Design and Analysis for Subjective Assessment of Visual and Taste Stimuli

Bareng A. S. Nonyane

Doctor of Philosophy
University of Edinburgh
2004

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(*Bareng A. S. Nonyane*)

Acknowledgements

I would like to thank my supervisors, Dr Chris Theobald and the late Mr Rob Kempton for their valuable advice throughout this project. I am grateful to Mr Kempton for suggesting the experiments and the types of data used in this project, and for providing the inspiration that has kept me going even after his sad and untimely death. I would like to express my gratitude to Dr Theobald for keeping his door always open for me, and for all his guidance and encouragement. I would also like to thank Dr Sarah Brocklehurst for her input.

I would like to thank all the staff and students at BioSS and the University of Edinburgh School of Mathematics, for their support and encouragement. I also thank all my beloved friends, who are too many to mention, for the good times we shared while I was working on this thesis.

Special thanks goes to my mum Ernestina for her support throughout my studies.

Thanks to the Hannah Research Institute for allowing me to use the apples data.

This work would not have been possible without the financial support of the Cecil Renaud Overseas Scholarship (South Africa).

Abstract

In areas such as agriculture and medicine, there is often a need for humans to carry out tasks of estimating the intensity of certain kinds of stimuli. A lot of experiments have been carried out by psychologists to study the performance of humans at such tasks. This thesis is concerned with quantitative visual assessments of plant disease severity and with food tasting studies. Judgements made by humans are prone to bias from several sources. Thus, the designs used for such experiments and the models used for analysis of such data need to account for this bias.

In studies where human assessors make judgements of long sequences of varying levels of a stimulus, sequential effects, such as carry-over from the previous stimulus and order of the stimulus in the sequence, are likely to arise. Designs which are balanced for order and carry-over have therefore been studied here, and a program which searches for sequences of one such design was written. The sequences generated by this program were then grouped according to certain invariance and optimality properties.

Calibration consists of comparing the performance of different measuring instruments that are used to measure similar samples of interest, as well as correcting for biases of some of the instruments. Here, humans were used as measuring instruments in a visual assessment experiment. A test experiment was carried out for which true stimulus intensities were known, and then calibration of responses from a subsequent similar experiment was done. This kind of calibration is known as absolute calibration because the true stimulus intensities in the test experiment were known. It was based on a Bayesian predictive method applied to a regression model of the responses on the true stimulus levels, with carry-over and order effects, as well as first order auto-correlation in the errors. A method to select the best assessors was based on the Shannon information criterion.

Data were analysed from a series of food tasting experiments in which a panel of assessors made judgements based on a number of attributes. Data from these experiments were combined in order to study assessor performance over time, and to use information about the assessors to improve analysis of their future performances. In this case, there was no standard measure of the intensity of attributes. Thus, the performance of each assessor was judged relative to the others in the panel, and this is called comparative calibration. Frequentist and Bayesian analyses were carried out based on a multiplicative model in which the responses were regressed on unknown parameters of the true attribute intensity for each food product, with the corresponding coefficients measuring the relative differences in the assessors' use of the scale.

Table of Contents

Chapter 1

Introduction

Repeated assessments of different kinds of stimuli by humans occur frequently in agricultural, biological and food research. Assessors are often presented with sequences of varying levels of the stimulus, and this results in repeated responses. Quantifying such assessments made by humans is an area that is referred to as *psychophysics*, as it has its roots in physiology and psychology. Experiments have been conducted over the years by psychophysicists with the aim of studying the way humans make such judgements. In these experiments, different kinds of stimuli such as sound, taste and visual stimuli have been studied. In psychophysics, the term *magnitude scaling* is often used to define the judgement of stimulus intensity.

Gescheider (1988) gave a review of some of the issues, such as different sources of bias in human judgement and the different models that have been proposed to describe the relationships between responses and stimuli. It has been found that human judgements are affected by the context in which they are carried out. There are often differences in the way assessors in judging panels use the scale, and each assessor's responses may suffer sequential effects such as order and carry-over from previous stimulus levels. The models and the designs used in these studies should therefore account for such effects.

Types of stimulus sequences which are appropriate for such repeated measures experiments have been suggested. These are sequential designs which are balanced for order and carry-over effects. As part of this thesis, a computer program was written in order to search for sequences that satisfy the properties of one such design proposed by Finney and Outhwaite (1956). A method of classifying the sequences generated by this program is suggested, as well as criteria for selecting optimal sequences.

In this thesis, assessments of visual and taste stimuli are studied. Visual assessments are studied in the context of plant pathology where repeated assessments of plant disease severity are carried out. Disease severity may be quantified by the percentage area of the plant organ, often a leaf, covered by disease damage, or by the number of spots caused by the disease (Krantz and Rotem (1988)). For this thesis, experiments were carried out in which subjects recorded visual assessments of images that mimic damaged leaves. The data from these experiments were investigated for evidence of carry-over and order effects as well as auto-correlation. A predictive calibration method of correcting the responses for individual biases was illustrated as well. For this method, a training experiment for which true percentage cover or counts of spots were known was carried out first. This gave information about the nature of assessors' bias, which was then used in correcting responses from a subsequent experiment using Bayesian predictive calibration. This form of calibration is called absolute calibration (Osborne (1991)). The best assessors for each task were selected using a method which is based on the Shannon information criterion.

For the taste stimulus, data from several apple-tasting experiments were obtained from the Hannah Research Institute in Ayr. The assessors gave their scores with respect to taste attributes such as sweetness. The main aim of the experiments was to monitor assessor performance over time, and thus data from these experiments were combined for analysis. In order to model attribute intensity, comparative calibration was carried out. That is, there was no standard measure to which to compare assessors' responses, and therefore each assessor's performance was measured relative to others in the panel. A multiplicative model which accounts for differences in the use of the scale was used for this. It was then shown how information on assessor performance from these experiments could be further used to improve analysis of future experiments.

1.1 Thesis Outline

In Chapter 2, a review of magnitude scaling experiments and findings in the psychophysics literature is given. Assessments of plant disease severity and sensory evaluation of foods are also introduced here.

Chapter 3 comprises a discussion of sequential designs and, in particular, those that were chosen for the visual assessment experiments in Chapter 4. The algorithm for the program that searches for some of these designs is also given here,

together with the classifications of design sequences generated by this program. Some criteria for choosing between sequences are also discussed.

In Chapter 4, the visual assessment experiments and their results are presented. Bayesian predictive calibration of individual assessor scores is demonstrated and an information criterion is used to select the best assessors.

Analysis of data from apple-tasting experiments is presented in Chapter 5. A Bayesian hierarchical model with multiplicative effects is used to analyse combined data from individual studies, and it is shown how information on assessor performance may be used for analysis of subsequent experiments.

Chapter 6 gives a summary of the aims of the thesis, how these were achieved and some suggestions for further work.

Chapter 2

Magnitude Scaling

2.1 Introduction

The task of assessing the intensity of a stimulus is often referred to in psychology as *magnitude scaling*, hence the title of this chapter. Quantifying such assessments made by humans is an area that is referred to as psychophysics as it has its roots in physiology and psychology. It involves studying the relationship between the response and the stimulus (Gacula and Singh (1984)), and the context in which judgements are made. A review of some of the work done on this is given in this chapter to describe the kinds of data, models and designs that have been proposed in the past. The application of magnitude scaling in assessment of plant disease severity is then discussed. A brief discussion of food tasting studies, which are also a special application of magnitude scaling, is also given.

2.2 Magnitude Scaling in Psychophysics

Magnitude scaling tasks are often affected by the context in which they are made. Also, because they are often carried out for sequences of varying stimulus levels, possible causes of bias include sequential effects. Sequential effects may be effects of the order of the stimulus level in a sequence and / or the interactions between successive stimuli or levels of stimulus in a sequence. These interactions are commonly known as *carry-over* effects and they may sometimes appear to be *assimilative* or *contrastive*. Carry-over is assimilative if the current response is biased towards the level of the preceding stimulus level and contrastive if the response is biased away from the preceding stimulus. It is important to note that carry-over is not always restricted to the immediately preceding stimulus level, but that it may also be from two or more preceding levels.

DiLollo (1964) did a study which showed how a contrastive effect was exhibited when groups of assessors were given tasks to judge a series of heavy objects and then shifted to a series of lighter objects, or vice versa. It turned out that a series that was judged first was used as a reference point for the second one, and a contrastive effect was exhibited. A positive contrastive effect was seen when assessors over-estimated a high weight series given that they had been through a low weight one before. A negative contrast occurred when they under-estimated low weight series after being through the high weight one before. This contrastive effect was seen to decrease as the length of the second series increased, implying that the effect of the previous series or reference level was forgotten over time.

Krueger (1972) investigated the perception of numerosity. The number of non-overlapping black dots on a white background, bunched together, was perceived as less than when the same number of dots was spread out. Krueger (1972) modelled the response as a power of numerosity, based on the power law proposed by Stevens (1957), and thus

$$R \approx kS^p,$$

where R is the response (perceived numerosity); S is the true stimulus level (numerosity); k is a constant and p is the numerosity exponent. The estimated value of p was found to be less than 1 (≈ 0.85), and there was an overall tendency to under-estimate the number of dots. In order to prove that R is indeed a power function of S in numerosity judgement, Krueger (1982) repeated the numerosity experiment where only a single judgement, instead of a sequence, was required. This ensured absence of sequential effects, and the results showed that the data did fit a power function with an exponent as high as when repeated judgements were made.

Lockhead and King (1983) argued that this model does not fully take into account the sequential effects in scaling tasks. They suggested a model in which the response depends on the current stimulus and memory of the previously observed stimulus, thus

$$R_t \approx S_t + a(M_{t-1} - S_{t-1}),$$

where R_t is the response at position t in a sequence, S_t is the current stimulus, a is a positive constant and M_{t-1} is the remembered value of the previous stimulus S_{t-1}, which is different from S_{t-1} itself. The quantification of this remembered value is not made clear in this paper.

The above model only takes into account the possibility of assimilation, and therefore, to account for contrastive effect, it was extended to

$$R_t \approx S_t + a(M_{t-1} - S_{t-1}) + b(\overline{M} - M_q),$$

where M_q is the average of recent $q(= t-2)$ memories; $\overline{M}$ is the average memory in the whole experiment, and b is a positive constant.

Morris and Rule (1988) conducted experiments where groups of assessors did either a numerosity or a length estimation task. Each of the tasks was carried out on a number of occasions in which sequences of stimulus levels were judged. They then calculated residuals of responses as

$$\log R_{tj} - \overline{\log R_t}, \tag{2.1}$$

where R_{tj} is response to stimulus t on occasion j, and $\overline{\log R_t}$ is the mean logarithm of that assessor's scores for stimulus t. The residuals on successive judgements were found to be positively correlated and also to be related to the position within the presentation sequence, which implied that there were trends in the sequence.

DeCarlo and Cross (1990) gave a wide review of models and theories of sequential effects in magnitude estimation as studied in the 1950's, 1960's and 1970's. They suggested models that account for cases in which assessors perform judgements relative to some frame of reference. Most of the models discussed are regression models with carry-over effects from previous stimuli or previous responses, and some auto-regressive error terms. These models were tested with various magnitude scaling experiments.

Some experiments also showed the effect of the length of time between presentation of stimulus levels within a sequence. DeCarlo (1992) termed this an *inter-trial interval*. DeCarlo (1994) and Sawyer and Wesenstein (1994) studied the relativity of judgement, that is, cases in which while assessing a sequence of stimuli, subjects tend to use previous responses as a point of reference for current ones. This was seen as another source of sequential effects. Auto-correlations of successive responses became larger when the point of reference was short term than when it was long term.

Sequential effects in a medical application of magnitude estimation of visual stimuli were investigated by Laming (1995): here pathologists repeatedly screen cervical smears to make diagnoses for cancer. Sequential effects were believed to

have caused a number of false positives. Two experiments mimicking the task of screening cervical smears were undertaken and an assimilative effect was discovered, indicating the positive correlation between successive diagnoses. The proposed solution to this was to have a library of cervical smear samples for which the correct diagnosis was known. These were inserted at random points of the sequence to be assessed, and at these points feedback was given after the score was entered. This improved pathologists' performance in general.

Experiments conducted by Schifferstein and Oudejans (1996) on judgement of saltiness of solutions showed a presence of contrastive sequential effects. They found that after a repetition of a task with one taste stimulus, a change to a different one results in over-estimation, and this is called *successive contrast*. Since this experiment involved a taste stimulus, it may also be seen as a case of sensory evaluation of food which is discussed in the next section.

2.3 Sensory Evaluation of Food

Sensory evaluation of food may be seen as a special case of magnitude estimation which involves judgement of sensory attributes such as texture, smell and sweetness of food and drink products. It differs from the psychophysics studies discussed above because there are normally many attributes of the same food product on which judgements are made. Also, it is hard to define one objective scale because the true stimulus intensity is often not known or hard to define unambiguously. Assessors tend to show a lot of variability in their use of the scale, and therefore the models used to analyse the data need to take this into account. Sensory evaluation of food plays an important role in food science and market research, for example, to establish consumer acceptance of new food products on the market.

Gacula and Singh (1984) gave a review of psychophysical aspects of sensory evaluation of food. These aspects were mainly the different kinds of scales used and the modelling of the relationship between the response and the stimulus. They also discussed possible block designs, analysis of variance models and some non-parametric methods of analysis. A compilation of more reviews of sensory evaluations of food is given in Piggot (1988).

Stone and Sidel (1993) gave a discussion of the practice of sensory evaluation, focusing on the planning and carrying out of sensory studies in order to obtain

meaningful data. The selection of assessors and scale types, setting up tasting environments or venues, decisions about designs followed to present the products, and some descriptive analysis of the data were also discussed.

2.4 Count and Area Estimation in Plant Disease Assessments

The motivation for the experiments and analyses carried out in Chapter 4 is the desired accuracy in visual assessments of diseases by plant pathologists. This is an interesting application of magnitude estimation, and studies have been carried out in plant pathology to try to improve accuracy. The terminology used in plant pathology, though, differs from that used by psychophysicists. Disease intensity is quantified in terms of *incidence*, which is the percentage of diseased plants or plant parts in a population, or in terms of *severity*, which is the percentage of a particular plant organ (e.g. leaf) covered by lesions of the disease. Severity is also sometimes measured by the number of patches caused by the disease on a plant organ. This project concentrates on disease severity as measured by count and percentage cover of lesions.

Krantz and Rotem (1988) gave an introductory discussion of the issues that are important in measuring disease intensity, highlighting the fact that quantifying disease intensity is a key to proper diagnosis. It also plays an important role in predicting crop yield (if yield is affected by diseases) and measuring susceptibility of certain plant varieties to diseases. An important first step towards disease assessments is the selection of the sample of plants or plant organs to be assessed. An example given by Krantz and Rotem (1988) is that if the objective is to establish whether yield loss due to rust is correlated with disease on flag leaves, then sampling units would be the flag leaves. Sampling also has to take into account the spatial distribution of the disease in order to get a representative sample.

There are other techniques that may be used to measure plant disease besides visual assessments by humans. Automated image analysis is one of them, and it may be more accurate than human visual assessments. Two problems with it are that it is costly and often involves destroying the plant. Remote sensing is also sometimes used and it is non-destructive as it involves looking at the sunlight reflection on diseased areas. The problem with it is that percentage reflectance relies on the sun angle and on leaf condition such as wetness. Thus, human

visual assessments are preferred when cheap and quick repeated assessments are required, but they are prone to subjective bias. A lot of work has been done with the aim of finding ways to improve these assessments, and one of these is discussed in Chapter 4.

Plant pathologists are concerned with what they call *accuracy*; the conformity to a given standard, and *precision*; which is consistency in scoring images of the same size repeatedly. Precision is often measured by the coefficient of variation obtained from the least squares regression of the responses on the true disease intensity. Nutter et al. (1993) compared the accuracy and precision of the three techniques of quantifying disease assessments on estimating Dollar Spot on bentgrass, and Parker et al. (1995) found that visual assessments gave more biased estimates than image analysis methods. Most studies show that over-estimation tends to occur most at low levels of severity.

Sherwood et al. (1983) discussed what they referred to as *illusions* that influence visual assessments of *Dactylis glomerata L.* (i.e. leaf spot on orchard grass). In psychophysics terminology, this would be referred to as context effects. They found that there was a general tendency to over-estimate the number of spots. Also, if two leaves with similar total percentage cover but differing numbers of spots are assessed, the one with a larger number of spots is perceived as having a larger total area of spots. Similar tendencies were discussed in the psychophysics work mentioned earlier by Krueger (1972) on perceived numerosity of black spots on a white background. The size of spots and the way they are scattered (bunched together or spaced out) affects perceived numerosity.

2.4.1 Modelling and improving scores of disease severity in plant pathology

Krantz and Rotem (1988) pointed out that often in plant pathology, the relationship between the estimates of plant disease severity and the true severity follows the Weber-Fechner law which states that the response is proportional to the log of the stimulus. As mentioned earlier, sometimes the objective of the disease assessments is to find a function that relates yield loss to disease severity. Shaw and Royle (1989) studied yield loss in wheat due to epidemics of foliar diseases caused by *Mycosphaerella graminicola*. Absolute estimates of disease severity for this purpose were obtained by regressing visual estimates on subsamples of leaves on which measurements were taken using image analysis. Nutter and Guan (2002) used information from visual assessments and remote sensing to quantify alfalfa

yield loss due to foliar diseases such as leaf spot. Remote sensing assessments were better at predicting yield loss than visual assessments. They showed that foliar diseases decreased yield significantly.

The desire to reduce bias in visual assessments led to the development of various training methods for human assessors. Conventional methods involve giving assessors a training set of photographs of diseased leaves for which true count or percentage cover is known, and then giving feedback on those after the assessors' scores have been entered. These kinds of methods are built into computer programs used for training assessors, such as DISTRAIN by Tomerlin and Howell (1988) and Disease.Pro by Nutter and Schultz (1995).

The work of Ferris et al. (2001) and Ferris (1999) was motivated by the count and percentage estimation of disease lesions on leaves. They looked at bias due to carry-over, and carried out experiments in which the stimuli were images of black dots on white square backgrounds. Such images were used to mimic disease lesions on leaves. Their experiments showed carry-over taking the form of assimilation at the 5% significance level for a single subject. Change-over designs balanced for carry-over were used to present the sequences of images, and different regression models which account for carry-over and autoregressive errors of responses were explored. These were combinations of models earlier discussed by Finney (1956), Stevens (1957), DeCarlo and Cross (1990) and DeCarlo (1994). Ferris et al. (2001) further proposed a proportional carry-over model where carry-over was modelled as a logistic function of the difference between two successive stimulus intensities.

Chapter 3

Sequentially-balanced Designs for Sensory Assessment Experiments

3.1 Introduction

The visual assessment experiments to be discussed in Chapter 4, like many other magnitude scaling experiments, give rise to sequences of responses at varying levels of a given stimulus. Like repeated measures in cross-over experiments, these experiments require appropriate designs to balance for sequential effects in the form of order and carry-over. Abeyasekera and Curnow (1984) discussed the importance of always adjusting for carry-over effects in the design and analysis of cross-over experiments, despite the increased variance resulting from this.

The designs used in the existing literature on magnitude scaling studies are seldom properly balanced for efficient estimation of the sequential effects that might arise. DeCarlo and Cross (1990) and DeCarlo (1992), for example, presented the levels of the stimulus of interest in randomised sequences which were not properly balanced. Balance in this context implies sequential balance such that, for n stimulus levels, all the possible n^2 ordered pairs of stimulus levels occur the same number of times in a design sequence.

This chapter discusses some possible sequential designs and how they differ in terms of balance for order and carry-over effects. A type of design proposed by Finney and Outhwaite (1956) is found to be the best, if it exists, as it has sequential balance. An algebraic way of constructing sequences under this design is not known, and therefore a C++ program was written to systematically search for all possible sequences for a given value of n. These may then be grouped into classes according to some invariance and optimality properties.

3.2 Sequential Designs

In experiments where long sequences of treatments are applied, there is a need to reduce any long term trend effects. This may be done by arranging treatments in replicates of treatments, which form blocks. In such experiments, there is also often a possibility of correlation between neighbouring responses. Neighbour-balance should therefore be accounted for in the design. This means that each treatment should be followed by every other treatment equally often. In other words, all pairs of treatments should occur equally often. In the case of magnitude scaling experiments, the direction of such balance is important: it has to be *sequential*, whereas in field experiments balance may be in either direction. Thus, in magnitude scaling experiments, balance is required for each individual contrast, and so it is not only *all pairs* but *all ordered pairs* of treatments that should occur equally often, say k times. Furthermore, Ferris et al. (2001) found that for their carry-over model, in order to estimate carry-over effects efficiently, all ordered pairs need to occur and also, each treatment should be preceded by another of the same type. This is referred to as *self-adjacency*.

What follows is a discussion of the designs that were explored for the visual assessment study. Initially, we were unaware of the literature on suitable one-dimensional designs, and thus some Latin square designs were manipulated to form one-dimensional sequences. These were then used in the visual assessment pilot study (Experiment 1, Chapter 4). Later on, though, suitable one-dimensional designs were discovered and found to be more appropriate. These were then used in Experiment 2. All these designs are now discussed as follows.

Williams (1949) proposed a Latin square design for experiments in which a series of treatments are applied to the same subject. Three possible cases of carry-over effects were discussed, namely carry-over from a single preceding treatment, carry-over from any number of preceding treatments and their interactions, and lastly, carry-over from two preceding treatments and their interactions. In this project, only the first case of carry-over was considered. The type of design proposed by Williams (1949) is such that the rows of a Latin square correspond to subjects, while the columns correspond to treatment order. The conditions are such that each treatment is preceded by every other equally often, and each treatment occurs equally often at each position in a square. If n is the number of treatments, conditions for balance can be achieved using only one Williams (1949) Latin square for even n, whereas for odd n, balance is achieved by a minimum of

two such Latin squares.

For the visual assessment experiments, the number of treatments (visual images of varying levels of cover) was seven, so two Latin squares were required. These were then manipulated by taking rows to form blocks of a one-dimensional sequence, with self-adjacencies occurring at the end and beginning of each block. In order to have an equal number of self-adjacencies for each treatment, the same treatment as the last one was placed at the beginning of the sequence. The response to this leading treatment was not included in the analysis, though. This is because this response does not have any carry-over effect, and as it does not belong to a complete block, it only gives information on that one particular treatment.

The transformation of this Williams (1949) Latin square design to a one-dimensional sequence is illustrated for $n = 7$ with the following two Latin squares.

1	2	5	4	3	6	7
2	3	6	5	4	7	1
3	4	7	6	5	1	2
4	5	1	7	6	2	3
5	6	2	1	7	3	4
6	7	3	2	1	4	5
7	1	4	3	2	5	6

1	6	3	5	7	4	2
2	7	4	6	1	5	3
3	1	5	7	2	6	4
4	2	6	1	3	7	5
5	3	7	2	4	1	6
6	4	1	3	5	2	7
7	5	2	4	6	3	1

A one-dimensional sequence of length 99 formed from the rows of these Latin squares is

1 1254367 7143256 6732145 5621734 4517623 3476512 2365471
1635742 2746153 3157264 4261375 5372416 6413527 7524631.

In this sequence, all ordered pairs and self-adjacencies occur twice each.

Williams (1952) suggested designs for field experiments in which plots are arranged in a one-dimensional sequence. Such designs are made up of $m \leq n$ blocks containing each of the n treatments only once. Williams (1952) defined two types of such a design. Type A is the one in which each treatment occurs equally often (k times) adjacent every other treatment. Type B is the one in which each treatment occurs equally often adjacent every treatment including itself (that is, one of its own kind). To ensure balance, an additional plot receiving the treatment applied to the last plot is placed at the beginning of the set of blocks. A Type A design is constructed such that $2m = k(n-1)$, and for Type B, $2m = kn$ and n must be even. In order to construct sequences under these designs, one may use a diagram in which all treatment symbols are arranged in a

circle, and then trace out a sequence by joining treatments by a continuous line so that all possible joins are made k times.

The problem with the designs suggested by Williams (1952) is that balance relates to un-ordered pairs and not ordered ones. So neighbour-balance is not sequential because there is no restriction on order of adjacency. An example for each type for $n = 4$ is given as

Type A:	2	1234	2314	3142	
Type B:	1	1234	4132	2413	3421.

As one can see from these examples, even if all pairs occur, they are not necessarily uniquely ordered. Type B has self-adjacencies but still, it has no restriction on order of adjacencies.

Finney and Outhwaite (1956) suggested Type I and Type II designs similar to Type B and Type A of Williams (1952), respectively. These designs were chosen for use in the visual assessment Experiment 2 in Chapter 4. They were proposed in the context of bioassay studies where it is often necessary that if there is a number of treatments to be tested, all be applied to a single subject, thus reducing variability for estimates of treatment effects between subjects. The definitions of these two types of designs are given as follows, where blocks are complete replicates or permutations of n treatments.

Type I sequences comprise an initial treatment followed by kn successive blocks of complete replicates. So, the initial treatment occurs $kn + 1$ times while all others occur kn times. Each of n^2 possible ordered pairs of successive treatments occurs k times.

Type II sequences comprise an initial treatment followed by $k(n-1)$ successive blocks of complete replicates. So, the initial treatment occurs $k(n-1)+1$ times while all others occur $k(n-1)$ times. Each of $n(n-1)$ possible ordered pairs of different treatments occurs k times.

Thus, every direct effect of every treatment occurs k times with the carry-over effect of each treatment including itself (for Type I) or of every other treatment (for Type II). Here again, the observation made on the initial treatment is not included in the analysis as it is only placed there to ensure balance. Williams (1949) Latin square designs can be changed into Type I designs with index $k = 2$ when n is odd, as was shown earlier. This is achieved by choosing any row to be

the first block and following it with the rest of the rows so that self-adjacencies occur at the ends and beginnings of blocks. An initial treatment, similar to the last treatment in a sequence, is then placed at the beginning of the sequence.

Finney and Outhwaite (1956) pointed out that the Type I sequences do not exist for $n = 3, 4, 5$, and that there exist many sequences for $n = 6$. They speculated that Type II sequences exist for all values of n. An example of a Type I sequence for $n = 6$ is given by

$$1 \quad 123456 \quad 635142 \quad 216543 \quad 315264 \quad 461325 \quad 536241$$

Sampford (1957) and Street and Street (1987) presented methods for the construction of the Finney and Outhwaite (1956) designs. Sampford (1957) discussed a general method of constructing certain classes of Type II sequences with index 1 and Type I sequences for index 2, and also looked at the analysis of data using these designs. For Type I designs with $k = 1$, he could not devise a general method of construction and showed that for $n > 2$ it is not possible to find a Latin square whose rows can form blocks of a sequence under this design. He gave examples of these kinds of sequences for values of n between 6 and 11 and for $n = 14, 18$ and 22. Sampford (1957) also identified a special case of a Type I design with index 1 and $n = 2r$ when r is odd, and one such design is known for r treatments.

Street and Street (1987) showed that the Type I designs with index 2 exist for all $n \geq 4$. They gave a general method for the construction of these which is based on an associated Latin square. They also mentioned that no general construction of Type I designs is known for index 1. For the Type II designs where the index is 1 and $n \geq 4$, they gave a general construction. This involves the use of a cyclic Latin square which is a Latin square whose rows are constructed by cyclic development of the initial row.

3.3 Systematic Search for Type I Sequences with Index $k = 1$

For the visual assessments Experiment 2 in Chapter 4, the Finney and Outhwaite (1956) Type I designs with index 1 were chosen. This is because they have self-adjacencies and they are easy to analyse as all ordered pairs exist. Also, this class of designs is more desirable than the one with the index k greater than 1 because

in some cases, very long sequences of treatments are impractical to use. Since no algebraic method of generating sequences under this class is known, a C++ program was written to search for the sequences systematically for any n, while keeping the first block in standard order: $1, ..., n$.

The symbols used by the program to denote treatments are integers $1, 2, ..., n$. Treatments of interest would then be allocated randomly to the symbols when the sequence is used for an experiment. It basically starts off with a vector of length n^2 filled with 1s, except for the first block which is filled with $1, ..., n$, in standard order. The last $n-1$ blocks are then filled by systematically making appropriate changes to the entries, doing appropriate checks to ensure that each symbol occurs once in each block, ensuring sequential balance and forming self-adjacencies. For simplicity, the program generates the sequences without the initial symbols at the beginning. These symbols are obviously 1's for all sequences since the first blocks are always in standard order and the last symbols are 1s. The program is given in Appendix A and a brief description of the algorithm is given below.

- Create a vector of length n^2 with 1s as entries. Put symbols 1 to n in the first n positions of this vector. These entries must be in standard order and they define the first block of the sequence.

- **BEGIN**: While you have not reached position n^2, continue to **Level Check 0**. If n^2 is reached, output the sequence and to search for another sequence, go back to the end of block 2, restoring entries to 1s as you go along, and then go to **Level Check 2**.

- **Level Check 0**: Move to the next position in the sequence. If sequence position mod$n = 1$ (i.e. beginning of block, called *block mark*), copy previous symbol to the current position to create a self-adjacency. Then go to **Adjacency Check**. Otherwise, if not at the beginning of a block then go to **Level Check 1**

- **Level Check 1**: Check if the current symbol has appeared in this block before. If so, go to **Level Check 2**, otherwise check if the current pair has occurred before in the sequence so far. If the pair has appeared before, go to **Level Check 2**, otherwise accept this move and go to **BEGIN**.

- **Adjacency Check**: Check if this self-adjacency has occurred in the sequence so far. If not, mark this position as the beginning of a block, move to the next position and go to **Level Check 1**. If this self-adjacency has

occurred in the sequence so far, then go back twice restoring entries to 1s, go back once more, restore *block mark* to (*block mark*$-n$) and go to **Level Check 1**.

- **Level Check 2**: If current position's entry is less than n, then increment it by 1 and go to **Level Check 1**. Otherwise change it to 1 and go back once. Check if at the beginning of a block. If not, increment current entry by 1 and go to **Level Check 1**. Otherwise, go back twice restoring entries to 1, go back once more and check if the symbol in this sequence position is n. If not, go to **Level Check 2** and if so, it means the search has gone all the way back to the first block, which means all possibilities have been exhausted for this value of n, and so write *"end of possibilities for this value of n"*. The program then terminates.

This program provides a way of searching and / or generating sequences for any value of n under the Type I design with index 1 in a way that was never available before. All the sequences generated are in standard order. For each of these of length n, $n!$ distinct sequences, which are not necessarily in standard order, can be obtained by permutation of the symbols $1, ..., n$. This program was tested, and produced such sequences, for all values of n between 6 and 20. Sampford (1957) gave examples for $n = 2, 6, 7, 8, 9, 10, 11, 14, 18$ and 22 only.

For $n = 6$ and 7 the program was run to completion, generating 324 and 175588 sequences, respectively. For $n = 8$, it was stopped when the number of sequences generated exceeded 7.6 million, and so for values of n greater than 8 it was just tested to see if it found any sequences, but was not run to completion. The use of this program thus partly solves the problem stated by Finney and Outhwaite (1956), who gave a detailed discussion of the case of Type II sequences for $n = 4$, as used in 4-point assays. They stated

> For greater values of t whatever the index, random selection from all sequences seems impossible until they have been systematically enumerated. In practice, choice of an arbitrary sequence followed by randomisation in respect of block permutation, reversal and letter permutation should be adequate.

In this thesis, a way of systematically enumerating Type 1 sequences with index 1 is given by the program and moreover, we look at some optimality properties

of these sequences in order to be able to choose the best ones to use in an experiment. One optimality property could be the one that is related to the number of occurrences of each of the treatments in each of the within-block positions. Also, it is of interest to find some classifications of these sequences so that only a subset of sequences for each value of n needs to be stored if the rest of the sequences can be generated from this subset. The next subsections discuss criteria for choosing the best design sequences and a way of classifying them.

3.3.1 Criteria for choosing design sequences

Ideally, a design sequence among a set generated by the program for a given value of n would be perfectly balanced in the sense that each treatment occurs in each of the n within-block positions only once throughout the sequence. This is not possible for this type of design as shown by Sampford (1957), but some sequences may be closer to achieving this balance than others. Thus a criterion is devised where for each sequence, the initial sequence is ignored and an $n \times n$ matrix is created in which the rows represent treatments and the columns represent the n within-block positions. Each cell then has the number of times the corresponding treatment occurs in the corresponding within-block position. Thus, consider the following design sequence for $n = 7$, without a leading symbol at the beginning.

$$1234567 \quad 7153624 \quad 4165273 \quad 3572146 \quad 6317425 \quad 5476132 \quad 2643751.$$

The corresponding incidence matrix, C, would be

$$\begin{bmatrix} 1 & 2 & 1 & 0 & 2 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 2 & 1 & 1 & 1 & 1 \end{bmatrix}$$

The requirement for self-adjacencies means that the first and last columns of this matrix always consist of 1s. A possible criterion value is then given by

$$\sum_i \sum_j (C_{ij} - 1)^2, \tag{3.1}$$

where i and j denote treatment and within-block subscripts, respectively. For a perfectly balanced sequence the value of (3.1) is zero, but in the case of Type I

sequences with index 1, a good sequence is the one that gives the minimum value for this criterion. For the purpose of this project, this criterion is referred to as the *sum of squares*.

Bradley and Yeh (1980) and Yeh et al. (1985) discussed trend-free and nearly trend-free designs, respectively. These are designs where direct treatment effects are orthogonal (or nearly-orthogonal) to previous treatment effects and to linear or higher order trends. They did this in the context of block designs. In the case of visual assessments, blocks only provide a method of constructing designs so that treatments occur roughly uniformly over the sequence, but they have no effect on the analysis. Also, assessors are not aware of these blocks. Thus block effects are not included in the visual assessments model. There would be block effects if this was a field situation where blocking corresponds to some differences in the environments. The criteria for trend-free designs proposed by Bradley and Yeh (1980) and Yeh et al. (1985) are therefore modified here so that they exclude block effects.

Consider the following model, in the context of sensory assessments where a total of $N = n^2$ responses per design sequence are analysed:

$$y_t = \mu + \sum_{i=1}^{n} \delta_t^i \tau_i + \sum_{a=1}^{p} \theta_a \phi_a(t) + \xi_t, \tag{3.2}$$

where y_t is the response to the treatment in sequence position t where $t = 1, ..., N$; μ is the overall mean; δ_t^i is an indicator variable equal to 1 when treatment i occurs in position t and equal to 0 otherwise, $i = 1...n$; τ_i is the ith treatment effect; $\phi_a(t)$ is an orthogonal polynomial of degree a, where $a = 1...p$; θ_a is a regression coefficient of the orthogonal polynomial and ξ_t is the error term. This model ignores the carry-over effects and that is because such effects are not expected to be very significant and also ignoring them simplifies the criterion. Following Bradley and Yeh (1980), a trend of order a is orthogonal to treatment allocations in this model if

$$\sum_{t=1}^{N} \delta_t^i \phi_a(t) = 0$$

for all i.

This is not achieved by design sequences of Type I with index 1 for a given value of n, because even though treatment allocations, or direct effects, are orthogonal

to block effects and to carry-over effects, they are not orthogonal to within-block positions, and thus only near-orthogonality is considered. For near-orthogonality of treatment effects with a given trend, the type of criterion proposed by Yeh et al. (1985) was used. That is, the design that minimises

$$\sum_{a=1}^{p}\sum_{i=1}^{n}\left\{\sum_{t=1}^{N}\delta_t^i\phi_a(t)\right\}^2, \tag{3.3}$$

among a set of designs, is the one that is most nearly trend-free (NTF). This requires that all treatment allocations be as nearly orthogonal to all specified trends as possible. In this project, only near-orthogonality to linear trend ($p = 1$) was considered and for this, (3.3) is equivalent to

$$\sum_{i=1}^{n}\left\{\sum_{t=1}^{N}\delta_t^i(t-\bar{t})\right\}^2, \tag{3.4}$$

since

$$\phi_1(t) = \frac{t-\bar{t}}{\sum t}, t = 1, ..., N \tag{3.5}$$

where $\bar{t} = \frac{N+1}{2}$.
The expression (3.4) is used instead of (3.3) because without the normalising constant, the criterion values are expressed as integers. So this simplified linear NTF criterion will be referred to as NTF1.

3.3.2 Design sequence classification

There are three reasons for putting these sequences into sets, namely

- to reduce storage space by storing representatives of the classes only (not necessary for $n = 6$, though);
- to examine how sequences are related in terms of invariance to the optimality criteria; and
- to possibly suggest an algebraic method for generating these designs.

In order to classify these sequences that are in standard order into sets, two transformations that may be applied to each of them were considered. These are *reversal* R and *block shifting* S. So S^b means shifting the first b blocks, $b = 0, ..., n$, of the sequence to the end and reversal involves reading the sequence in reverse, from the end to the beginning. After each of these transformations, the first block of the new sequence is changed into standard order and corresponding mappings are done for the rest of the blocks. Since the new sequences are in standard order the leading symbols are 1s, as before.

If I denotes an operation that leaves the sequence unchanged, then it is easy to see that

$$\mathrm{R}^2 = \mathrm{S}^n = \mathrm{I}$$

and that

$$\mathrm{S}^b\mathrm{R} = \mathrm{RS}^{n-b}, b = 0, ..., n-1.$$

This means that any combination of these two transformations is expressible as either S^b and RS^b for some b in $0, ..., n-1$, and there can only be at most $2n$ such transformations. Thus, a transformation set made up of a sequence and all other sequences generated from it by these transformations, is expected to have a maximum of $2n$ sequences. Finney and Outhwaite (1956) discussed the application of these transformations to all possible sequences for a given value of n, whereas here, the transformations are discussed as applied to sequences which are in standard order only. Hence according to Finney and Outhwaite (1956), a transformation set has a maximum of $2kn(n!)$ sequences. Here, a transformation set of sequences in standard order has a maximum of $2kn$ sequences, but as mentioned earlier, each of these can produce up to $n!$ sequences by symbol permutation. This would then result in a maximum of $2kn(n!)$ sequences per transformation set for a given value of n.

The case of $n = 6$, for which 324 sequences in standard order were generated by the program, was used to illustrate the classification of sequences into such sets. For this case, there were 28 distinct sets: 26 of those comprised 12 different sequences while two had 6 different sequences each. For the two sets with 6 sequences each, it turned out that for each sequence, $S^3 = I$, resulting in repetitions.

In order to classify the sequences, a systematic approach was followed. The sequences were stored in a file *sequence file*. A brief summary of the algorithm for the classification program is as follows

- Open *sequence file* in which sequences for $n = 6$ are stored in rows 1 to 324. Define two pointers to this file, *pointer 1* and *pointer 2*.

- Read the first design sequence as pointed by *pointer 1*. This sequence is now called current sequence. While *pointer 1* has not reached the end of the file continue as follows.

- For each of the possible $2n$ transformations S^b and RS^b, for b in $0, ..., n-1$, do:

  apply the transformation to current sequence and obtain a new sequence; using *pointer 2*, go through the *sequence file* to look for a matching sequence. When a match is found, report its position in file, that is 1,...,324, and go back to apply another transformation to current.

- After all the transformations and matchings for current sequence, go back to read the next current sequence from *pointer 1*.

The 28 sequences which are representatives of the sets are given in Table 3.1. The first column of this gives labels of representatives; the next 6 columns are the blocks of the sequences and the last two columns are their corresponding values of the two criteria, sums of squares and NTF1. The brief algorithm for forming these sets explains why there are numerical gaps in the labels for the representatives. The representative labels with a * are those whose classes have only 6 unique sequences. The values of the two criteria given were found to be invariant to the reversal and shifting operations for a given set. Orthogonality to higher trends may be considered for NTF in (3.3) as well, but the quadratic trend criterion is not invariant under the two operations of reversal and shifting, and therefore using it would make it difficult to choose between sets.

Table 3.1: Representative sequences of Type I, $k = 1$ for $n = 6$

Class							SS	NTF1
1	123456	613254	415263	351462	243165	536421	24	252
2	123456	613254	462153	365142	241635	526431	18	256
3	123456	613524	415362	251643	314265	546321	14	254
4	123456	613524	415362	251643	321465	542631	14	254
5	123456	613524	415362	265143	316425	546321	22	244
6	123456	613524	416253	364215	514632	265431	20	248
7	123456	613524	416253	365142	264315	546321	14	246
8	123456	613524	421653	326415	514362	254631	18	238
9	123456	613524	425163	315462	2 14365	532641	20	270
10	123456	613524	426315	516432	214653	362541	18	254
12	123456	613524	432165	514263	315462	253641	20	270
13	123456	613524	436215	531642	251463	326541	16	252
15	123456	613542	241653	362514	463215	526431	18	252
16	123456	613542	251463	315264	432165	536241	18	274
18	123456	613542	251463	326415	531624	436521	20	244
20	123456	613542	264153	314625	516324	436521	22	266
23*	123456	614253	315264	432165	541362	246351	20	268
26*	123456	614253	315264	465132	241635	543621	16	244
27	123456	614253	315462	241635	513264	436521	16	262
29	123456	614253	315462	264135	516324	436521	16	246
30	123456	614253	316524	463215	541362	264351	18	254
32	123456	614253	354162	246315	513264	436521	14	234
33	123456	614253	354162	246315	521364	432651	14	248
35	123456	614352	215364	462513	324165	542631	14	254
40	123456	614352	253164	426513	362415	546321	14	230
42	123456	614352	265413	364215	516324	462531	18	254
43	123456	614352	265413	364215	531624	463251	22	254
54	123456	615324	425163	314652	264135	543621	18	230

Chapter 4

Predictive Calibration of Quantitative Visual Assessments

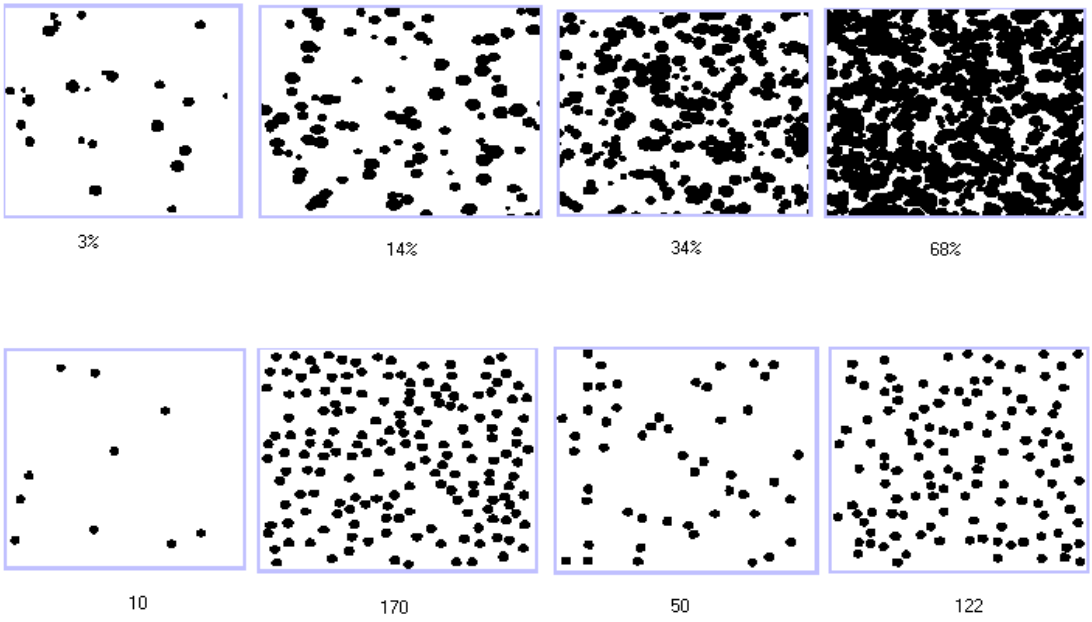
4.1 Introduction

This chapter gives a discussion of the visual assessment experiments which were carried out to mimic the assessments of disease severity in plant pathology. The aims here are to improve the apparatus and the procedure for these experiments as they were carried out by Ferris et al. (2001); to illustrate a predictive calibration method used to correct assessors' responses for bias, and also to illustrate an information criterion for choosing the best assessors in a panel. Images of black dots on white background, similar to those of Ferris et al. (2001) were used. Examples of these images are given in Figure 4.1, where the first row has samples for percentage cover estimation and the second one has samples for count estimation.

A program was written to automate the presentation of sequences of images on a PC monitor, instead of presenting images on an overhead projector as was done by Ferris et al. (2001). The designs used are as discussed in Chapter 3. First, a pilot study was carried out in which only the cover estimation task was done. Then a second experiment was carried out with both the cover and count estimation tasks, and for this, a change in the method of entering the scores was made.

A Bayesian predictive calibration method was used to correct assessors' scores for bias. This requires that an assessor first carries out a test experiment for which true cover or count levels are known. Then, a parametric model is fitted to the scores obtained from this in order to evaluate the extent of the bias. This model includes bias due to a preceding image level, position of an image in a sequence and auto-correlation among errors of the responses. A subsequent experiment is then carried out and the scores from this are corrected for bias using information from

Figure 4.1: Sample images for cover and count



the test experiment. As well as correcting the scores for bias, a criterion based on the Shannon measure of information was used to measure assessor performance so that the best assessors may be selected.

4.2 Experiment 1: Pilot study

This experiment involved estimation of percentage area of white squares covered by black circular dots as seen in Figure 4.1. The subjects who took part in it were 2nd year students taking a Statistical Inference module and aged between 18 and 22. There were 12 females and 13 males. Their vision was either good or corrected to good. The incentive offered was £25 for the most accurate participant, where accuracy was measured by the value of the mean difference between the scores and the expected levels of cover.

A design followed for presenting images in the pilot study was the one by Williams (1949) as described in Chapter 3. This is a Latin square design which, for the purpose of this experiment, was converted into a sequential one, balanced for order and carry-over effects and having self-adjacencies for each treatment, treatments in this case being the varying levels of cover of the dots. Seven nominal levels

of cover were chosen, and these were 5, 10, 17, 26, 37, 50 and 65%. These were chosen to cover the range of disease severity only up to 65% because it was thought to be the range for which it is most important to make the right distinction and diagnosis of severity. Any severity above 65% would simply lead to a conclusion that the plant or plant organ in question is severely diseased and the appropriate action would be taken depending on the purpose of the disease assessment.

The seven cover levels were then randomly allocated to the numbers 1 to 7 in the design sequence. A program written in Visual Basic (idea by R. Kempton and program written by Alec Mann) was used to generate images according to these levels, and each student could run this on their own PC. Each of the black dots generated had a radius of between 0.01 and 0.025 pixels $\times$ the width of the whole image, which was 768 pixels. For each nominal level of cover, this program generated a pool of images with percentage cover equal to nominal level $\pm 1\%$. Sequences were then generated by randomly picking images from these pools, following the design. The students were not aware of the existence of these nominal levels. Since an initial image level was placed at the beginning of the sequence for balance in self-adjacencies, each subject viewed 99 ($49 \times 2 + 1$) images per sequence, but as mentioned before, the response to this initial image was not included in the analysis. As there were two sessions with a break in-between, each student viewed and gave estimates of cover for 198 images in total.

The program started off with a training session at the beginning, where six images were displayed with levels of cover in the same range as the levels used in the experiment. Subjects were given 6 seconds to view each image and to enter their score using the keyboard. After this, the expected percentage cover was revealed and the subjects prompted to press the return key to view the next one. After this training session, the program then proceeded to the main experiment with the following instructions given on the screen.

> You will be presented with images of black circles on a white square with a grey surround and you are requested to enter your estimate of the percentage (0-100) cover of the circles. You will be given 6 seconds, with a beep at the 4th one, for each image. After entering the score, wait for the next image - you do not need to press the return key! If you make a mistake do not panic, simply concentrate on the next image. You will see a sequence of 198 images with a break after the first 99. During this break, you will be asked to press a key when

ready to continue. Please note that you will not be told the actual score after each image as it was done in the training session.

Ferris et al. (2001) presented similar images to groups of subjects on an overhead projector, and this is thought to have introduced bias because the subjects were not the same distance away from the images. This time, such bias was to be reduced as each person viewed automatically-generated images on their own monitor. Also, for Ferris et al. (2001), some mistakes happened when placing the transparencies on the projector. Hence some responses were omitted and the correspondence between subsequent responses and the images was uncertain. In this project, using the keyboard for entering the responses and the automatic storage of these responses to a result file reduced data entry errors, such as missing responses and non-correspondence between responses and true cover levels.

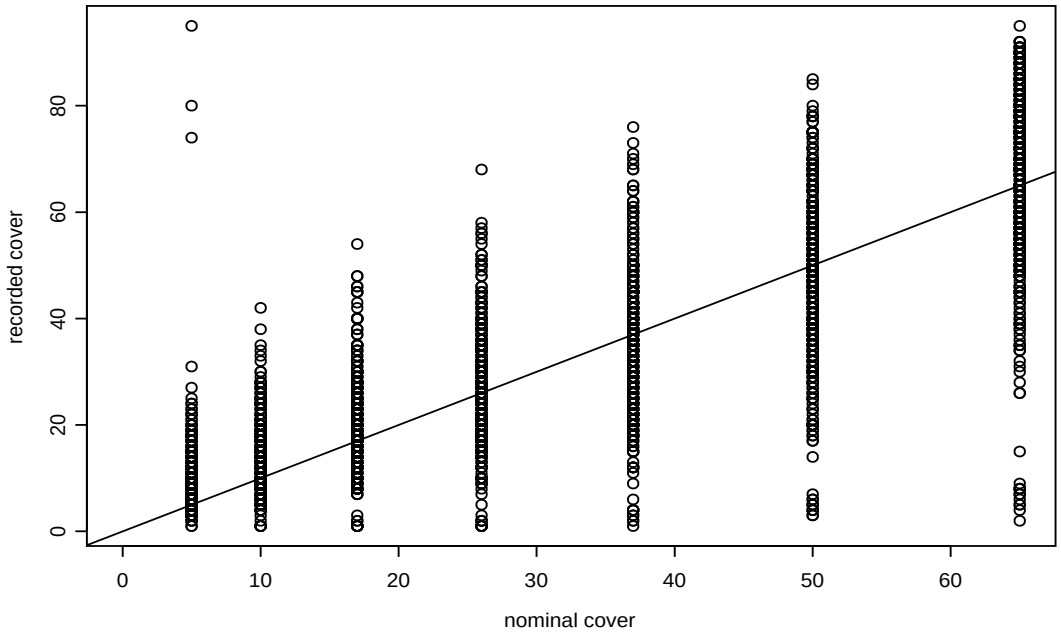
4.2.1 The data

For each subject, a result file containing columns of nominal levels and recorded responses were stored. Data from 2 of the 25 subjects who took part were discarded because one of them did not complete the experiment, and the other one had too many errors that indicated that he got completely confused with the task.

The plot of the responses versus nominal level of cover, as well as a line of equality, in Figure 4.2 shows an overall roughly linear increase in response with the increase in nominal cover. There is some bias for all levels of cover, and the variance increases as the expected levels increase. This figure also shows some outlying observations. Most outliers in high nominal levels of cover tend to be single-digit scores, so some of these were thought to have occurred when a person entered the first digit of a two-digit score and ran out time before entering the second digit. High outliers for the 5% nominal level were possibly from assessors accidentally typing an extra digit for a one-digit response.

If analysis of these data was required, the outliers would be treated according to whatever method of treating outliers seems suitable, for example, removing them. There were 39 missing scores in the data set and that was an indication of either how slow the subjects were in responding to the images within 6 seconds, or how inefficient the method of data entry was, or both. Also, it could be an indication of loss of concentration during the course of the experiment.

Figure 4.2: Nominal and recorded cover for Experiment 1 with line of equality



4.3 Experiment 2

From the results of the pilot study, it was decided that some changes needed to be made to improve the experimental apparatus and procedure for Experiment 2. The main problem was thought to be the way in which scores were entered. The use of a keyboard was seen to be prone to errors, possibly because the subjects had to keep changing focus from the monitor to the keyboard and back. This was then changed to using a mouse pointer and dragging it along a bar representing a scale, 0 to 100, which indicated relative cover of the dots on the image. This bar appeared below each image and once the subject had decided on their score they left-clicked their mouse at the corresponding point to confirm the score.

In this experiment, a second task of estimating the number of dots was introduced. For this, scores were also entered by dragging a mouse pointer along the scale bar. So, numbers appeared on the bar and changed accordingly as the mouse pointer moved along it, and the response was confirmed by left-clicking the mouse. The same nominal levels as for the pilot study were used for the cover task, that is 5, 10, 17, 26, 37, 50 and 65%; for the count task, the nominal levels were 17, 27, 40, 60, 95, 135 and 200. The scale bar for count, on which the scores were entered, was graduated from 1 to 250. Unlike in the pilot study, the actual levels

generated by the program were within 10% of the nominal levels. For the images used in the count task, the dots on the white squares were disjoint, while those for the cover task were allowed to overlap. Also, the number of images viewed in the training sessions of this experiment was increased from 6 to 9.

The subjects were eight fourth-year Statistics students and 17 second-year Mathematics and Statistics students at the University of Edinburgh. They were aged between 18 and 24, and some of them had taken part in the pilot study. The incentive offered was three prizes worth £25, £15 and £10 for three participants with the lowest mean difference between their responses and the true levels, in both tasks.

The design used was a sequentially balanced design of Finney and Outhwaite (1956) as discussed in Chapter 3. Four sequences for $n = 7$ had been enumerated at the time of this experiment, before a program was written to search for all possible sequences. One of these four was then randomly selected for each task and each subject. The subjects carried out one task selected at random and had a 3 minute break before moving on to the second one which, of course, was preceded by its own training session.

4.3.1 The data

The result files from this experiment had columns for the true levels (that is, within 10% of the nominal levels), responses and nominal levels, unlike in the pilot study where only nominal levels and responses were stored. The true levels were recorded for use as explanatory variables in fitting models. So, here, nominal levels were used only for generating the design and for preliminary analysis described in the next subsections. There was a need to stabilise the variance of the responses, and this was done by choosing a suitable transformation for both the responses and true levels. The data recorded in the cover task are basically proportions, that is, there is an upper and a lower bound to the scale used to make judgement. Because of this, a logit transformation seemed a reasonable choice. This was tried out together with the logarithmic and square-root transformations and the logit transformation stabilised the variances most.

According to DeCarlo and Cross (1990), the log transformation of data obtained from estimation of loudness of sound tones, as well as area estimation of dark circles of varying sizes, seemed reasonable. They found that a plot on a log-log scale showed a linear increase in the log of the responses. They also observed some

autocorrelation in the errors of the responses. In the case of cover estimation in the visual assessment study, the log transformation did not stabilise the between-subject variances across all levels of cover. Instead, it decreased between-subject variation at high levels and increased it at lower levels. A logit transformation was used instead, because the data are basically proportions. For count estimation, the log transformation stabilised the variance well. The plots of responses versus true levels for both tasks and their respective transformations are given in Figure 4.3. As can be seen from these plots, there are very few outliers, and there were very few missing values too. This indicates that the new method of data entry was more efficient than the one used in the pilot study.

Assuming normality for the transformed data, the regression model (4.1) was fitted to the data for each task.

$$y_{it} = \tau_i + \beta_1 x_{it} + \beta_2 x_{it-1} + \beta_3 t + \gamma_{it} + \omega_{it,t-1} + \epsilon_{it}, \tag{4.1}$$

where y_{it} is the transformed response of subject i at position t of the sequence; x_{it} is the true level at position t with coefficient β_1; x_{it-1} is the true level at position $t-1$, which allows for a carry-over effect from the previous level, with coefficient β_2; the effect of the position in the sequence is given by t with coefficient β_3; γ_{it} denotes the interaction between assessor and true level; $\omega_{it,t-1}$ denotes the interaction between the current level and the immediately preceding one; and ϵ_{it} is the residual effect assumed to have a normal distribution with mean 0 and variance σ^2.

Tables 4.1 and 4.2 give respective analyses of variance results from the above model. For count estimation, all effects are significant at 1% level. The previous effect level in cover estimation is not significant, and there is a non-significant interaction between current and previous cover. The probability plots showed that the assumption of normality holds for the residual transformed data from both tasks. In order to check the independence assumption of errors, the partial auto-correlation function of the residuals was plotted: in both tasks, it turned out that auto-correlation at lag 1 is highly significant. Thus an auto-regressive model of order 1 AR(1), for the errors, would be suitable.

One may be interested in the effect that fitting the terms in the above anova model in different orders would have. This is a question of the degree of multicollinearity in the data, which is the extent to which regressors are correlated with each other. When there is multicollinearity, the order in which the terms are added matters

Figure 4.3: Plots of responses versus true levels for count and cover tasks and their corresponding transformations

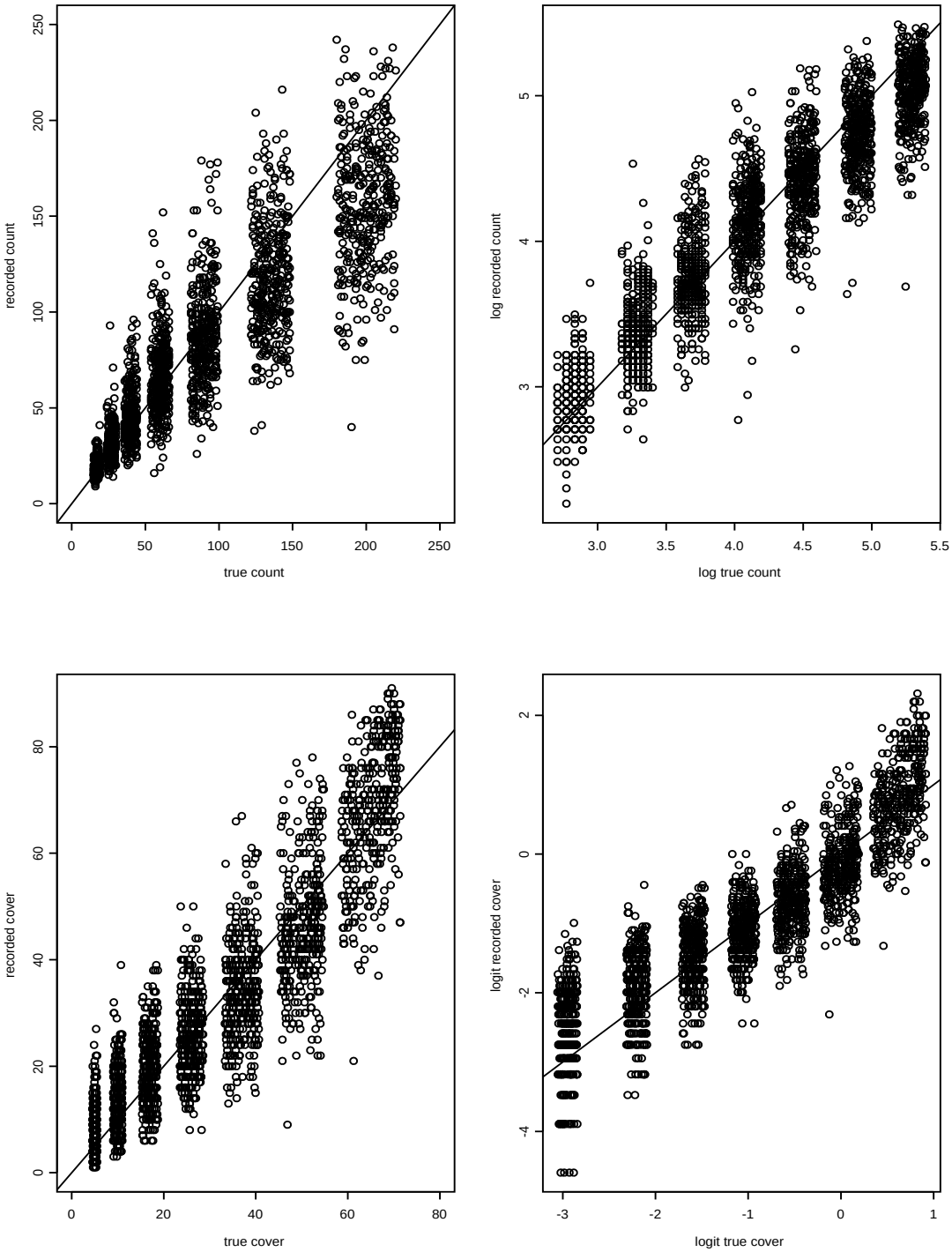


Table 4.1: Analysis of variance for logs of count data

Source	Df	Sum of sq	Mean sq	F-Value	Pr(F)
Subject	24	38.14	1.59	33.26	<0.0001
Current true level	1	1167.61	1167.61	24437.27	<0.0001
Previous level	1	0.44	0.44	9.12	0.0026
Image number	1	3.96	3.96	82.81	<0.0001
Subject × current level	24	7.65	0.32	6.67	<0.0001
Current level × previous level	1	2.05	2.05	42.89	<0.0001
Residuals	2379	113.67	0.05		

because the effect of each regressor is calculated relative to the effects of others already in the model. The visual assessment experiments were designed such that multicollinearity is minimised. This was done by using the Finney and Outhwaite (1956) designs where current and previous image levels are orthogonal to each other, and therefore the order in which these two are added to the model, relative to each other, does not matter. The arrangement of image levels in complete blocks gives a roughly uniform allocation of current and previous image levels throughout the sequence, thus these two terms are expected to be roughly orthogonal to the order effect. Finally, the subject effect is also made roughly orthogonal to the rest of the terms by a random selection of the sequences for each of the subjects. Thus, when the order of the addition of regressors is changed in this model, the sum of squares do not differ substantially. When comparing the sum of squares for each of these regressors when fitted individually, and when fitted together with the others, one can see that there is only a slight difference in these values. For the count data, individual sum of squares values for subject, current level, previous level and order are 38.14, 1167.20, 0.34 and 3.84, respectively, and for cover, these are 168.63, 2615.72, 0.11 and 6.19, respectively. These do not differ much from the regressors' estimates of sum of squares in the full models shown in tables 4.1 and 4.2. In both tasks, the previous level effect is seen to be nonsignificant when fitted on its own, and that makes sense because, as a carry-over effect in the model, it is only interpretable when fitted together with a current effect term.

In order to investigate the nature of carry-over as in Ferris et al. (2001), the mean bias for transformed data (*bias=response−true level*) was calculated for current nominal level by previous nominal level and given in Tables 4.3 and 4.4. These tables are such that carry-over is in the form of assimilation if, for a given row, all or most values of bias to the right of the diagonal are larger than the diagonal

Table 4.2: Analysis of variance for logits of cover data

Source	Df	Sum of Sq	Mean Sq	F-Value	Pr(F)
Subject	24	168.63	7.03	40.28	<0.0001
Current true level	1	2618.77	2618.77	15014.73	<0.0001
Previous level	1	0.49	0.49	2.81	0.094
Image number	1	6.78	6.78	38.85	<0.0001
Subject × current level	24	52.41	2.18	12.52	<0.0001
Current level × previous level	1	0.01	0.01	0.05	0.83
Residuals	2395	417.72	0.17		

Table 4.3: Mean bias for current by previous nominal levels for count estimation

	Previous nominal						
Current nominal	17	27	40	60	90	135	200
17	0.17	0.13	0.09	0.11	0.09	0.07	0.06
27	0.14	0.16	0.10	0.12	0.05	0.09	0.02
40	0.04	0.07	0.12	0.13	0.18	0.04	−0.01
60	−0.04	0.02	0.05	0.07	0.06	0.07	0.06
90	−0.17	−0.18	0.09	−0.05	0.01	−0.00	−0.04
135	−0.21	−0.19	−0.21	−0.16	−0.17	−0.11	0.08
200	−0.31	−0.27	−0.29	−0.24	−0.27	−0.22	−0.23

entry and those to left are lower. If, for a given row, values to the right of the diagonal are lower and those to the left are higher than the diagonal, then a contrastive effect is exhibited.

In the counts data, assimilation is mainly seen when the previous image is lower than the current one, that is bias values to the left of the diagonal are less than corresponding diagonal entries. This is referred to as negative assimilation. Otherwise, a contrastive effect is exhibited. For the cover data, there is no consistent

Table 4.4: Mean bias for current by previous nominal levels for cover estimation

	Previous Nominal						
Current Nominal	5	10	17	26	37	50	65
5	0.49	0.43	0.36	0.41	0.33	0.39	0.45
10	0.33	0.17	0.29	0.23	0.25	0.33	0.32
17	0.15	0.09	0.14	0.10	0.09	0.17	0.21
26	−0.16	−0.05	0.05	−0.04	−0.03	−0.08	0.02
37	−0.21	−0.12	−0.21	−0.13	−0.17	−0.12	−0.02
50	−0.23	−0.23	−0.12	−0.17	−0.19	−0.21	−0.03
65	0.17	0.24	0.23	0.25	0.19	0.07	0.12

assimilative or contrastive effect. These results differ from those of Ferris et al. (2001) where assimilative effects were observed.

4.3.2 Parametric model for calibration

As mentioned before, the aim is to be able to carry out calibration for individual assessors, assuming a particular model for the responses. Based on the results of the experiments above, a parametric regression model similar to (4.1) was assumed for the transformed data from both tasks. This was simplified by disregarding the interaction terms, while the errors were assumed to be correlated. Thus, the transformed response, y, is assumed to have an expectation which is a linear function of the transformed current level, and transformed previous level, which is the carry-over term in the model. The position of an image in a sequence was seen to have an effect on the response and it is therefore included in this model as well. This could be seen as an effect of fatigue as assessors perform the task repeatedly, or maybe a learning effect. The errors were assumed to be correlated following an AR(1) process. In order to reduce the instability that might be caused by collinearity in the predictors (Hocking (2003), Chapter5), the intercept is assumed to be the overall expected level of cover or count, and the predictors are centred about their means. Hence, the model used for calibration is given as follows.

$$y_t = \beta_0 + \beta_1(x_t - \overline{x}) + \beta_2(x_{t-1} - \overline{x}) + \beta_3(t - \bar{t}) + \varepsilon_t \tag{4.2}$$

with

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t, t = 1, ..., N \tag{4.3}$$

where the subscript t denotes the position of an image level in the sequence, x_t denotes stimulus level at position t, ρ is the autocorrelation parameter and u_ts are normally and independently distributed, that is $NID(0, \sigma^2\ )$.

4.4 Calibration

Osborne (1991) gave a broad review of statistical calibration as applied in various contexts. Definitions of two kinds of calibration were also given, namely absolute calibration and comparative calibration. Absolute calibration is the one where true or correct standard measures are available in order to correct the scores, whereas in comparative calibration, such standard measures do not exist and

performance of each measuring instrument is measured relative to the others. Absolute calibration is discussed in this chapter, while comparative calibration and a corresponding model will be seen in the next chapter which involves food tasting studies. Absolute calibration used here is similar to scientific laboratory calibration where two types of measuring instruments may be available and one of them is precise but costly and slow or impractical to use, and the other one is quicker and easy to use, but less accurate. So, test measurements are made in order to evaluate the extent of bias and variability exhibited by the latter instrument. Then, information from the accurate instrument is used to correct the less accurate measurements. So, the easy-to-use instrument can then be used in future day to day measurements because information on how to correct its scores for biases would be available.

Aitchison and Dunsmore (1975) gave introductory examples of laboratory calibration and then discussed a Bayesian predictive method of carrying out statistical calibration. Predictive calibration proceeds in such a way that a measuring instrument, which in this case is a human assessor, takes part in a test experiment which closely resembles the actual experiment to be undertaken in future. A parametric regression model is then fitted to model the relationship between response and true stimulus. Assuming that biases stay fairly consistent over a particular period of time, parameter estimates from this model are used to correct biases in subsequent experiments.

The test experiment, often referred to as the *calibration experiment*, results in responses y_t, in a vector $\mathbf{y}$, to the true stimuli x_t, in vector $\mathbf{x}$, so that the data are denoted by $\mathbf{z} = (\mathbf{x}, \mathbf{y})$. When a measurement is made in future, the response, denoted by y_*, is given for some unknown stimulus x_*. The aim then is to infer the value of x_* assuming that (x_*,y_*) follow the same probability distribution as the elements of $(\mathbf{x}, \mathbf{y})$ with parameter vector θ. For the visual assessment experiments, calibration as discussed by Aitchison and Dunsmore (1975) was generalised to apply to the future response and stimulus as vectors, instead of scalars. This is because the model assumes responses have carry-over effects and auto-correlation in the errors, so it would not be appropriate to do calibration for individual future scores.

A frequentist approach to this would proceed as an inverse regression. Model (4.2) would be fitted to the data from the test experiment using the generalised least squares method, in order to obtain the estimates $\hat{\theta} = (\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\rho}, \hat{\sigma})$.

The expectation of the future response y_{*t}, assuming that the two experiments have the same number of responses, is then given by

$$E(\hat{y}_{*t}) = \hat{\beta}_0 + \hat{\beta}_1(x_{*t} - \overline{x}_*) + \hat{\beta}_2(x_{*t-1} - \overline{x}_*) + \hat{\beta}_3(t - \overline{t});\ t = 1, ..., N,$$

where N is the total number of responses in a sequence. In order to estimate a vector of future stimuli $\mathbf{x}_*$, one would solve a system of N linear equations

$$\lambda_t = E(y_{*t}) - \hat{\beta}_0 = \hat{\beta}_1 x_{*t} + \hat{\beta}_2 x_{*t-1} + \hat{\beta}_3 t.$$

The problem with this approach is that there are $N + 1$ that need to be solved for. Also, when some information about the possible mean value and precision of the future stimulus is available, it needs to be incorporated in the estimation of x_* (Aitchison and Dunsmore (1975), Section 10.7). A Bayesian approach was therefore followed for the visual assessment study. Generalisation of this to a vector of future stimuli was made possible by the availability of Markov chain Monte Carlo (MCMC) methods which allow for sampling from intractable probability distribution functions.

Bayesian calibration may be seen as a form of inverse regression where the predictive distribution plays a key role. The essence of the predictive distribution was highlighted by Roberts (1965) and Geisser (1971). Aitchison (1975) showed that for samples of finite size, the goodness of fit criterion favours a parametric model which is fitted using the predictive approach, to the one fitted using the estimative method. In the estimative approach, unknown parameters are replaced by some efficient estimates such as maximum likelihood estimates.

A posterior distribution of the parameter vector θ, after observing data $(\mathbf{x}, \mathbf{y})$ from the calibration experiment is given through Bayes theorem as

$$p(\theta|\mathbf{x}, \mathbf{y}) = \frac{p(\theta)p(\mathbf{x}, \mathbf{y}|\theta)}{p(\mathbf{x}, \mathbf{y})}, \tag{4.4}$$

where $p(.)$ denotes a probability density, and the probability density for the future response vector $\mathbf{y}_*$ is given by

$$p(\mathbf{y}_*|\mathbf{z}, \mathbf{x}_*) = \int p(\mathbf{y}_*|\mathbf{x}_*, \theta)p(\theta|\mathbf{x}, \mathbf{y})d\theta. \tag{4.5}$$

This is called the predictive distribution of $\mathbf{y}_*$.

From this, a probability distribution for inferring $\mathbf{x}_*$, called the calibrative distribution can be defined. First, a prior distribution for $\mathbf{x}_*$ is defined. This depends

on the kind of calibrative experiment used. Aitchison and Dunsmore (1975) defined two kinds of calibration experiments, namely, *natural* and *designed* (or *controlled*). A natural experiment is one in which the explanatory variables x occur naturally, for example, soil water content. In this case, the prior for x_* is assumed to depend on some parameter ψ, which determines the generation of x, and also on θ. A designed calibration experiment is one in which the values of the explanatory variable are specifically chosen to possibly cover the range of future explanatory variables. In this case, there is no information available about the generation of x and thus, the prior for x_* is independent of any parameter and is denoted by $p(x_*)$. The visual assessment calibration experiment is a designed one and thus, the joint density of $\mathbf{x}_*$ and $\mathbf{y}_*$ is given as $p(\mathbf{x}_*, \mathbf{y}_*|\theta) = p(\mathbf{x}_*)p(\mathbf{y}_*|\mathbf{x}_*, \theta)$. The calibrative density is then derived thus

$$\begin{aligned} p(\mathbf{x}_*, \theta|\mathbf{y}_*, \mathbf{x}, \mathbf{y}) &\propto p(\mathbf{x}_*)p(\theta)p(\mathbf{y}_*|\mathbf{x}_*, \theta)p(\mathbf{y}|\mathbf{x}, \theta) \\ &\propto p(\mathbf{x}_*)p(\mathbf{y}_*|\mathbf{x}_*, \theta)p(\theta|\mathbf{x}, \mathbf{y}), \end{aligned}$$

and integrating out θ gives

$$p(\mathbf{x}_*|\mathbf{y}_*, \mathbf{y}, \mathbf{x}) \propto p(\mathbf{x}_*)p(\mathbf{y}_*|\mathbf{x}_*, \mathbf{x}, \mathbf{y}), \tag{4.6}$$

where $p(\mathbf{y}_*|\mathbf{x}_*, \mathbf{x}, \mathbf{y})$ is the predictive distribution of the future response. It should be kept in mind that Model (4.2) includes the covariate t, and so, these probability distributions also depend on t.

The main difficulty encountered when doing calibration for the visual assessments was with the choice of a suitable prior distribution for $\mathbf{x}_*$. In practice, the prior distribution for $\mathbf{x}_*$ is influenced by the context and area of application, and so does the range of stimulus chosen for a designed calibration experiment. Prior distributions for severity of plant disease may be determined by the plant pathologists' observation of disease severity during the current and previous seasons. In this case a Normal prior was assumed with the mean equal to that of $\mathbf{x}$ in the calibration experiment, and the standard error as double that of $\mathbf{x}$ in the calibration experiment, to allow for the possibility of high variance in future scores.

Normal prior distributions were assumed for the coefficients $\beta_0, \beta_1, \beta_2$ and β_3. Two possibilities for the values of their prior expectations were considered. First, it may be assumed *a-priori*, for Model (4.2), that an assessor is unbiased: that there are no carry-over or order effects. The prior expectations would then be

Table 4.5: Prior 1 distributions for the parameters in Model (4.2)

Parameter	Cover task	Count task
Expected response	$\beta_0 \sim N(-1.093, 2)$	$\beta_0 \sim N(4.085, 2)$
Coefficient for current level	$\beta_1 \sim N(1, 2)$	$\beta_1 \sim N(1, 2)$
Coefficient for previous level	$\beta_2 \sim N(0, 5)$	$\beta_2 \sim N(0, 5)$
Coefficient for seq. position	$\beta_3 \sim N(0, 5)$	$\beta_3 \sim N(0, 5)$
Standard deviation	$20 \times 1.33\sigma^{-2} \sim \chi^2(20)$	$20 \times 0.55\sigma^{-2} \sim \chi^2(20)$
Autocorrelation parameter	$\frac{1}{2}(\rho + 1) \sim Be(5, 5)$	$\frac{1}{2}(\rho + 1) \sim Be(5, 5)$
Future stimulus	$x_* \sim N(-1.09, 0.74)$	$x_* \sim N(4.09, 1.49)$

Table 4.6: Prior 2 distributions for the parameters in Model (4.2)

Parameter	Cover task	Count task
Expected response	$\beta_0 \sim N(-1.008, 0.79 \times 10^{-4})$	$\beta_0 \sim N(4.062, 0.21 \times 10^{-4})$
Coef. for current level	$\beta_1 \sim N(0.888, 0.59 \times 10^{-4})$	$\beta_1 \sim N(0.846, 0.31 \times 10^{-4})$
Coef. for previous level	$\beta_2 \sim N(0.012, 0.59 \times 10^{-4})$	$\beta_2 \sim N(0.017, 0.31 \times 10^{-4})$
Coef. for seq. position	$\beta_3 \sim N(0.002, 0.9 \times 10^{-7})$	$\beta_3 \sim N(-0.001, 0.4 \times 10^{7})$
Standard deviation	$20 \times 1.33\sigma^{-2} \sim \chi^2(20)$	$20 \times 0.55\sigma^{-2} \sim \chi^2(20)$
Autocorrelation	$\frac{1}{2}(\rho + 1) \sim Be(5, 5)$	$\frac{1}{2}(\rho + 1) \sim Be(5, 5)$
Future stimulus	$x_* \sim N(-1.09, 0.74)$	$x_* \sim N(4.09, 1.49)$

$\beta_0 = \overline{x}, \beta_1 = 1, \beta_2 = 0, \beta_3 = 0$. This is shown in Table 4.5 and will be referred to as Prior 1. Another option would be to assume prior expectations and variances of the coefficients from their estimate after fitting Model (4.2) to the data from Experiment 2. These priors are shown in Table 4.6 and will be referred to as Prior 2. In both cases, the correlation parameter, ρ, has a beta prior, centred around 0 because very little auto-correlation was exhibited in Experiment 2 and the same was expected here. It is assumed that, like in the design of the test experiment, x_{*0} is the same as x_{*1}. The prior distribution for σ^2 was assumed to be an inverse gamma which is equivalent to $ds^2\sigma^{-2} \sim \chi^2(d)$ where s is the estimate of the sample variance and d is the corresponding number of degrees of freedom. The values of these are taken from the analysis of data from Experiment 2 as well.

A program called WinBUGS (Spiegelhalter et al. (1996), Section 2.3) was used to fit the Bayesian regression model, sampling from the full conditional predictive and calibrative distributions. This is a Gibbs sampler program, with some Markov chain Monte Carlo sampling methods, that is available free from *http://www.mrc-bsu.cam.ac.uk/bugs*. In order to illustrate calibration in visual assessments, two sets of data were required from each assessor. That is, an experiment was run as

Table 4.7: Assessor mean square errors for count data

		Calibrated response		
	Recorded	Model (4.2)	Model (4.2)	Model (4.7)
Assessor	response	Prior 1	Prior 2	Prior 1
Alex	6.77	6.56	7.52	6.75
Alexander	7.73	6.94	7.83	7.04
Ayona	10.02	8.19	9.11	8.21
Isthri	5.44	5.18	5.28	5.22
Tumi	9.36	6.72	6.61	6.93

a calibration experiment and the parametric model (4.2) fitted, with the above prior distributions, to assess bias and variability. A similar experiment was run again to obtain responses for which true stimulus levels were to be inferred. In other words, the second run of the experiment represented a future performance of an assessor in a field of work. Five assessors, one of whom had taken part in Experiment 2, took part in this and they did the calibration and the second experiment one week apart. The resultant posterior means of the future stimuli were obtained from samples of 10000 iterations after a burn-in of 5000 for each assessor.

4.4.1 Calibration results

It was possible to examine whether there was an improvement in the responses after calibration, because the program used for the visual assessment experiments recorded true count and cover levels. The mean square error is defined as the mean of the squared differences between the transformed response and the transformed true level. So, here, mean square errors for recorded responses and calibrated responses were calculated in order to see if calibration has resulted in an improvement of the responses. Calibrated responses are taken to be the posterior means of $\mathbf{x}_*$, sampled from the calibrative distribution $p(\mathbf{x}_*|\mathbf{y}_*, \mathbf{y}, \mathbf{x})$ by Winbugs. The values of the mean square errors for these calibrated responses under Model (4.2) under the two priors are given in the second and third columns of Tables 4.7 and 4.8 for count and cover tasks, respectively (the values in the fourth column will be discussed later).

When comparing values of mean square errors for the calibrated responses to the recorded ones, it can be seen that for the count task there is a reduction in the mean square error due to calibration for each assessors under Prior 1, although it is slight for three of them. Under Prior 2, calibration resulted in reduction of

Table 4.8: Assessor mean square errors for cover data

		Calibrated response		
	Recorded	Model (4.2)	Model (4.2)	Model (4.7)
Assessor	response	Prior 1	Prior 2	Prior 1
Alex	28.64	17.35	20.216	18.31
Alexander	26.10	29.28	28.22	28.68
Ayona	81.36	28.92	34.65	27.06
Isthri	21.96	23.84	17.51	20.92
Tumi	83.42	24.59	58.47	23.06

mean square error for only 3 of the assessors. For the cover task however, there is a substantial reduction in the mean square error for 3 out of the 5 assessors under Prior 1. Under Prior 2 the reduction for four out of the 5 assessors is not as substantial as for Prior 1. So, Prior 1 seems to perform better than Prior 2.

The two assessors Alexander and Isthri showed a very interesting characteristic of the way this calibration method performs. This is clearly seen in the plots of the data. These are in Figures 4.4 and 4.5, where columns correspond to plots of test data, future experiment data and calibrated response data; and rows correspond the five assessors: Alex, Alexander, Ayona, Isthri and Tumi, in that order.

The interesting characteristic seen in the cover task is that, for Alexander and Isthri (rows 2 and 4 in Table 4.5) in particular, test data shows that lower levels are significantly biased in the opposite direction of the biases of the same levels in the future experiment. Thus, for those levels, calibration results in correction of the scores in a wrong direction. In other words, the direction of bias in the future experiment is assumed to be the same as it was in the test experiment, but because this is not true, this method resulted in even more bias.

The calibration procedure was tested for robustness to changes in the prior distribution of the future stimulus level x_*. When the prior variance of x_* was doubled, the calibration did not at all improve the responses and in some cases it worsened them. On the other hand, when the prior variance was reduced, the calibrated responses were drawn towards the mean, hence introducing bias, particularly for the bottom and top levels of the stimulus.

Another test of the calibration procedure was done by changing the regression model assumed for calibration. The carry-over from the previous stimulus level and order effects were removed, assuming only that errors were correlated, following an AR(1) model. This resulted in the model:

Figure 4.4: Plots of logs of responses versus log of true count; columns correspond to test, future and calibrated data; rows correspond to assessors in alphabetical order of their names

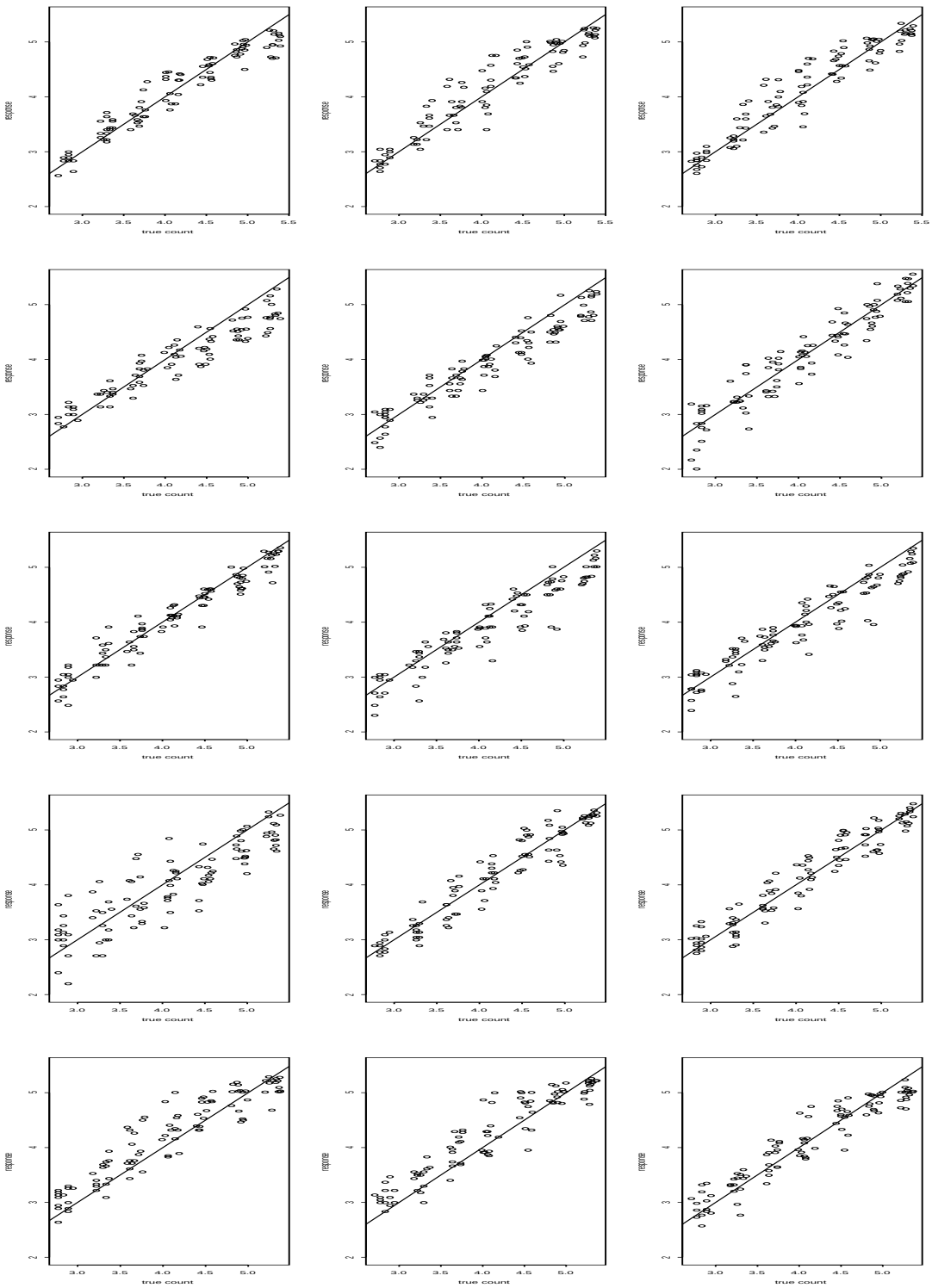
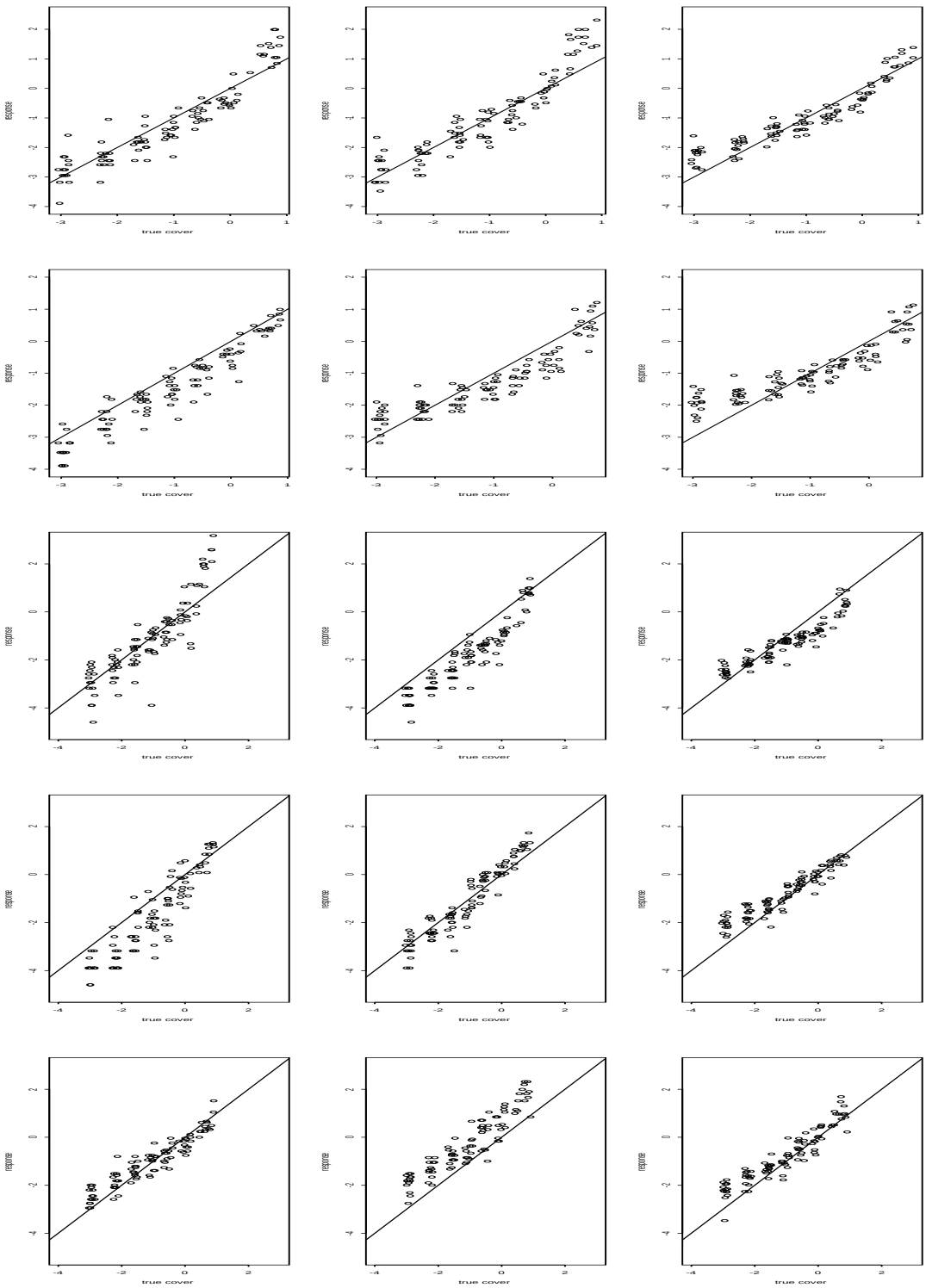


Figure 4.5: Plots of logits of responses versus logits of true cover; columns correspond to test, future and calibrated data; rows correspond to assessors in alphabetical order of their names



$$y_t = \beta_0 + \beta_1(x_t - \overline{x}) + \varepsilon_t, \tag{4.7}$$

with

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t. \tag{4.8}$$

This model was fitted using Prior 1 because this prior was seen to perform better than Prior 2. The mean square errors of the calibrated responses under this model are shown in the fourth columns of Tables 4.7 and 4.8. The result of this change in the model for calibration reflects what was shown by the analysis of variance for the two tasks in Experiment 2. For the count task, analysis of variance showed carry-over from the previous image level to be significant. Thus, calibration under a model without this carry-over term does not improve the responses as well as calibration under the model with the carry-over term. Hence the mean square errors of calibrated responses under Model (4.7) are higher than those under Model (4.2).

On the other hand, for the cover task, carry-over from the previous image level was not significant, as shown by the analysis of variance results. Thus, calibration under a model without this carry-over term improves the responses (for 4 out of the 5 subjects) more than calibration under the model with the carry-over term. Hence the mean square errors of calibrated responses under Model (4.7) are lower than those under Model (4.2).

4.5 Selecting the Best Assessors

It is sometimes of interest to select individuals who perform well among a panel of assessors. Spezzaferri (1985) proposed a criterion for choosing the best instruments among a set of measuring instruments, based on calibration experiments. Shannon information theory is used in this criterion in order to measure how much information an instrument's predictive distribution of a future response y_* gives about the true future stimulus x_*. Shannon and Weaver (1963) first defined an entropy of a distribution, which is a measure of its uncertainty, for a continuous random variable x with probability density $p(x)$, as

$$H = -\int_{\infty}^{\infty} p(x)\log p(x)dx. \tag{4.9}$$

Kullback and Leibler (1951) generalised this idea of information such that for $p_1(x)$ and $p_2(x)$ related to some hypotheses 1 and 2, respectively,

$$\log \frac{p_1(x)}{p_2(x)} \tag{4.10}$$

is the information in x for discrimination between hypotheses 1 and 2 (also discussed in Lindley (1956)). Roberts (1965) and Geisser (1971) both emphasised the usefulness of the information contained in the predictive distribution. This forms the basis of the criterion: if an assessor is good, their calibrative density $p(\mathbf{x}_*|\mathbf{x}, \mathbf{y}, \mathbf{y}_*)$, which depends on $\mathbf{y}_*$, provides a lot of information about the likely value of $\mathbf{x}_*$, relative to the prior $p(\mathbf{x}_*)$ which is independent of y_*. The calibration experiment provides data $\mathbf{z} = (\mathbf{x}, \mathbf{y})$, and the joint predictive distribution of $\mathbf{x}_*$ and $\mathbf{y}_*$ is given by

$$p(\mathbf{x}_*, \mathbf{y}_*|\mathbf{x}, \mathbf{y}) = p(\mathbf{y}_*|\mathbf{x}, \mathbf{y})p(\mathbf{x}_*|\mathbf{x}, \mathbf{y}, \mathbf{y}_*), \tag{4.11}$$

where $p(\mathbf{x}_*|\mathbf{x}, \mathbf{y}, \mathbf{y}_*)$ is the calibrative density, which is a posterior density of $\mathbf{x}_*$ after observing $\mathbf{x}, \mathbf{y}, \mathbf{y}_*$.

The measure of information for an assessor, as given by Spezzaferri (1985), is thus

$$I = \int\int p(\mathbf{y}_*|\mathbf{x}, \mathbf{y})p(\mathbf{x}_*|\mathbf{x}, \mathbf{y}, \mathbf{y}_*) \log \frac{p(\mathbf{x}_*|\mathbf{x}, \mathbf{y}, \mathbf{y}_*)}{p(\mathbf{x}_*)} d\mathbf{x}_* d\mathbf{y}_*. \tag{4.12}$$

So, an individual with a high value of I is regarded as a good assessor. In order to be able to evaluate I, it is simplified by first realizing that since the visual assessment experiments are a designed type of experiments, $p(\mathbf{x}_*|\mathbf{x}, \mathbf{y})$ is equal to $p(\mathbf{x}_*)$, so that

$$p(\mathbf{y}_*|\mathbf{x}, \mathbf{y})p(\mathbf{x}_*|\mathbf{x}, \mathbf{y}, \mathbf{y}_*) = p(\mathbf{y}_*, \mathbf{x}_*|\mathbf{x}, \mathbf{y}, \mathbf{y}_*) = p(\mathbf{y}_*|\mathbf{x}, \mathbf{y}, \mathbf{x}_*)p(\mathbf{x}_*).$$

From this, it can be deduced that

$$\frac{p(\mathbf{x}_*|\mathbf{x}, \mathbf{y}, \mathbf{y}_*)}{p(\mathbf{x}_*)} = \frac{p(\mathbf{y}_*|\mathbf{x}, \mathbf{y}, \mathbf{x}_*)}{p(\mathbf{y}_*|\mathbf{x}, \mathbf{y})}$$

and thus (4.12) becomes

$$\int\int p(\mathbf{y}_*, \mathbf{x}_*|\mathbf{x}, \mathbf{y}) \log \frac{p(\mathbf{y}_*|\mathbf{x}, \mathbf{y}, \mathbf{x}_*)}{p(\mathbf{y}_*|\mathbf{x}, \mathbf{y})} d\mathbf{x}_* d\mathbf{y}_*$$

which is equivalent to

$$\int\int p(\mathbf{y}_*|\mathbf{x}, \mathbf{y}, \mathbf{x}_*)p(\mathbf{x}_*) \log p(\mathbf{y}_*|\mathbf{x}, \mathbf{y}, \mathbf{x}_*) d\mathbf{x}_* d\mathbf{y}_* - \int\int p(\mathbf{y}_*, \mathbf{x}_*|\mathbf{x}, \mathbf{y}) \log p(\mathbf{y}_*|\mathbf{x}, \mathbf{y}) d\mathbf{x}_* d\mathbf{y}_*.$$

This simplifies to

$$\int p(\mathbf{x}_*) \int p(\mathbf{y}_*|\mathbf{x},\mathbf{y},\mathbf{x}_*) \log p(\mathbf{y}_*|\mathbf{x},\mathbf{y},\mathbf{x}_*) d\mathbf{x}_* d\mathbf{y}_* - \int p(\mathbf{y}_*|\mathbf{x},\mathbf{y}) \log p(\mathbf{y}_*|\mathbf{x},\mathbf{y}) d\mathbf{y}_*. \tag{4.13}$$

For the case of univariate Normal linear regression with suitable prior distributions, $p(\mathbf{y}_*|\mathbf{x},\mathbf{y},\mathbf{x}_*)$ was shown to be a Student density by Aitchison and Dunsmore (1975). In this project though, (4.13) has been evaluated by approximating the distributions of $\mathbf{y}_*$ given $\mathbf{x},\mathbf{y},\mathbf{x}_*$ and $\mathbf{y}_*$ given only $\mathbf{x},\mathbf{y}$ using Normal densities. Thus the first term in (4.13) becomes

$$-\frac{q}{2}(\ln(2\pi)+1) - \frac{1}{2}\int p(\mathbf{x}_*) \ln\det \mathrm{V}(\mathbf{y}_*|\mathbf{x},\mathbf{y},\mathbf{x}_*) d\mathbf{x}_*,$$

and the second term becomes

$$\frac{q}{2}(\ln(2\pi)+1) + \frac{1}{2}\ln\det \mathrm{V}(\mathbf{y}_*|\mathbf{x},\mathbf{y})$$

where V denotes the variance matrix, and q is the number of random variables y_* that the normal pdf approximates. Thus, (4.13) becomes

$$\frac{1}{2}\left[\ln\det \mathrm{V}(\mathbf{y}_*|\mathbf{x},\mathbf{y}) - \int p(\mathbf{x}_*) \ln\det \mathrm{V}(\mathbf{y}_*|\mathbf{x},\mathbf{y},\mathbf{x}_*) d\mathbf{x}_*\right]. \tag{4.14}$$

It is then possible to approximate (4.14) using Markov chain Monte Carlo methods. For the visual assessments model, this was restricted to just the first two values of the stimulus x_{0*} and x_{1*}, instead of a larger vector. In particular, a set of standard normal quantiles of $p(x_*)$ was used to generate y_{1*} from the predictive distribution given $\mathbf{x}$ and$\mathbf{y}$. Fixed quantiles were used for illustration here instead of taking x_{0*} and x_{1*} from the normal prior distribution defined for model (4.2). This is because with a number of fixed quantiles, only a small covariance matrix needs to be evaluated and this does not require as many MCMC iterations as for random variables from a Normal distribution. The values of the information criterion (4.14) for the five assessors and each of the tasks, using each of the two priors, are given in Table 4.9

The ordering of assessors with respect to their values of the information criterion for count, under both priors, is Ayona, Alex, Alexander, Tumi and Isthri, in descending order of performance. For cover estimation task, under both priors, the descending order of performance is Tumi, Alex, Alexander, Ayona and Isthri. When comparing this ordering to the plots of the test data in the first columns of Figures 4.4 and 4.5, one can see that for both tasks, an assessor who is ranked the

Table 4.9: Values of the information criterion (4.14)

	Count task		Cover task	
Assessor	Prior 1	Prior 2	Prior 1	Prior 2
Alex	1.115	1.436	0.243	0.596
Alexander	1.098	1.407	0.219	0.552
Ayona	1.202	1.456	0.192	0.359
Isthri	0.707	0.914	0.118	0.249
Tumi	0.929	1.035	0.332	0.903

best is the one whose scores show most consistent bias (the difference between the response and the true level) and consistent variance (random variation in scores of the same nominal level) across all levels of the stimulus. The criterion under the two priors gave the same ordering of assessors.

If a frequentist approach was followed to select the best assessor, the value of R-squared would be used, and this is obtained from fitting a least squares regression model to the data. Alternatively, to account for auto-correlation in the data, a generalised linear model may be fitted and the value of the deviance would be used to choose the best assessor. It was of interest to compare the information criterion measure discussed above to these frequentist measures. This was done using count data from the 25 assessors who took part in Experiment 2. The Bayesian model used was (4.2): the one used for the 5 assessors who took part in the calibration study. It was fitted based on Prior 1, but the prior variances in this case were made more vague by halving the degrees of freedom.

Figure 4.6 shows the plot of the 25 assessors' information criterion values for the count task, against their R-squared values; Figure 4.7 shows the plot of the information criterion values against their deviance values; and Figure 4.8 shows the plots of logs of the responses versus logs of true stimulus levels for each assessor. The values of these criteria agree with the plots of the data as one can see that the assessors who are ranked the best show very little bias and variability in their responses. There is a very close relationship between the information criterion and the two frequentist criteria of R-squared and deviance. Assessor 19 is ranked the best while assessor 12 is the worst.

From (4.14), the first term in brackets may be regarded as the natural log of the residual sum of squares, while the second part is the natural log of the total sum of squares. Thus it may be seen that the information criterion is approximately

Figure 4.6: Information criterion versus R-squared

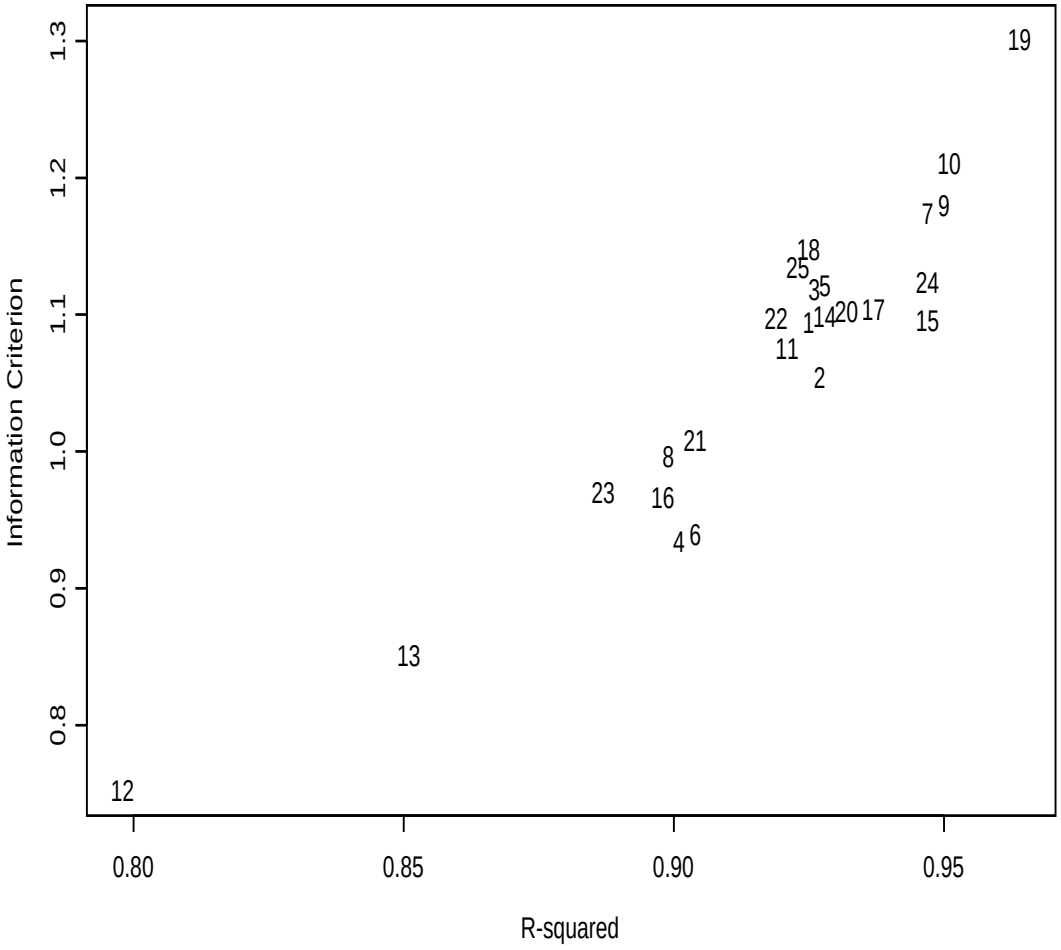


Figure 4.7: Information criterion versus deviance

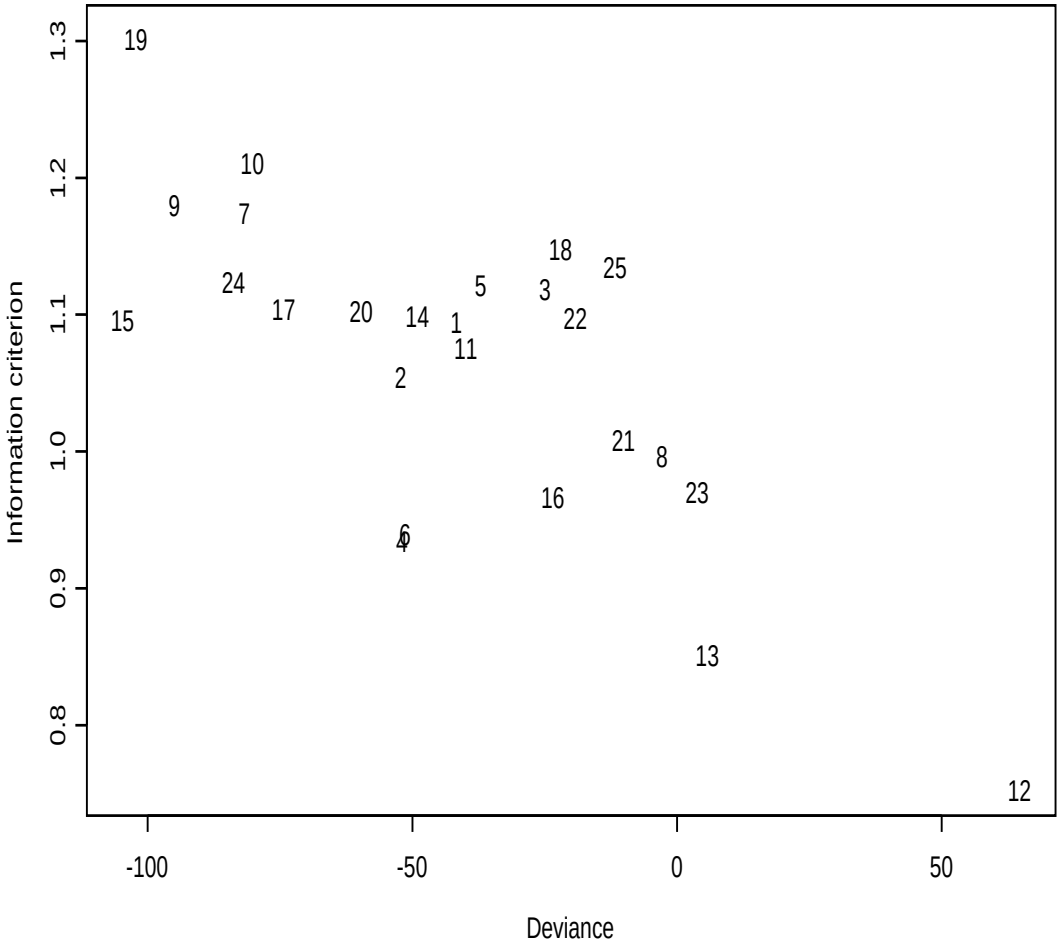
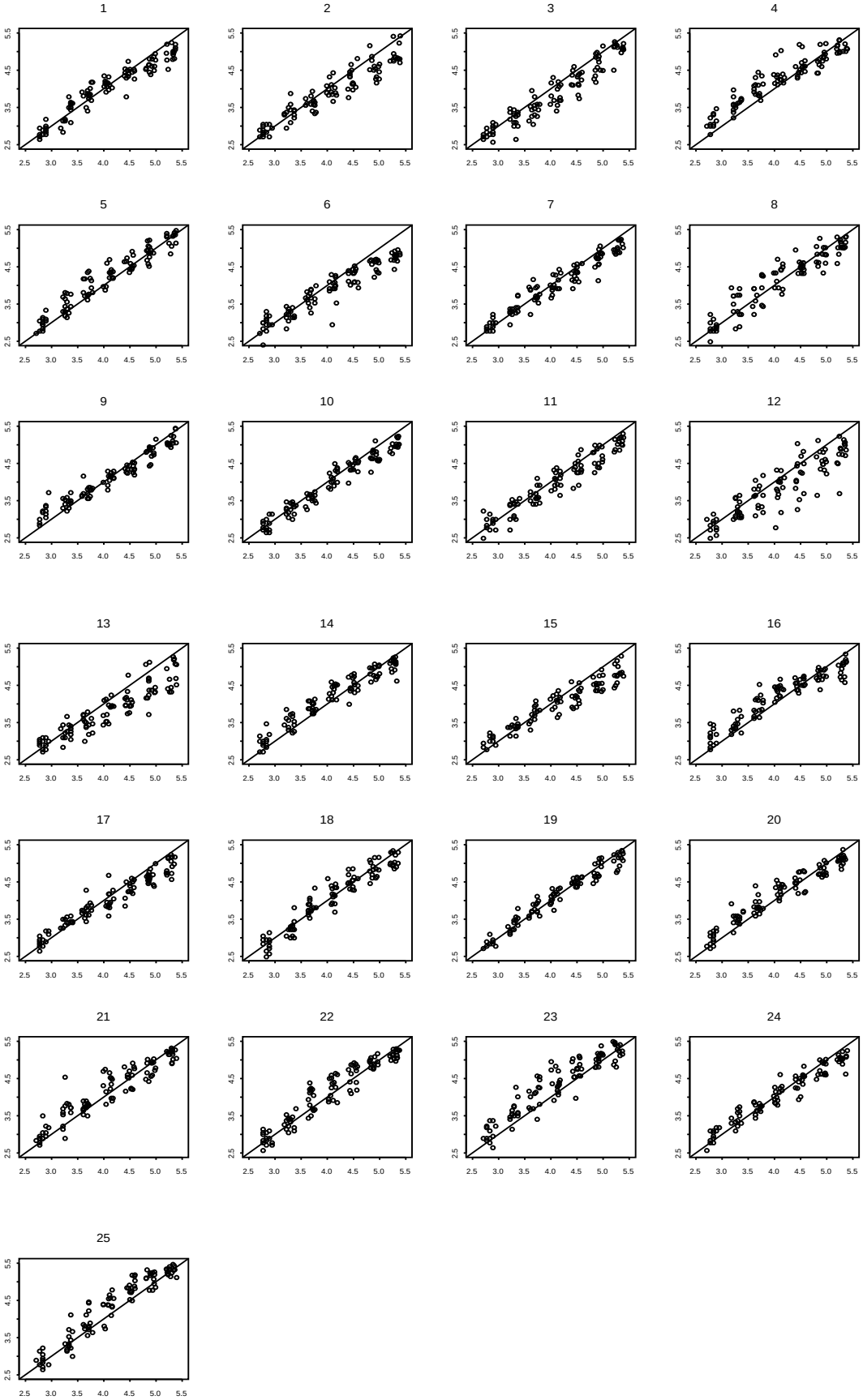


Figure 4.8: Plots of log responses versus log true count for Experiment 2



$$-0.5ln(1 - \frac{residualSS}{totalSS}) \approx -0.5ln(Rsquared). \quad (4.15)$$

Hence plotting the information criterion versus $-0.5ln(Rsquared)$ showed a perfectly linear relationship.

4.6 Summary

The computer aided visual assessment experiments carried out here were a great improvement from the way they had been done by Ferris et al. (2001). Using personal computers to run the experiment was shown to have a lot of advantages, and conducting a pilot study made it possible to identify some of the weaknesses of the program. Thus, the improvement in the way scores were entered reduced outliers and missing values substantially. The Bayesian predictive calibration presented by Aitchison and Dunsmore (1975) was successfully generalised to calibration for a vector of scores with auto-regressive errors, and this was possible due to the availability of Bayesian MCMC methods. The calibration does indeed improve the scores, but it performs well when biases are consistent between the test and subsequent performances. The assumption of consistent bias between the two experiments, when this is not the case, is one drawback of the method, but bias might become more consistent if assessors are tested repeatedly. The information criterion for selecting the best assessors gives more credit to assessors who are reliable, in a sense that their biases stay consistent throughout all levels, than to inconsistent assessors.

Chapter 5

Analysis of Food Tasting Studies

5.1 Sensory Evaluation Studies

The food industry is increasingly involved in food tasting experiments for reasons such as establishing consumer preferences of different food and drink varieties on the market. Usually, series of food tasting experiments are conducted with a panel of trained assessors who give their scores or rankings on food products. These scores are made with respect to certain attributes. Choosing and training panel members are some of the important issues in these studies. Piggot and Hunter (1999) gave a summary of the issues to be considered when recruiting, selecting and training assessors, and the main rule is that assessors have to be reliable, consistent, and have discriminating power. When assessors have been chosen, their performance may then be evaluated regularly over time. This chapter discusses a model for assessor performance, used in the analysis of the apple tasting data described below.

5.2 Apple Data

The data examined in this chapter were provided by the Hannah Research Institute in Ayr. Apple tasting is one of the many food tasting experiments conducted there. The aim was to monitor assessor performance over time, and thus a series of apple tasting experiments was conducted over several years. Datasets from 45 of the experiments spanning the period March 1996 to April 1998 were available for analysis. For these studies, 14 trained assessors were involved, but not all of them were present at every experiment. An incidence matrix for experiments×assessors, showing attendance is given in Appendix B with ones indicating presence and spaces indicating absence. As few as 4 and as many as 11 distinct apple varieties were used in each experiment and their choice depended

on the availability on the local market at the particular time.

There is some information that was either not recorded at all or not consistently recorded about the apple varieties throughout the study period. This includes some of the information about varieties' country of origin, method of packaging (either loose or in a bag) and presentation (peeled or not peeled). For ease of analysis, this problem was simplified by referring to varieties within experiments as *products* and treating them as unique to every experiment. This resulted in up to 11 unique products per experiment. Thus, one variety could correspond to two or more products within an experiment.

The design followed for presenting apples to assessors was such that 8 different apple variates were tasted per experiment. These were split into 2 groups of 4 varieties each, and these were referred to as sessions. The ordering of apples in each session was randomised. Furthermore, this design was replicated 3 times. Thus each assessor would give 24 responses in total. For some data sets, though, the design was not strictly followed during the experiment, and there were some imbalances such as one session per replicate. The length of break-time between sessions and replicates was not reported.

Each assessor was allocated a private booth in which to do the tasting, and scores were entered by dragging a mouse pointer along a scale bar on a computer screen to indicate ones score on a scale from 0 to 100, for each attribute. Each product was scored on 11 attributes, namely, fruity, sweet, acidic, bitter, presence of perfumed smell, floral sensation, after-taste, persistence, hard, crunchy and overall acceptability of the product. Analysis in this project was done only on the sweetness score to illustrate the statistical techniques proposed. The reason for this choice of attribute is that the scores associated with it did not have a very high proportion of zeros, hence less skewed data than for other attributes

We do combined analyses of these experiments in order to

- Model assessor performance across experiments.
- Use a predictive method to adjust mean product effects for missing assessors in each experiment.
- Incorporate information on assessors' use of the scale to improve precision of analysis of future experiments.

5.3 Models and Analysis of Sensory Studies

The challenge in modelling sensory studies is that assessors' degrees of psychological differences are not reflected in the same degrees of the scale used. This is because different assessors use different amounts and different parts of the scale. Thus, the type of scale used greatly influences the analysis methods. Also, there is, almost always, no standard measure of attribute intensity that assessors' scores may be compared to. Gay and Mead (1992) discussed some of the problems associated with the different scales and the corresponding statistical analyses. Firstly, they pointed out that when scores are obtained from an interval line scale, it is common practice to use analysis of variance, assuming that the following four assumptions apply to the data.

- Scores are independent within assessors
- There is equal random variation between assessors
- Scores are normal
- Scores are made on the same scale of measurement for all assessors.

A model under these assumptions, for a given replicate, is given by

$$y_{ij} = \alpha_i + \theta_j + \epsilon_{ij}, \tag{5.1}$$

where y_{ij} is the score for assessor i corresponding to product j; α_i is the effect of assessor i; θ_j is the effect of product j and ϵ_{ij} is the random error assumed to be distributed $N(0, \sigma^2)$.

If these assumptions are violated, analysis could just be based on the product rankings given by the means of the assessors. That, though, is not adequate as it only allows one to test product differences and it does not allow for estimation of product means. Gay and Mead (1992) went further to suggest a modification of anova Model 5.1 for the case when the assumption of same scale of measurement is violated. This violation almost always occurs when one looks at interval scaling. So this model, they suggest, may be modified so that multiplicative coefficients are added in order to model differences in scale use between assessors or between blocks, if a block design is used. They termed the relative scale use *assessor expansiveness*. The model then becomes nonlinear when such coefficients are

added, and an iterative fitting method may be used for parameter estimation. Gay and Mead (1992) did not give this modified model in their report, but from their definition, it seems to be similar to the multiplicative model often used in plant variety trials (this is discussed further in the next subsection). It is given by

$$y_{ij} = \alpha_i + \beta_i \theta_j + \epsilon_{ij}, \tag{5.2}$$

where β_i is the coefficient of assessor i which models assessor expansiveness, and the rest of the parameters are as defined before.

Nonlinear modelling has problems such as non-convergence when few observations are available per assessor. Gay and Mead (1992) suggested a method of conducting an experiment in order to deal with this problem: they called this a *self-adjusting scale* method. This method works in such a way that products are arranged in either complete or incomplete blocks. For each block, assessors taste all products and then they choose two products which they perceive as extreme in intensity, in opposite directions, with respect to an attribute of interest. Finally, the rest of the products in that block are placed on the scale relative to those two. Thus, data used for analysis are the product ranking information from the blocks. The method is called self-adjusting scaling because assessors are asked to use the same scale units for blocks that differ in their within-block product differences. So, it is assumed that the scale self-adjusts to each block. A model similar to (5.2) is fitted to data from this method, with the expansiveness associated with blocks, and the following two-stage iterative method is used.

- Step 1: Initialise coefficients of expansiveness to 1.
- Step 2: Estimate product means by weighted regression of observed scores on the latest block expansiveness coefficients.
- Step 3: Update block expansiveness coefficients based on the latest estimates of product means. The expansiveness of scoring within blocks is related to product differences.
- Iterate between Steps 2 and 3 until convergence to final true product differences as perceived by the assessors.

The self-adjusting scale method is easy to use even when the assessors have not been well trained. The modelling suggested here is very similar to multiplicative

modelling for combining variety trial data. An analogy drawn between models for sensory data sensory studies and variety trials by other authors, will be discussed in the later subsections as it forms the basis of the analysis of the data under study in this project.

Mead and Gay (1995) gave a thorough discussion of sequential designs of sensory trials when information from previous trials is used for design and analysis of new trials. This is information such as block size, replication, position and carry-over effects. In contrast, in this project, information on assessor performance is used to improve analysis of data from future experiments, and this is done in the Bayesian framework.

The literature has a wide range of studies which are conducted with the main aim of evaluating assessor performance. As an example, Naes (1998) described three plotting techniques that can be used to detect individual differences between assessors when the sensory evaluation involves rankings. The first one is called an *eggshell plot* which is obtained by calculating some consensus ranking (for example, mean ranking for each product on a given attribute) of the whole panel, and then plotting each assessors' ranks to see how close they lie to this consensus ranking. The second one involves doing analysis of variance for individual assessors and then using the F-values for the test of equality of product effects, as a measure of the assessor's ability to detect differences among products. The third one is just a plot of means of raw data for all product effects and assessors, and the variation among scores for all products. This is basically a combination of the first two as it shows both mean consensus and variation within products.

Rossi (2001) discussed two measures of assessor performance which are *repeatability* and *reproducibility*. The former is the ability to score the same product consistently for each attribute, while the latter implies the ability to score products the same, on average, as other members of the panel. These measures arose from the context of analytical (or chemical) laboratory evaluation where inter-laboratory performance is monitored. They are both obtained from descriptive statistics for assessor and product combinations.

McEwan et al. (2002) followed a similar technique used in testing laboratory performance called *proficiency testing*, to compare results of different sensory panels. As part of a large European study, they had standard panels against which others were tested. Their measures involved doing a generalised Procrustes analysis to get a number of significant sensory dimensions for each panel, and they

also looked at how well assessors within a panel agreed with each other and with the overall consensus.

In order to model assessor performance in the apple experiments, heterogeneity of product and error variance between assessors is modelled with a multiplicative model similar to that suggested by Gay and Mead (1992). Also, few graphical methods, and none of the multivariate methods, as used by many sensory researchers, were explored here because the aim was not to classify products for commercial purposes but to correct for missing assessors and to improve precision in analysis of future experiments.

5.4 Multiplicative Interaction Models

Models with multiplicative interaction terms are widely used in analysing data from crop variety trials, and they have recently been adapted for analysing data from sensory studies. In crop variety trials, varieties are grown in different environments and the crop yields from these compared. Environments in this case mean either locations or a combination of locations and years of planting, if trials are carried out over several years. Usually, data from these trials are incomplete, thus making analysis not so straightforward. An additive model for such data would be

$$y_{ij} = \alpha_i + \theta_j + \epsilon_{ij}, \tag{5.3}$$

where y_{ij} is the mean yield from variety i in environment j, α_i is the effect of variety i, θ_j is the effect of environment j, and ϵ_{ij} is the random error term, assumed to be normally distributed with zero mean and variance σ^2. When trials are replicated, that is each variety is planted in several plots, the above model could then have a term for an interaction between variety and environment η_{ij}, and a third subscript would be added for replicate r so that the response is denoted by y_{ijr}. Such an interaction may be heterogeneous depending on the varieties' sensitivity to changes in environments. When such heterogeneity exists, an additive model like (5.3) above is inadequate for estimating standard errors of variety comparisons. The works of Yates and Cochran (1938) and Finlay and Wilkinson (1963) gave rise to a model which takes this heterogeneity into account by adding a multiplicative term to (5.3), and this is given by

$$y_{ij} = \alpha_i + \beta_i\theta_j + \epsilon_{ij}. \tag{5.4}$$

So, the interaction between variety and environment is modelled by a regression on location effects with coefficients β_i known as variety sensitivity parameters. Here, the environment effects are unobserved explanatory variables, and when they are treated as fixed effects they have to be estimated as well. The model can be generalised by assuming unequal residual variances for varieties so that $\text{var}(\epsilon_{ij}) = \sigma_i^2$. For the sake of parameter identifiability, the θ_j's are constrained to sum to zero while β_is have mean 1. Relative values of the β_is are then used to compare varieties: a variety for which $\beta_i > 1$ is seen as being more sensitive to environments and one with $\beta_i < 1$ is less sensitive, among other varieties in a trial.

Digby (1979) described a method of least squares that can be used to fit this model with equal residual variances when variety $\times$ environment data tables are incomplete. This involves alternating between estimation of variety-specific parameters and estimation of environment-specific parameters. On the other hand, Oman (1991) showed how maximum likelihood (ML) via Fisher's scoring algorithm can be used to estimate parameters in this model when data are possibly incomplete. He demonstrated this using plant genetics data collected to study the combined effects, on the firmness of tomatoes, of a genotype called NOR and other polygenes. Families of tomatoes are groups of tomatoes from the same plant and they are taken as random, while the genotypes are fixed within families. So both fixed and random effects enter the model multiplicatively, hence it becomes a mixed multiplicative effects model.

Gogel et al. (1995) used a mixed multiplicative model similar to the one of Oman (1991) but they used Restricted Maximum Likelihood (REML) to estimate variance parameters. This is because ML tends to give more biased estimates of variance parameters than REML when the number of fixed effects is large relative to the data. Their iterative method of parameter estimation involved using the *average information* (AI) algorithm.

In reaction to Gogel et al. (1995), Piepho (1997) showed how estimation procedures by Oman (1991) and Gogel et al. (1995) used to fit the Finlay and Wilkinson (1963) model can be used for a more general one by Mandel (1971), given as

$$y_{ij} = \mu + \alpha_i + \theta_j + \gamma\beta_i\omega_j + \epsilon_{ij}, \tag{5.5}$$

where β_i and ω_j are parameters for variety i and environment j respectively.

This is done with the environment effects regarded as random, so that ω_j and θ_j are independent and normally distributed with zero means and variances σ^2_ω and σ^2_θ, respectively. This is an *additive main effects multiplicative interaction* (AMMI) model widely used in biological and agricultural applications, according to Piepho (1997). It is considered less restrictive than the one defined above because the environment effect is not the only independent variable of regression, but a combined effect of varieties and environments is. The ranking of varieties based on *sensitivities* in the Finlay and Wilkinson (1963) model and that based on *coefficients* in the Mandel (1971) model gives the same results.

Nabugoomu et al. (1999) gave a generalisation of the Finlay and Wilkinson (1963) model for use in analysis of variety $\times$ locations across years data. Their general multiplicative model is given by

$$y_{ijk} = \alpha_i + \phi_k + \gamma_{ik} + \beta_{ik}\theta_{jk} + \epsilon_{ijk}, \tag{5.6}$$

where the subscript k is indicative of the year. Hence, y_{ijk} is the yield for variety i in location j nested within year k; ϕ_k is the year effect; γ_{ik} is the variety-by-year interaction and the rest of the parameters are similar to those of the multiplicative models defined above. The year and location j in year k effects θ_{jk} are regarded as random. If sensitivities are consistent across years then $\beta_{ik} = \beta_i$.

Nabugoomu et al. (1999) proposed a method which is a generalisation of the modified least squares algorithm of Digby (1979) for three-way incomplete data. They also used REML based on the Fisher scoring scheme for estimation of variety means and sensitivity coefficients when locations are regarded as random. These two methods gave similar results.

5.4.1 Multiplicative models for sensory studies

The approach followed by Brockhoff and Skovgaard (1994) and Smith et al. (2003) is that of taking models for sensory data as particular cases of the analogy that exists between models for variety trials and those for comparing methods of measurement. So, human assessors act as measuring instruments in sensory studies, so that the differences between food products are realized through assessors' scores. Thus, assessors and food products are analogous to plant varieties and locations or locations in years, respectively. In sensory studies, though, there are often more design attributes that need to be accounted for, such as carry-over from a previous product and order of presentation.

On the other hand, Theobald and Mallinson (1978) discussed these models for comparing methods of measurement in the case of comparative calibration, that is, a comparison of measuring instruments in which none is regarded as standard. This is true of most studies of sensory evaluation of food. These models, also referred to as calibration equations, were discussed by Theobald and Mallinson (1978), both as structural and functional relations. As a structural relation, the true response is expressed as a linear function of some hypothetical standard measurement F_j, which is a random variable. Thus, for the jth sample to be measured,

$$y_{ij} = \alpha_i + \beta_i F_j + \epsilon_{ij} \tag{5.7}$$

where y_{ij} is the response from instrument i; α_i is the ith instrument effect; the coefficient β_i is referred to as a calibration factor for the ith instrument; F_j is the hypothetical standard measurement of sample j and the random error $\epsilon_{ij} \sim N(0, \sigma_i^2)$. When F_j and ϵ_{ij} are both assumed to be normally distributed, this becomes a factor analysis model with one common factor. If the calibration model is a functional relationship, then the true response is a linear function of a scalar parameter θ_j for the jth sample, and thus for the ith instrument, the model becomes

$$y_{ij} = \alpha_i + \beta_i \theta_j + \epsilon_{ij} \tag{5.8}$$

$i = 1, ..., n$; and for parameter identifiability, the restrictions $\sum \theta_j = 0$ and $\sum \theta_j^2 = n$ are imposed so that α_i becomes the expected value of the sample mean $\overline{y}_{i.}$. Theobald and Mallinson (1978) used maximum likelihood to estimate parameters in this model. As a measure of the performance of an instrument, they defined precision π_i as $\frac{\beta_i}{\sigma_i}$. Modelling sensory panel data is a case of comparative calibration, as there is no standard measurement for the attributes of interest. Although the terminology used by Brockhoff and Skovgaard (1994), Brockhoff (1997) and Smith et al. (2003) is mainly taken from the literature on multiplicative models for variety trials, their models are really the general cases of the functional calibration models discussed by Theobald and Mallinson (1978).

Brockhoff and Skovgaard (1994) presented the use of the multiplicative model, with fixed effects, like the one in (5.8) to account for the heterogeneity of product and error variance between assessors. The β_is are constrained to have mean 1 and are seen as measures of assessors' discriminating ability. They were defined by Mead and Gay (1995) as *assessor expansiveness*. These two definitions have

the same interpretation that if an assessor is good, they will have low residual variance σ_i^2 while using a big part of the scale in order to clearly show differences between products, and therefore their value of β_i will be greater than 1. For such assessors, the value of the precision $\pi_i = \frac{\beta_i}{\sigma_i}$ will be high. It is easy to see that, for replicated tasting, high values of the F-test statistics of the product differences correspond to high values of π_i.

Brockhoff and Skovgaard (1994) and Brockhoff (1997) illustrated how a multiplicative model can be generalised to account for the whole design structure when each product is tasted more than once, that is, in a replicated experiment. Thus when replication effects are indexed by r, the model has the form

$$y_{ijr} = \alpha_i + \beta_i(\theta_j + \gamma_r) + \epsilon_{ijr}, \tag{5.9}$$

where γ_r is a replication effect and this, like the product effect, is realized through the assessors' use of the scale. Brockhoff (1997) also proposed some significance tests of the multiplicative model to test for common assessor variances and common assessor expansiveness.

Smith et al. (2003) used the multiplicative model for products assumed to be a random sample from a larger population of products. This led to a mixed multiplicative model with assessor effects assumed fixed and products as random. They used REML for parameter estimation in this case.

5.5 Analysis of Individual Tasting Experiments

Analysis of variance (ANOVA) of the individual apple tasting data sets, for the sweetness attribute, showed that out of 45 data sets, only 11 have significant replicate effects; 6 have significant session effects and 3 have significant order effects, all tested at 5% level of significance. These three effects were therefore not included in the models for analysing these data. The additive model for an individual dataset is thus given by

$$y_{ijr} = \alpha_i + \theta_j + \epsilon_{ijr}, \tag{5.10}$$

where y_{ijr} is the sweetness score from assessor i, tasting product j in replicate r; α_i is the effect of assessor i; θ_j is the effect of product j and ϵ_{ijr} is a random error term assumed to be normally distributed with mean 0 and variance σ^2.

The above model does not take into account the heterogeneous interaction between the assessor and product effects. This interaction, shown to be partly made up of differences in assessor expansiveness, can be seen in the plots of mean sweetness scores versus products for four of the experiments in Figure 5.1. The colours do not correspond to the same assessors in all four plots because attendance was not the same for all experiments as shown by the incidence matrix in Appendix B. The plots show general consistency between assessors in the ordering of products on the scale from 0 to 100, but there are differences in expansiveness.

A model with multiplicative interaction effects was therefore chosen instead of (5.10). This is similar to the one given in (5.4) as

$$y_{ijr} = \alpha_i + \beta_i \theta_j + \epsilon_{ijr}, \tag{5.11}$$

with the ϵ_{ijr} distributed independently as $N(0, \sigma_i^2)$, so that there is heterogeneous residual variances for assessors. For parameter identifiability, the condition $\sum_j \theta_j = 0$ was applied and for the sake of interpretation, as before, the β_i's were constrained to have mean 1. So when using this model, assessor performance can be measured by expansiveness and precision.

To fit the fixed effects multiplicative model (5.11) with the conditions on the parameters mentioned above, an alternating regressions scheme originally proposed by Digby (1979) and generalised by Nabugoomu et al. (1999) was followed. This scheme, described below, is appropriate in this case because some data sets are incomplete. Here, it is generalised so that the σ_i differ between assessors.

The alternating regressions algorithm for fitting the fixed effects multiplicative model (5.11) is as follows.

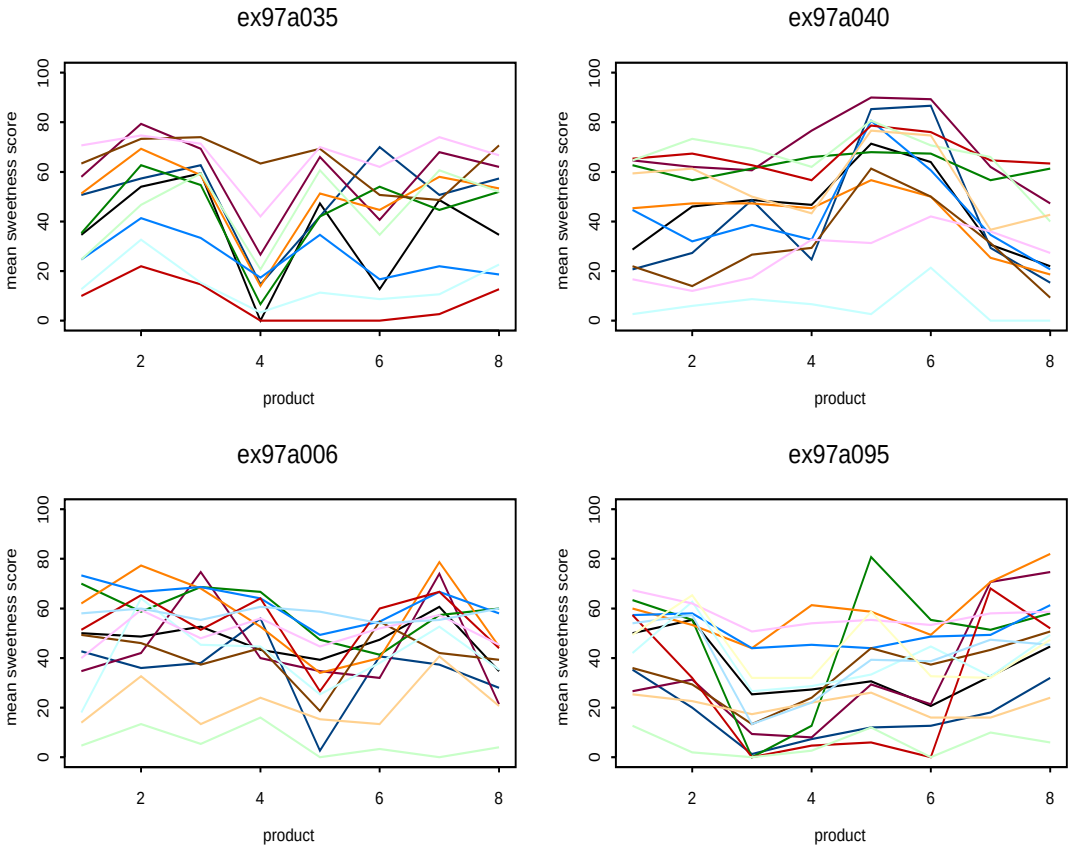
Step 0: Estimate product effects $\hat{\theta}_j$ from model (5.10) by least squares.

Step 1: Create a variate $M_j = \hat{\theta}_j$ for product j. Regress y_{ijr} on M_j for each i and r for which y_{ijr} is not missing. This gives the estimates $\hat{\alpha}_i$, $\hat{\beta}_i$ and $\hat{\sigma}_i$.

Step 2: Scale $\hat{\beta}_i$s to unit mean. Obtain new estimates of θ_j by regressing $y_{ijr} - \hat{\alpha}_i$ on the $\hat{\beta}_i$ through the origin, with weights $\hat{\sigma}_i^{-2}$, for each j and r for which y_{ijr} is not missing.

Step 3: Repeat Step 1 and Step 2 until convergence.

Figure 5.1: Plots of mean sweetness score vs products for 4 experiments



Convergence is assumed when successive values of the estimates differ by not more than 0.0001. The parameter estimates from this analysis are plotted for each of the assessors across experiments in Figure 5.2. These are presented in graphs showing respectively, $\hat{\alpha}_i$s, $\hat{\beta}_i$s and $\hat{\sigma}_i$s. For each of these parameters, the plots are done for the assessors split into two groups. This is done so that the patterns of changes in estimates may be clearly seen for each assessor.

One can observe that estimates of β_is and σ_is within each experiment are highly variable and therefore, there is a need for a model to account for this variability. Also, these estimates do not seem to have converged. To check whether the alternating regressions algorithm converges to maximum likelihood estimates is not straightforward because there are too many parameters to estimate, and some data sets are incomplete.

As an attempt to get maximum likelihood estimates, a log-likelihood function for the model was derived analytically, and a numerical procedure was used to maximise this with respect to the parameters. This failed because the function was too complex. Brockhoff and Skovgaard (1994) also followed the same procedure of deriving the log-likelihood analytically and then maximising it numerically but they still could not prove that the maxima of the function converged to maximum likelihood estimates. As a solution to these problems, a hierarchical random effects model was considered and fitted using the Bayesian approach. This helped to model the variability in the data, and also to achieve one of the aims of the project which was to use information about assessors to model data from future experiments. The random effects model is also better because of the incompleteness of the data sets.

5.5.1 Bayesian hierarchical model for individual tasting experiments

The apple data sets have a multilevel structure: there are several responses from each assessor who tastes products in two different sessions of each of the three replications. In such data, parameters specific to one level may depend on parameters specific to other levels, and hierarchical models allow for this complexity to be modelled. In the Bayesian approach, parameters are taken as randomly sampled from certain distributions, and the models are often described using a directed acyclic graph (DAG). This is a graph that shows conditional dependence between parameters in a Bayesian model, and it was thoroughly discussed by Gilks et al. (1996). The WinBUGS program introduced in Chapter 4, allows one to define a model using this graph.

A hierarchical random effects model with multiplicative effects for a single tasting experiment (5.11) is given by the DAG in Figure 5.3, where all the data and parameters are represented by nodes. Stochastic nodes are represented by circles: if there were fixed ones they would be represented by square nodes. The nodes

Figure 5.2: Plots of frequentist parameter estimates from the multiplicative model (5.11) of individual data sets

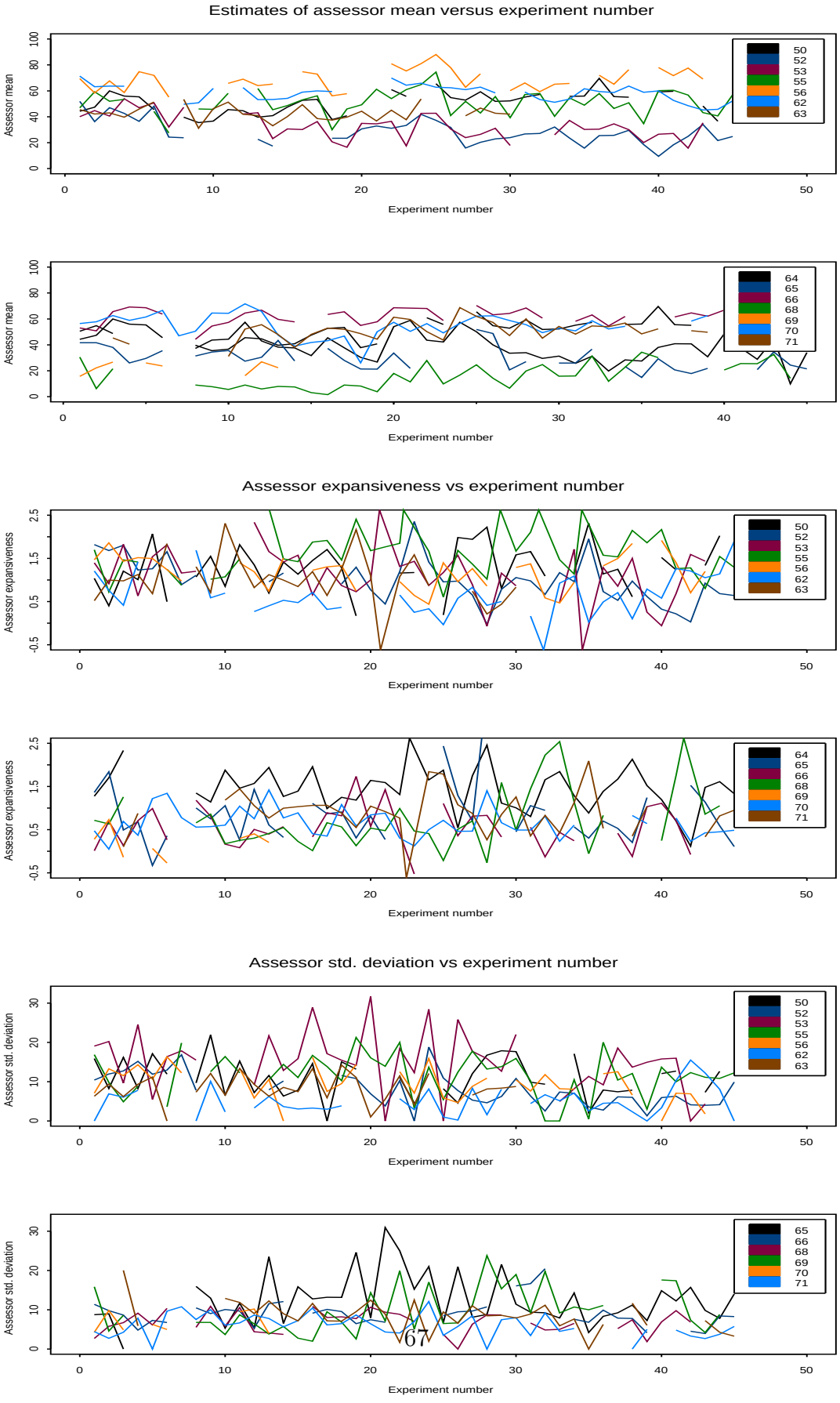


Table 5.1: Prior distributions for the multiplicative model (5.11)

$\alpha_i \sim N(\mu_A, \sigma_A^2)$	$\mu_A \sim N(m_A, r_A^2)$	$d_A s_A^2 \sigma_A^{-2} \sim \chi^2(d_A)$
$\beta_i \sim N(1, \sigma_B^2)$	$d_B s_B^2 \sigma_B^{-2} \sim \chi^2(d_B)$	
$\theta_j \sim N(0, \sigma_P^2)$	$d_P s_P^2 \sigma_P^{-2} \sim \chi^2(d_P)$	
$ds^2\sigma_i^{-2} \sim \chi^2(d)$		

are joined by directed lines to show conditional dependences. Solid arrows show probabilistic dependences, while dashed arrows show deterministic ones.

In Figure 5.3, the sweetness score, y_{ijr}, has expectation μ_{ijr} and standard deviation σ_i. The mean μ_{ijr} is in turn determined by the assessor effect α_i, the expansiveness β_i and the product effect θ_j. The β_is are taken to have prior mean 1 so that they are interpreted as relative measures of assessor expansiveness. The subscripts A, B, P are used to denote populations of assessor effects, assessor expansiveness and product effects, respectively. The form of the hierarchical priors is given in Table 5.1, and the values of means and variances for these, given in Table 5.2, were obtained from an expert who was involved in the apple tasting studies (Tony Hunter, personal communication). He mentioned that because of the high variability in these studies, there is very little certainty in the information, hence the low degrees of freedom. The $\sim$ means distributed as or independently distributed as. The ds in the prior definitions of variance parameters are the numbers of degrees of freedom corresponding to the prior estimates of sample variances, s^2. A prior $d_* s_*^2 \sigma_*^{-2} \sim \chi^2(d_*)$ is equivalent to $\sigma_*^{-2} \sim Gamma(d_*/2, d_* s_*^2/2)$ where * denotes a corresponding subscript.

The additive random effects model which assumes equal assessor variances was fitted and compared to the multiplicative one above. Similar priors as for the multiplicative model were used. This additive model was given earlier as

$$y_{ijr} = \alpha_i + \theta_j + \epsilon_{ijr}, \tag{5.12}$$

where $\epsilon_{ijr} \sim N(0, \sigma^2)$.

Figure 5.3: DAG of a multiplicative model (5.11) for a single tasting experiment

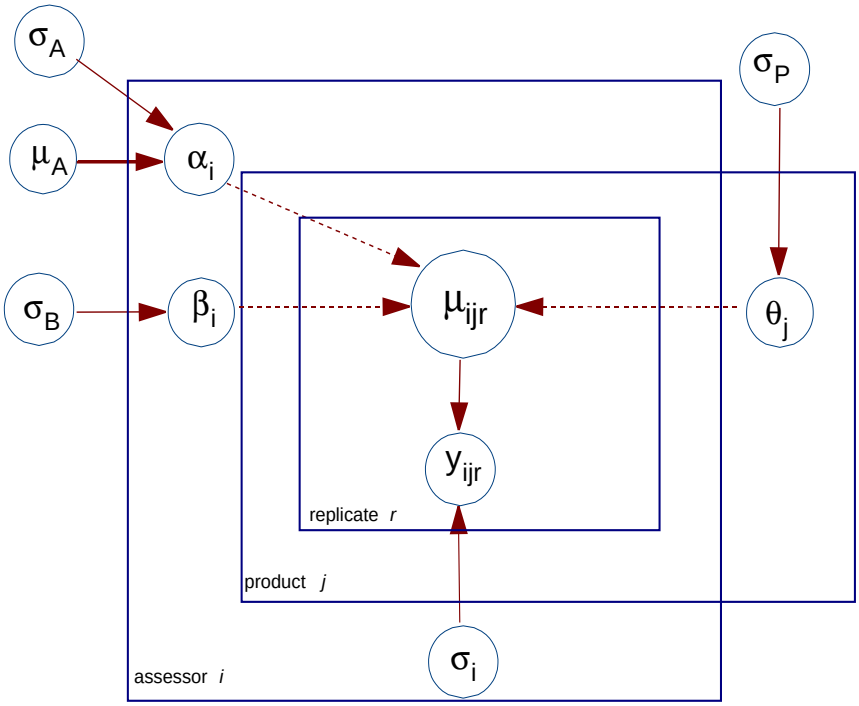


Table 5.2: Parameter values defining prior distributions for model (5.11)

Parameter	Mean	Variance	Estimate	Degrees of freedom
μ_A	33	30		
s_A^2			264	7
s_B^2			0.14	9
s_P^2			39	7
s^2			100	54

Model fit was determined by the deviance (-2log-likelihood) obtained from the MCMC analysis. However, for hierarchical models, the number of parameters is not clearly defined: thus, Spiegelhalter et al. (2002) derived an effective number of degrees of freedom p_D, using a decision theoretic approach. This measures the extent of model complexity, and thus adding it to the posterior mean deviance gives a statistic that penalises a model for complexity. This statistic, called the deviance information criterion (DIC), can be used to compare models. The model with the smallest value of DIC would best fit a replicate dataset of the same structure as the one currently observed. This criterion is discussed further in Spiegelhalter et al. (2002).

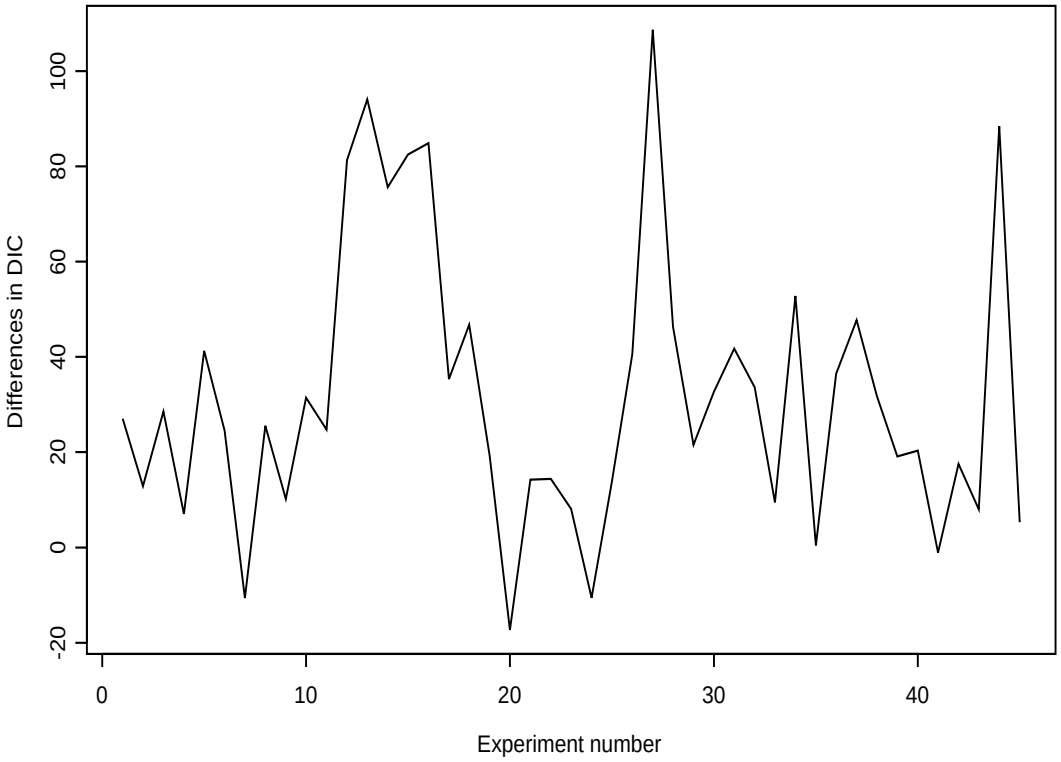
The values of DIC for each model and each of the 45 data sets were obtained from 10000 samples after a burn-in of 5000. A plot of differences in the values of DIC (that is, DIC for Model (5.12) $-$ DIC for Model (5.11)) is given in Figure 5.4. All but three of the differences are positive, showing that the DIC for the analysis of variance model with equal residual variances is higher than that for the multiplicative model. Therefore, the multiplicative model where different residual variances are allowed for the assessors, fits the data better.

Posterior means of the parameters from fitting the random effects multiplicative model (5.11) to the individual data sets are plotted in Figures 5.5 and 5.6. Here again, for each parameter, the plots are made for assessors split into two groups. When comparing these to the estimates from the frequentist analysis plotted in Figure 5.2, one can see that they are less variable within experiments, and hence it now makes sense to plot the estimates of assessor precision π_i. Overall, assessor means and standard deviations stay fairly constant across all experiments although there is some variability in the estimates of expansiveness and precision.

5.6 Analysis of Combined Tasting Data

Data from all 45 experiments were combined in order to perform a meta-analysis. Instead of using the raw data for this, means were calculated over replicates within combinations of assessor and product in each experiment. This causes a slight loss of efficiency because of missing values, but it is the usual procedure for combining data for meta-analysis in other areas such as variety trials studies. From now on, the *sweetness score* in the context of combined analysis will refer to these means over replicates.

Figure 5.4: Differences in DIC values of models (5.12) and (5.11)



An analysis of variance model for these data is given by

$$y_{ijk} = \alpha_i + \phi_k + \theta_{j(k)} + \gamma_{ik} + \zeta_{ij(k)} + \epsilon_{ijk}, \tag{5.13}$$

where y_{ijk} is the mean response of assessor i for product j in experiment k; α_i is the effect of assessor i; ϕ_k is the effect of experiment k; $\theta_{j(k)}$ is the effect of product j which is nested within experiment k; γ_{ik} is the interaction term between assessor i and experiment k; $\zeta_{ij(k)}$ is the interaction between assessor i and product j within experiment k and ϵ_{ijk} is the random error term assumed to be normal with mean 0 and variance σ^2.

As in the individual data analyses, the interaction between assessor and product within experiment may be modelled as multiplicative. Thus, the multiplicative model (5.11) is generalised so that the stimulus structure, made up of products and experiments in which they occur, is realized through assessor expansiveness (Brockhoff (1997)). This generalised model is given by

Figure 5.5: Plots of posterior means of the parameters of Model (5.11) of individual data sets

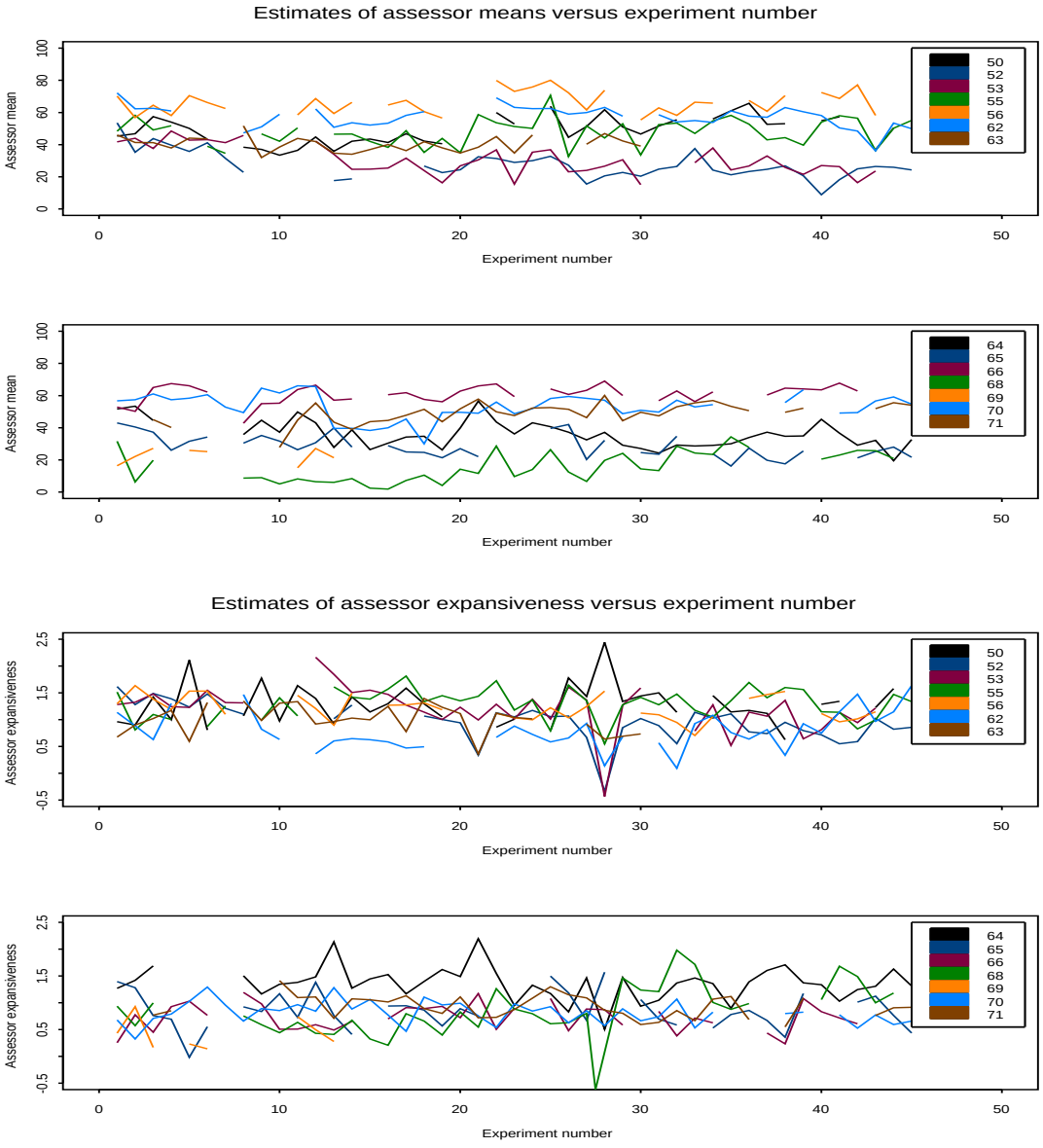
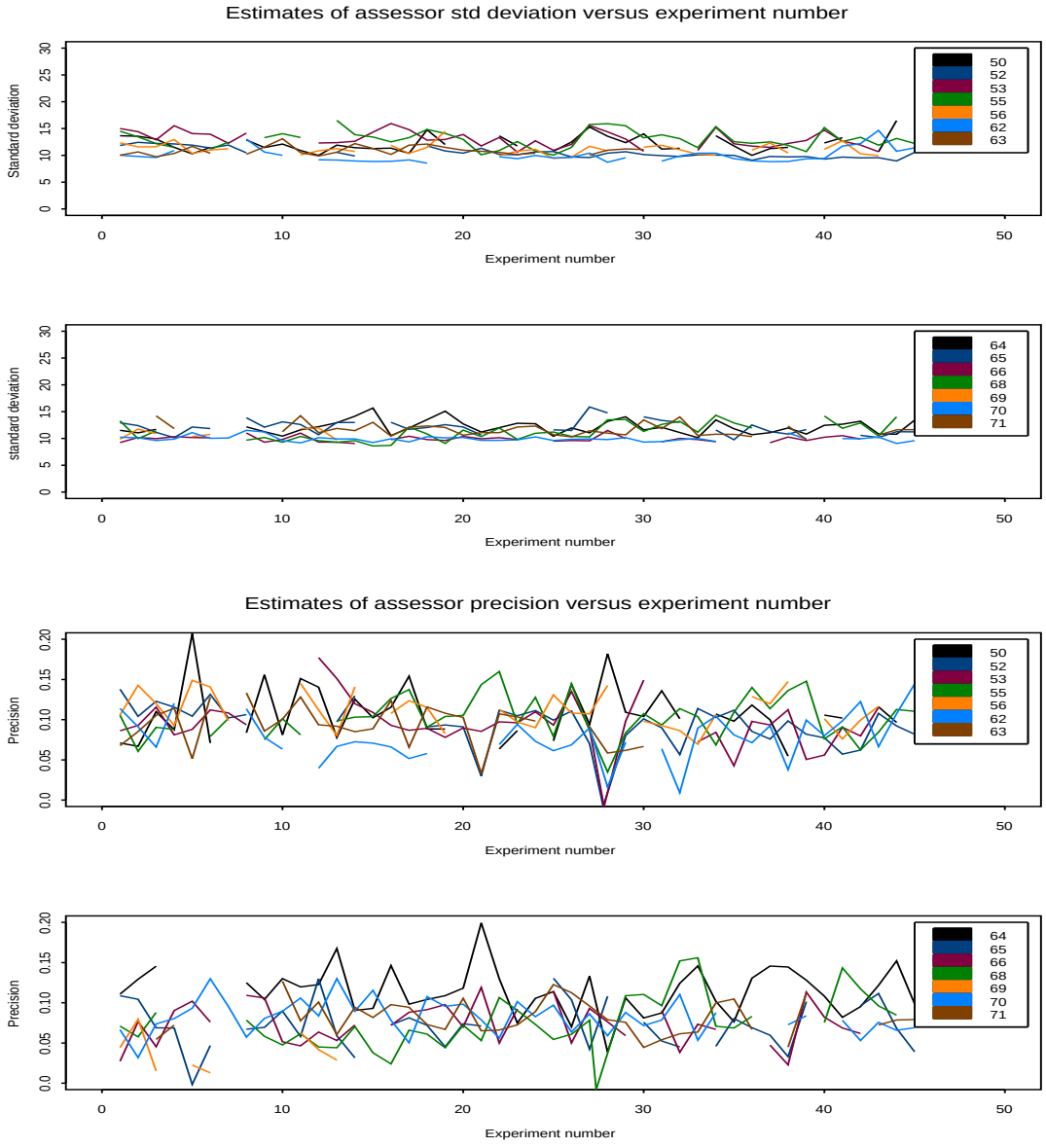


Figure 5.6: Plots of posterior means of the parameters of Model (5.11) of individual data sets



$$y_{ijk} = \alpha_i + \gamma_{ik} + \beta_{ik}(\theta_{j(k)} + \phi_k) + \epsilon_{ijk}, \tag{5.14}$$

where β_{ik} is assessor i's expansiveness in experiment k; $\epsilon_{ijk} \sim N(0, \sigma_i^2)$ and the other parameters are as defined before. This model is analogous to Model (5.6) used by Nabugoomu et al. (1999) for across-years analysis of variety trials, with our experiment effect corresponding to their year effects. In model (5.6) though, the expansiveness is not applied to the year effects ϕ_k and the explanation for that difference was summarised as follows (Mike Talbot, personal communication).

> In plant variety trials, factors that influence sensitivity to locations are often different from those associated with seasonal effects and information on sensitivity to location is more important to growers than sensitivity to seasonal effects.

In contrast, for the apple tasting, products from many different seasons throughout the year (over several experiments) are tasted, and it is important to know how assessors score them, given that they are grown in different seasons.

Model (5.14) was fitted assuming all effects to be fixed using a generalised version of the alternating regressions algorithm used for individual experiments analysis. This was not appropriate because there are too many parameters to be estimated; there are no reasonable asymptotic assumptions and there are many sources of variation. So, as for the individual experiments, a Bayesian random effects model was used.

5.6.1 Bayesian hierarchical model for combined data

It is possible to reparametrize Model (5.14) in order to speed up convergence during MCMC sampling, and this is called *hierarchical centring* (Gelfand et al. (1995)). So, the assessor effect in experiment k, α_{ik}, was centred on α_i, while the product effect $\theta_{j(k)}$ was centred on ϕ_k, and β_{ik} on β_i. This resulted in the model

$$y_{ijk} = \alpha_{ik} + \beta_{ik}\theta_{j(k)} + \epsilon_{ijk}. \tag{5.15}$$

A directed acyclic graph (DAG) for model (5.15) is given in Figure 5.7, and Table 5.3 shows the priors placed on its parameters.

Figure 5.7: DAG of multiplicative model (5.15) for combining data over experiments

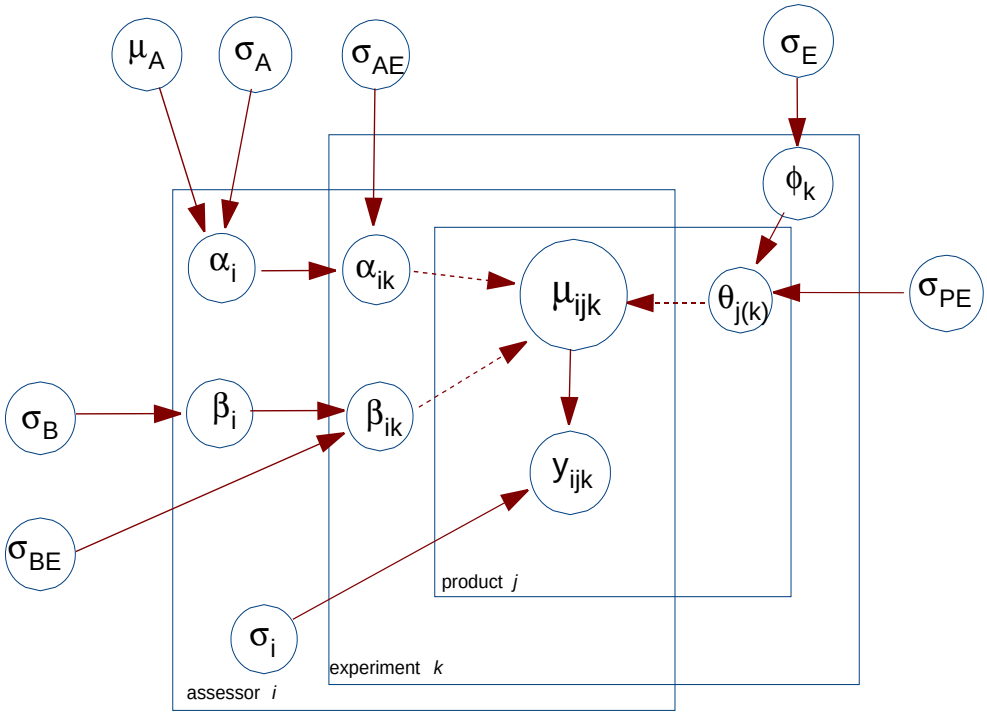


Table 5.3: Prior distributions for the multiplicative model (5.15)

$\alpha_{ik} \sim N(\alpha_i, \sigma^2_{AE})$	$\alpha_i \sim N(\mu_A, \sigma^2_A)$	$\mu_A \sim N(m_A, r^2_A)$
$d_{AE} s^2_{AE} \sigma^{-2}_{AE} \sim \chi^2(d_{AE})$	$d_A s^2_A \sigma_a^{-2} \sim \chi^2(d_A)$	
$\beta_{ik} \sim N(\beta_i, \sigma^2_{BE})$	$\beta_i \sim N(1, \sigma^2_B)$	$d_{BE} s^2_{BE} \sigma^{-2}_{BE} \sim \chi^2(d_{BE})$
$d_B s^2_B \sigma^{-2}_B \sim \chi^2(d_B)$		
$\theta_{j(k)} \sim N(\phi_k, \sigma^2_{PE})$	$\phi_k \sim N(0, \sigma^2_E)$	$d_{PE} s^2_{PE} \sigma^{-2}_{PE} \sim \chi^2(d_{PE})$
$d_E s^2_E \sigma^{-2}_E \sim \chi^2(d_E)$	$d s^2 \sigma_i^{-2} \sim \chi^2(d)$	

Table 5.4: Parameter values defining prior distributions for Model (5.15)

Parameter	Mean	Variance	Estimate	Degrees of freedom
μ_A	33	30		
s^2_{AE}			66	4.4
s^2_A			264	7
s^2_{BE}			0.0225	4.4
s^2_B			0.14	9
s^2_{PE}			39	30.8
s^2_E			39	4.4
s^2			50	30.8

The subscripts A, B, P and E are used to denote populations of assessors, assessor expansiveness, products and experiments, respectively. Here, again, the same expert's prior values of parameters as for the individual data analysis were uses. He had only had experience analysing data sets individually, though, hence for combined analysis, his prior information on parameters from experiment populations was not very certain and it had to be down-weighted. Down-weighting was achieved by dividing the degrees of freedom corresponding to variance parameters by 10, and this was then taken as the expert prior information. So, the values for the priors used in the combined analysis are given in Table 5.4.

As mentioned earlier, one of the purposes of doing a combined analysis was to predict average product effects $\theta_{j(k)}$ over assessors within experiments, adjusted for missing assessors. This was achieved by defining the nodes for the assessor-specific parameters as if all assessors were present at every experiment. In that way, estimates of α_{ik} and β_{ik} were generated even though assessor i did not take part in experiment k. Thus, the overall posterior product means obtained at the end of the MCMC sampling were adjusted for the missing assessors.

5.6.2 Results of combined analysis

Any trend over experiments, may be seen by plotting the posterior means of experiment effect for data D, that is $E(\phi_k|D)$. This is given in Figure 5.8. In order to see changes in experiment effects, that are free from any product effects, the plot of $E(\overline{\alpha}_{ik} - \alpha_i|D)$ is presented in Figure 5.9. This shows changes in experiment effects which result in the changes (increases or decrease) in posterior means of the assessors. There was no significant auto-correlation between experiment effects as revealed by their partial auto-correlation plot.

Figure 5.8: Plot of posterior means of experiment effects ϕ_k

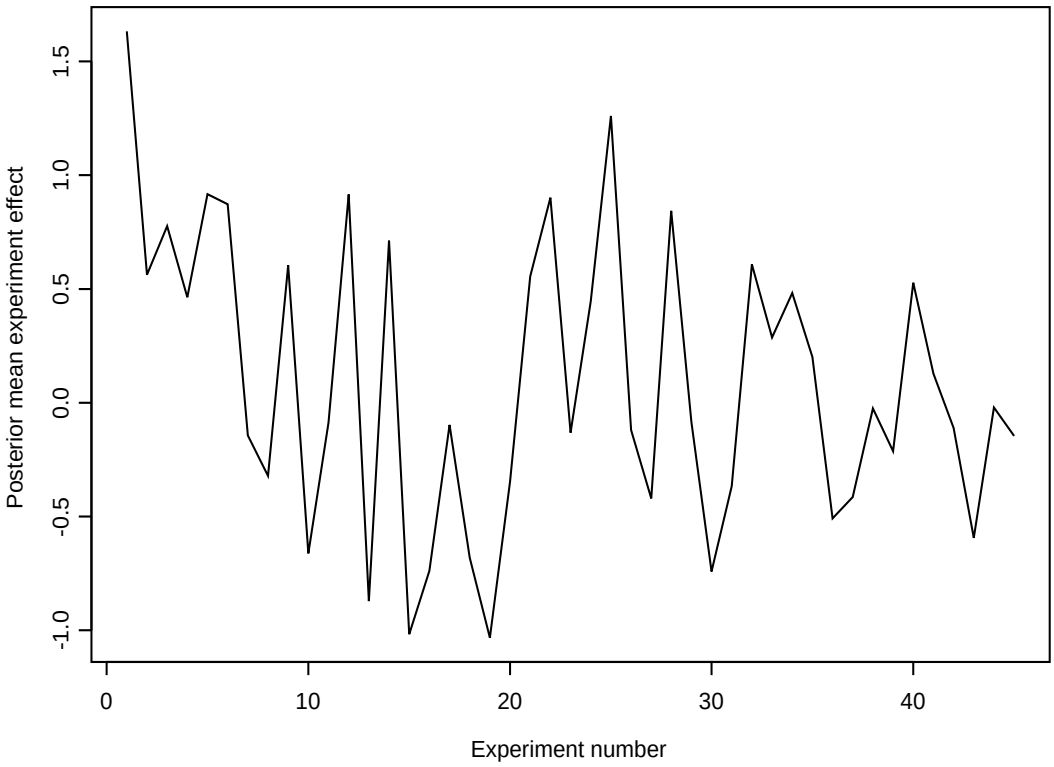


Figure 5.9: Plot of posterior experiment effects which result in changes in assessor means

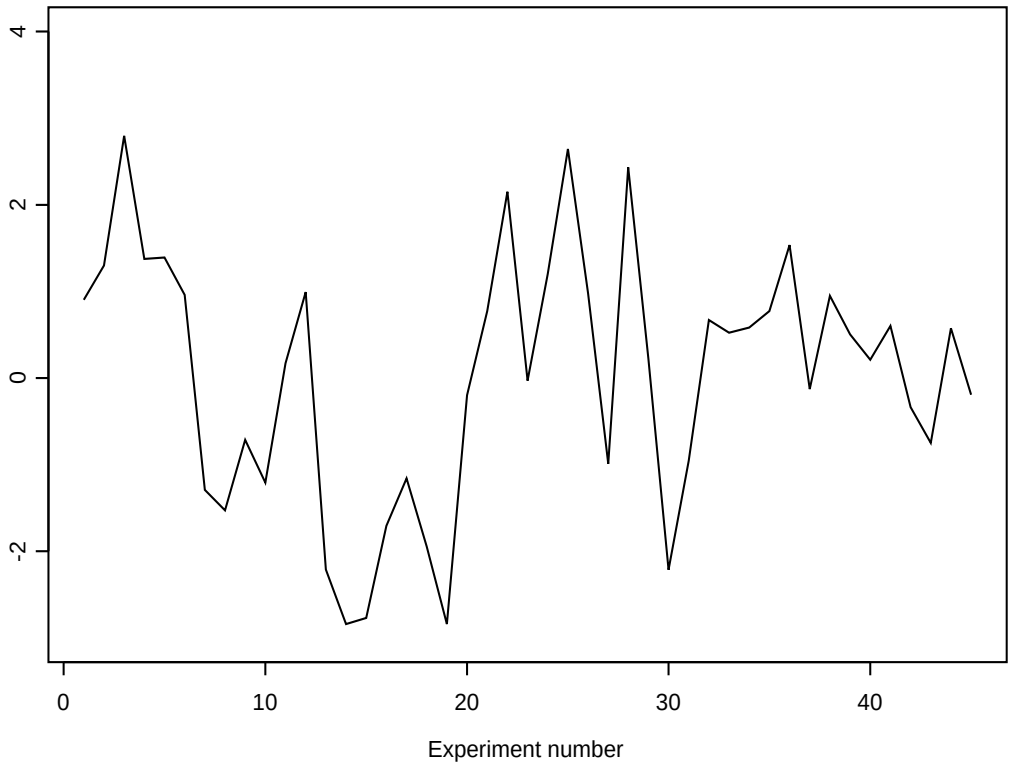


Table 5.5: Posterior means (and s.e.) of the parameters of Model (5.15)

Assessor	$\hat{a}_i$(s.e)	β_i(s.e)	σ_i(s.e)	π_i(s.e)
50	48.61 (2.29)	1.26 (0.15)	9.92 (0.46)	0.127 (0.017)
52	26.58 (2.14)	1.01 (0.14)	7.18 (0.38)	0.141 (0.020)
53	30.11 (2.39)	1.31 (0.16)	13.48 (0.64)	0.097 (0.013)
55	47.78 (2.13)	1.31 (0.16)	12.46 (0.56)	0.106 (0.014)
56	65.08 (2.32)	1.19 (0.15)	8.18 (0.43)	0.146 (0.021)
62	57.16 (2.15)	0.78 (0.14)	5.41 (0.28)	0.146 (0.027)
63	40.40 (2.44)	1.03 (0.15)	7.69 (0.43)	0.135 (0.022)
64	35.79 (2.10)	1.44 (0.16)	10.17 (0.51)	0.142 (0.015)
65	28.12 (2.22)	0.80 (0.14)	9.66 (0.43)	0.083 (0.015)
66	60.79 (2.16)	0.65 (0.13)	6.49 (0.29)	0.101 (0.021)
68	15.58 (2.05)	0.82 (0.14)	9.58 (0.49)	0.085 (0.016)
69	22.18 (4.16)	0.54 (0.21)	6.68 (0.53)	0.081 (0.032)
70	53.19 (2.04)	0.81 (0.12)	5.19 (0.26)	0.156 (0.025)
71	47.90 (2.23)	0.93 (0.15)	9.19 (0.45)	0.102 (0.018)

The posterior means of the α_i, β_i, σ_i and π_i and their standard errors are given in Table 5.5. The assessor means have a range of about 50. Assessor 69's estimates are highly variable, and that is due to the fact that she pulled out of the study after the first 17 experiments in which she gave very low scores anyway. It may be wise to omit her responses from the analysis. Overall, assessors 52, 56, 62, 63, 64 and 70 are shown by their estimates of precision to have high discriminating abilities.

5.7 Analysis of Future Experiments

Posterior distributions of the parameters obtained from the analysis of past data may be used as priors for analysing data from future experiments. In this case, past data are the combined data from the 45 apple tasting experiments. The effects of the future experiment are assumed to have been drawn from the same population of experiment effects as the past ones. Such an analysis is illustrated here by assuming there is one future experiment in which all of the assessors from the past experiments take part.

The DAG in Figure 5.7 is extended to include data from a future experiment. The extended DAG is given in Figure 5.10. Parameters specific to the future experiment are subscripted with a $*$, and m indexes future products, so that the future response from assessor i is denoted by y_{im*} for a future product $\theta_{m(*)}$. These future product effects are assumed to be drawn from the same population

Table 5.6: Parameter values defining expert and diffuse prior distributions for analysing future experiment model in 5.10

		Variance			Degrees of freedom	
Parameter	Mean	Expert	Diffuse	Estimate	Expert	Diffuse
μ_A	33	30	30			
s^2_{AE}				66	4.4	0.88
s^2_A				264	7	1.4
s^2_{BE}				0.0225	4.4	0.88
s^2_B				0.14	9	1.8
s^2_{PE}				39	30.8	6.16
s^2_E				39	4.4	0.88
s^2				50	30.8	6.16

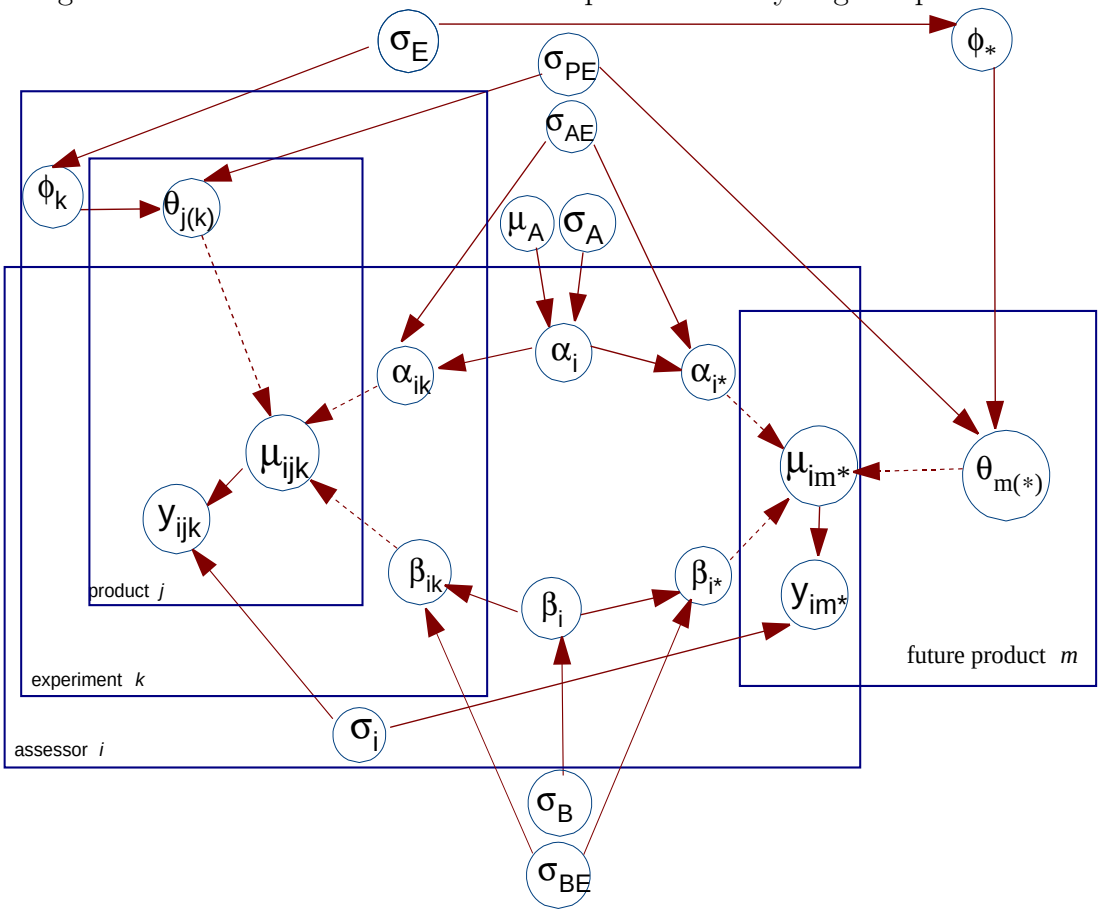
of product effects as the ones in the past, hence having the same distribution with variance σ^2_{PE}. Thus,

$$y_{im*} = \alpha_{i*} + \beta_{i*}\theta_{m(*)} + \epsilon_{im*}. \tag{5.16}$$

Model (5.16) may be further extended to a case in which combined analysis done for two or more future experiments. In that case, the DAG would have an extra box around the future products' box, to indicate a factor *future experiment*. Also, it is possible to extend this model to a case in which future assessors include some who did not take part in the past experiments. These assessors would be assumed to have come from the same population as the past assessors. The future product box would then be overlapping with a separate one indicating a factor *new future assessor*. In this thesis, though, analysis of the future experiment is restricted to the model shown in 5.10

This analysis was illustrated using similar prior distributions as for Model (5.15) of the combined analysis, and in order to test for the robustness of this analysis to prior information, it was carried out using more vague prior information as well. This vague prior information was obtained by dividing the degrees of freedom corresponding to the variance parameters by 5. Table 5.6 shows these values of the priors. The priors with heading 'Expert' denote the expert prior information used in the combined analysis in the past section while the priors labelled 'Diffuse' are the vague ones.

Figure 5.10: DAG of model for future experiment analysis given past data



5.7.1 Results of analysis of a future experiment

As shown in the DAG, this is the case of one future experiment, and future assessors are all those who took part in the past experiments. Often, in food tasting experiments, the aim is to establish differences between products as perceived by human assessors, and therefore, it is desirable that the average posterior variance of the product differences, $\theta_{a(*)} - \theta_{b(*)}$, be as small as possible. The aim of using past information to analyse future data is therefore to reduce this variance, compared to what it would be if the past information was not incorporated into the analysis.

A measure of average posterior variance of product differences is calculated in a similar way as in Theobald et al. (2002). They do this in the context of variety trials where, for p varieties, the posterior variances of the $p(p-1)$ differences between variety yields are of interest. In their case though, the same varieties occur in the trial data and, potentially, in the future target site as well, whereas in food tasting, the future products are different from those tasted in the past. So, given past data D, the average posterior variance of future product differences is given by

$$\frac{1}{p(p-1)} \sum_{a \neq b} \text{var}(\theta_{a(*)} - \theta_{b(*)}|D) \tag{5.17}$$

which is equivalent to

$$\frac{2}{p-1} \left\{ \sum_{j} \text{var}(\theta_{j*}|D) - p \ \text{var}(\bar{\theta}_*|D) \right\}. \tag{5.18}$$

From fitting the model in Figure 5.10 using WinBUGS, the value of $\text{var}(\theta_{j*}|D)$ was obtained from the MCMC sampling and an extra node was created for $\bar{\theta}_*$ in order to obtain $\text{var}(\bar{\theta}_*|D)$. Fitting the multiplicative model (5.11) to a future data set is expected to give less precise analysis when the prior distributions used are not posterior distributions from analysing past data. In that case, the value of (5.18) is expected be larger than when the data set is analysed using the model in Figure 5.10. This was illustrated by using the combined apple data sets as past data and data from experiment *ex97a095* as a future experiment. This data set is one of the 45 that form a combined data set, so it is used twice, but its influence on the posterior distribution should be slight as there are so many of them. Data

Table 5.7: Values of (5.18) from analysing data from *ex97a095*

	Expert prior	Diffuse prior
Model(5.11)	9.51	12.41
Model in Figure 5.10	4.49	6.09

from this experiment were balanced and all 14 assessors were present. This makes it easy to incorporate posterior distributions specific to each assessor from past analysis, as prior information for their corresponding future analysis. In order to avoid autocorrelation in samples when using vague priors, the MCMC chain was run for 100 000 iterations, taking only 1 in every 10 samples. The analysis resulted in the values of (5.18) given in Table 5.7, under the two priors.

It can be seen from Table 5.7 that the analysis based on diffuse priors gives higher posterior variances of the product differences. In both cases though, analysis of future data using posterior distributions from past analysis as prior information results in over 50% reduction in posterior variance of the product differences.

5.8 Summary

The three main aims of combining apple data from the different experiments have been achieved, based on the assumptions of a multiplicative model. These aims were to study assessor performance over time, to adjust product effect predictions when some assessors are not present, and to use the information about assessors in the past data to obtain more precise analyses of future experiments. The Bayesian random effects approach performs better at modelling individual data sets than the fixed effects model. It has been seen to give posterior estimates of the parameters which are less variable. The model with multiplicative effects is good for modelling the assessor $\times$ product variance heterogeneity. It can be expected to perform better than the analysis of variance model when there is differential use of the scale, which is measured by the expansiveness coefficient β_i, for each assessor.

Bayesian hierarchical models have a lot of advantages, including ease of modelling heteroscedasticity and incompleteness in the data. The effects of assessors that were not present were easily predicted by borrowing strength from their responses in other experiments. Incorporating information from past analysis into future analysis has been seen to increase precision by reducing the variance of product

differences. The disadvantage in Bayesian modelling is the difficulty in eliciting prior information. As seen in the analysis of apple data, some prior information from experts may not be very reliable.

Chapter 6

Conclusions and Further Work

6.1 The Aims of the Project

This project focused on three themes, namely, sequential designs for repeated measures data, analysis of visual assessment data and analysis of food tasting data.

First, a review of existing designs was done to identify the most suitable type for studies that involve repeated assessments of stimulus intensity. The problem with the type of design identified to be the most suitable one was that a general algebraic method of generating sequences under it is not known. This makes it difficult to recommend this design for use in experiments. Thus the aim was to find a systematic method of generating sequences for a given number of treatments.

Secondly, a continuation and improvement of the visual assessment experiments conducted by Ferris et al. (2001) was the motivation behind the experiments carried out in this project. Designs, apparatus and procedures for carrying out these experiments had to be improved. In addition, there has always been a need to find ways of reducing bias in such tasks. Subjective bias in assessments is one of the main concerns in areas such as plant breeding, because it is very important to assess the extent of damage caused by diseases correctly. Another aim of this part of the project was to suggest a method for selecting the most reliable assessors in a panel, when a linear bias correction can be applied to the responses of each of them.

Thirdly, the apple tasting data obtained from the Hannah Research Institute was to be analysed in order to be able to assess the performance of assessors over time and to adjust product effect predictions for missing assessors in experiments. It

was also important to illustrate a way of using information on assessors to analyse data from their future performances.

6.2 How the Aims were Achieved

Sequential designs balanced for order and carry-over effects were found to be the most suitable type of designs for the repeated measurement experiments in magnitude scaling studies. Type I sequential designs proposed by Finney and Outhwaite (1956) were chosen because sequences under these designs comprise all possible ordered pairs and self-adjacencies. Such designs with index 1 are particularly preferred because such balance is achieved with just a single sequence of length n^2. So, if the number of treatments to be included is large, and the index is greater than one, the sequence may become too long and impractical to use on a single subject.

A C++ program which searches for Type I designs with index 1 was written, and it produced sequences for the number of treatments between 6 and 20. The arrangement of the treatments in blocks under these designs gives approximate balance across the whole sequence, although complete orthogonality of treatment allocation to any trend cannot be achieved within the blocks. Thus, if one is interested in selecting the best of the sequences generated, the two criteria for near-orthogonality, NTF1 and the sums of squares, may be used. Also, as shown for the case of 6 treatments, sequences may be placed in groups so that only one representative from each group needs to be stored. Other members of the group may then be derived by applying the transformations of reversal and shifting or a combination of these two. What has been achieved here is a solution to the problem of generating these sequences, and also, one does not need to choose randomly among these since a near-orthogonality criterion may be used.

For the visual assessment experiments, a program was written that allowed assessors to view images on their individual monitors, and to enter their scores using the mouse pointer, while the scores were automatically saved to a file. This was a great improvement to the way these experiments had been carried out by Ferris et al. (2001), as it greatly reduced errors in data recording and bias due to varying distances between assessors and the images being presented. Carry-over from the previous stimulus level was also found to be less significant in this project than had been found before.

Bayesian predictive calibration was successfully generalised to a vector of future scores under a model with carry-over effects and auto-correlation in the errors. The type of calibration carried out here was absolute calibration, because the true count and cover levels in the calibration experiment were known. It worked successfully in cases where bias remained consistent between the calibration and future experiments. It was also seen how important it is that the assumptions of the calibration model be as valid as possible. This was seen in the calibration for the cover task. In that case, carry-over was not expected to be significant as the previous results had shown, and so calibration worked better when based on the model without carry-over effects.

The criterion for selecting the best assessors was the one proposed by Spezzaferri (1985), based on the Shannon measure of information. This was approximated by assuming normality for the predictive distribution of the future response y_* for given future stimulus x_*. MCMC methods were then used to estimate the value of the criterion. The availability of MCMC methods made it easier to estimate the criterion than what Spezzaferri (1985) had to do in the 80s. Then, a Student prior distribution was assumed for the future stimulus vector, and this was approximated by a multivariate normal distribution with the same mean vector and covariance matrix, and the resulting expression for the criterion was then evaluated by Taylor's expansion. This information criterion strongly penalised assessors whose performances were not consistently biased or variable throughout all levels of the stimulus. There is a strong correlation between this criterion and R-squared and deviance from the frequentist analysis. The difference between these is that, unlike the two frequentist criteria, the information criterion has an influence of the prior information about the assessors performance.

In analysing apple tasting data, a comparative calibration model, in contrast with absolute calibration for the visual assessments, had to be used. This was done by a Bayesian random effects multiplicative model which allowed for the quantification of true product effects through the scores entered by assessors. The differential scale use was quantified by the expansiveness coefficients and relative values of these were used to measure the assessors' discriminating ability. When data from the 45 different experiments were combined, the hierarchical nature of the combined data set and the imbalance in the number of assessors and products per experiment was modelled under the Bayesian framework. Thus, predictions of product effects were adjusted for missing assessors by borrowing strength from their data in other experiments. The change in the experiment

effects seemed to correspond to changes in posterior means of assessor effects as shown by the plots, but the experiments did not seem to be correlated in any way.

In sensory evaluation of food studies, it is often most important to determine product differences as precisely as possible. Using past information as priors for analysis of data from a similar future experiment was shown to improve precision by reducing the variance of product differences.

6.3 Further Work

With regard to sequential designs, sequences under the Type I designs with index 2 can be generated using the algebraic methods presented by Sampford (1957) and Street and Street (1987). If one is interested in finding all such sequences for a given value of n, though, the program used to search for sequences with index 1 may be extended for this purpose.

It is possible that, by looking at sequences generated by the program for various numbers of sequences, further work, involving combinatorics, might eventually reveal some algebraic method of generating the Type I sequences with index 1.

When correcting the visual assessment scores for bias, predictive calibration was shown to work well only when the nature of bias stayed constant between the calibration and future experiments. It thus seems worth doing further research on how calibration could be made robust to changes in the nature of bias.

For the apple tasting data, using past information to analyse future data was illustrated for just one future experiment. This may be done for a combined analysis of several future experiments as well as some future assessors who were not involved in the past experiments.

Some of the other apple attributes that were measured had responses which were highly skewed and had large proportions of zeros. Thus a normality assumption used for sweetness scores would not be appropriate for these, and different distributions for the responses would have to be assumed. One example would be to assume a latent Gaussian model for these.

Appendix A

Program for Systematic Search of Type I Sequences with Index 1

```
/*This program generates all possible Type I sequences according to
 the systematic procedure described in Chapter 3. The value of n may be
changed accordingly
*/
#include <math.h>
#include<iostream.h>
#include<fstream.h>

const int treatno=7;  //this is the number of treatments n
int treat[]={1,2,3,4,5,6,7}; //1...n
int bmark;   //marks beginning of block
int seq[treatno*treatno]; //stores sequence
int status=1;
int status2=1;
int goback;
int options=1;

int check1(int s[treatno*treatno],int curr,int beg,int pos);
int check2(int s[treatno*treatno],int curr,int prev,int cbmark);
void printing(int s[treatno*treatno]);

main(){
            //initialization
 for (int j=0; j<treatno*treatno;j++)
  {
```

```
  seq[j] = treat[j];
  if (j>=treatno) seq[j]=1;
 }

/* now continue from n+1 to end of sequence */

bmark=treatno;    //i.e at j=n (because vectors in C++ start at 0!)
int j=bmark;
while (j<treatno*treatno+1) //while not end of sequence
   {
    if (options==0){cout<<"endofoptions";break;}//no more options
    if (j==treatno*treatno)
      {
       int k;
       printing(seq); //print sequence obtained
       for (k=treatno*treatno-1;k<2*treatno-1;k--)
         {seq[k]=1;}   //so we'll end up at beg of n-1th block
       j=k;  // start from here
       if (seq[j]<treatno)
         {seq[j]++; status=1;} // go to TEST 1
       else
         {                  //go back twice and start from there
          for (int k=1;k<3;k++)
          {seq[j]=1; j=j-1;} //now we are 3 steps back
          status=1;   //go to TEST 1
         }
       } //end of 'if (j==treatno*treatno)'


   /*
    because c++ vectors start at 0, j mod n is not 1 so we have to say
    if j+1 mod n is 1 then j is beginning of block
   /*
     status=1;  //reset status
     if ((j+1) % treatno==1)
       {
         seq[j]=seq[j-1];  //copy prev treat to create a self-adjacency
```

```
          {                    //if self-adjacecny is repeated
            for (int k=1;k<3;k++) //go back twice changing to 1
                {seq[j]=1; j=j-1;} //so we are 3 steps back

            int go=1;
            while (go!=0)
                {
                 if(seq[j]<treatno) // then check if hit n
                    {seq[j]++;status=1;go=0; break;}//go to TEST 1
                 else
                   {
                     seq[j]=1; j=j-1;
                     if ((j+1) % treatno==1) //hit beginning of block?
                       {
                        for (int k=1;k<3;k++)//go back twice changing to 1
                           {seq[j]=1; j=j-1;} //so we are 3 steps back
                        if (j==treatno){cout<<"Endofoptions";options=0;
                                                    break;}
                         else continue;
                       } //end of if((j+1) ...

                     else {seq[j]++;status=1;break;}

                   }// end of else
                 }//end of while go
          } //end of if check2

         else
            {              //self-adjacency not repeated
              bmark=j;     //so we are fine at beg of block
              status=1;    //and move to next point
              j++;
            }
      } //end of beg of block?if((j+1) ....

/* if not at the beginning of a block, test next symbol for validity */
```

```
/*TEST 1*/
  while (status!=0) //go back to check one as long as we need to
      {
       if (check1(seq,seq[j],bmark,j) == 1) //if symbol was repeated
         {
          int go=1;
          while (go!=0)
            {
              if(seq[j]<treatno) // then check if hit n
                 {seq[j]++;status=1;go=0; break;}//go to (2)
              else
                {
                  seq[j]=1; j=j-1;
                  if ((j+1) % treatno==1) //hit beg of block?so failed
                     {
                       if (j==treatno) {cout<<"Endofoptions";status=0;
                                                 options=0; break;}
                       else
                        {
                         for (int k=1;k<3;k++)
                                  //go back twice, restoring to 1
                            {seq[j]=1; j=j-1;}//so we are 3 steps back
                         bmark=bmark-treatno; //change value of bmark
                         continue;
                        }
                     } //end of if ((j+1) % treatno==1)

                  else {status=1;continue;}
                }//else

             }//end of while (go!=0)

          }//end of if (check1(seq,seq[j....
       /* symbol not repeated so we check for pairs */
      else if (check2(seq,seq[j],seq[j-1],bmark-1) == 1)
          {                   //if self-adjacency repeated
           int go=1;
```

```
                while (go!=0)
                   {
                    if(seq[j]<treatno) //then check if reached first block
                      {seq[j]++;status=1;go=0; break;}//go to TEST 1

                    else
                      {
                       seq[j]=1; j=j-1;
                       if ((j+1) % treatno==1) //hit beg of block?so failed
                         {
                           if (j==treatno) {cout<<"Endofoptions";status=0;
                                                    options=0;break;}
                           else
                            {
                            for (int k=1;k<3;k++)//go back twice restoring to 1
                              {seq[j]=1; j=j-1;} //so we are 3 steps back
                               bmark=bmark-treatno; //change value of bmark
                               continue;
                            }
                          } //end of if ((j+1) % treatno==1)
                           else {status=1;continue;}

                        }//else
                       }//end of while (go!=0)
                     } //end of if check2

                  else  //pair not repeated, so accepted, go to beginning
                   {j++;status=0; break;}

                } //while status

      } //of while (j<treatno*treatno+1)


} //end of main
```

```
/* a function for printing out sequencese */

void printing(int s[treatno*treatno])
{
int k=1;
for (int i=0; i<treatno*treatno;i++)
     {
       cout<<s[i]<<' ';
      if ((i+1) % treatno ==0) {cout<<' ';} //spaces in between blocks
     }
 cout<<endl;
}

/* a function to check if a letter already exists */
int check1(int s[treatno*treatno],int curr,int beg,int pos)
{
  for (int k=beg;k<pos;k++) //check if symbol was repeated or not
    {
         if (s[k]==curr) return 1;
    }
      return 0;
}

/* a function to check if a pair already exists in sequence */
int check2(int s[treatno*treatno],int curr,int prev,int cbmark)
{
  for (int i=0; i<cbmark; i++)
      {
        if (s[i]==prev)
          {
            if (s[i+1]==curr)
            return 1; // pair already exists
          }
      }
  return 0;    //otherwise pair does not exist yet
}
```

Appendix B

Apple Experiments Attendance Table

Incidence matrix of attendance at tasting experiments

		Assessor Number													
	Experiment	50	52	53	55	56	62	63	64	65	66	68	69	70	71
1	ex96a078	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	ex97a006	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	ex97a011	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	ex97a014	1	1	1	1	1	1	1		1	1			1	1
5	ex97a015	1	1	1	1	1		1		1	1	1	1	1	
6	ex97a016	1	1	1	1	1		1		1	1		1	1	
7	ex97a018		1	1	1	1		1	1	1				1	1
8	ex97a026	1	1	1			1	1	1	1	1	1	1	1	
9	ex97a030	1			1	1	1	1	1	1	1	1		1	
10	ex97a032	1			1		1	1	1	1	1	1		1	1
11	ex97a035	1			1	1		1	1	1	1	1	1	1	1
12	ex97a040	1		1		1	1	1	1	1	1	1	1	1	1
13	ex97a095	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14	ex97a097	1	1	1	1	1	1	1	1	1	1	1		1	1
15	ex97a103	1		1	1		1	1	1			1	1	1	1
16	ex97a105	1		1	1	1	1	1	1	1	1	1	1	1	1
17	ex97a107	1		1	1	1	1	1	1	1	1	1		1	1
18	ex98a002	1	1	1	1	1	1	1	1	1	1	1		1	1
19	ex98a005	1	1	1	1	1		1	1	1	1	1		1	1
20	ex98a008	1	1	1	1			1	1	1	1	1		1	1
21	ex98a009	1	1	1	1			1	1	1	1	1		1	1
22	ex98a012	1	1	1	1	1	1	1	1		1	1		1	1

23	ex98a013	1	1	1	1	1	1	1	1	1	1	1	1	1
24	ex98a016		1	1	1	1	1	1	1			1	1	1
25	ex98a024	1	1	1	1	1	1		1	1	1	1	1	1
26	ex98a026	1	1	1	1	1	1		1	1	1	1	1	1
27	ex98a030	1	1	1	1	1	1	1	1	1	1	1	1	1
28	ex98a036	1	1	1	1	1	1	1	1	1	1	1	1	1
29	ex98a038	1	1	1	1	1	1	1	1		1	1	1	1
30	ex98a040	1	1	1	1	1		1	1	1		1	1	1
31	ex98a042	1	1		1	1	1		1	1	1	1	1	1
32	ex98a045	1	1		1	1	1		1	1	1	1	1	1
33	ex98a054		1	1	1	1	1		1		1	1	1	1
34	ex98a056	1	1	1	1	1	1		1	1	1	1	1	1
35	ex98a060	1	1	1	1		1		1	1		1		1
36	ex98a063	1	1	1	1	1	1	1	1	1	1	1	1	1
37	ex98a065	1	1	1	1	1	1		1	1	1	1		
38	ex98a069	1	1	1	1	1	1		1	1	1	1	1	1
39	ex98a073		1	1	1		1		1	1	1		1	1
40	ex98a079	1	1	1	1	1	1		1		1	1		1
41	ex98a082	1	1	1	1	1	1		1		1	1	1	
42	ex98a085	1	1	1	1	1	1		1	1	1	1	1	
43	ex98a091	1	1	1	1	1	1		1	1		1	1	1
44	ex98a098	1	1		1		1		1	1	1	1	1	1
45	ex98a106		1		1	1	1	1	1	1			1	1

Bibliography

Abeyasekera, S. and Curnow, R. (1984). The desirability of adjusting for residual effects in a cross-over design. *Biometrics*, 40:1071–1078.

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika*, 62:547–554.

Aitchison, J. and Dunsmore, I. R. (1975). *Statistical Prediction Analysis*. Cambridge University Press.

Bradley, R. A. and Yeh, C. (1980). Trend-free block designs: Theory. *Annals of Statistics*, 8:883–893.

Brockhoff, P. M. (1997). Statistical testing of individual differences in sensory profiling. In *5th European Conference on Food Industry and Statistics*.

Brockhoff, P. M. and Skovgaard, M. (1994). Modelling individual differences between assessors in sensory evaluations. *Food Quality and Preference*, 5:215–224.

DeCarlo, L. T. (1992). Intertrial interval and sequential effects in magnitude scaling. *Journal of Experimental Psychology: Human Perception and Performance*, 18:1080–1088.

DeCarlo, L. T. (1994). A dynamic theory of proportional judgment: Context and judgement of length, heaviness and roughness. *Journal of Experimental Psychology:Human Perception and Performance*, 20:372–381.

DeCarlo, L. T. and Cross, D. V. (1990). Sequential effects in magnitude scaling: Models and theory. *Journal of Experimental Psychology: General*, 119:375–396.

Digby, P. G. N. (1979). Modified joint regression analysis for incomplete variety by environment data. *Journal of Agricultural Science*, 93:81–86.

DiLollo, V. (1964). Contrast effects in the judgment of lifted weights. *Journal of Experimental Psychology*, 68:383–387.

Ferris, S. (1999). *Response Models and Efficient Designs for Change-Over Experiments with Treatment Carryover*. PhD thesis, University of Edinburgh.

Ferris, S. J., Kempton, R. A., Deary, I. J., Austin, E. J., and Shotter, M. V. (2001). Carryover bias in visual assessment. *Perception*, 30:1363–1373.

Finlay, K. W. and Wilkinson, G. N. (1963). The analysis of adaptation in a plant-breeding programme. *Australian Journal of Agricultural Research*, 14:742–754.

Finney, D. J. (1956). Crossover designs in bioassay. *Proceedings of the Royal Society B*, 145:42–61.

Finney, D. J. and Outhwaite, A. D. (1956). Serially balanced sequences in bioassay. *Proceedings of the Royal Society B*, 145:493–507.

Gacula, M. C. and Singh, J. (1984). *Statistical Methods in Food and Consumer Research*. London: Academic Press Inc.

Gay, C. and Mead, R. (1992). A statistical appraisal of the problem of sensory measurement. *Journal of Sensory Studies*, 7:205–228.

Geisser, S. (1971). *Foundations of Statistical Inference*. Toronto: Holt, Rinehart and Winston.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995). Efficient parametrizations for normal linear mixed models. *Biometrika*, 82:479–488.

Gescheider, G. A. (1988). Psychophysical scaling. *Annual Reviews of Psychology*, 39:169–200.

Gilks, W., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Gogel, B. J., Cullis, B. R., and Verbyla, A. P. (1995). REML estimation of multplicative effects in multi-environment variety trials. *Biometrics*, 51:744–749.

Hocking, R. R. (2003). *Methods and Applications of Linear Models*. John Wiley and Sons, New Jersey, second edition.

Krantz, J. and Rotem, J., editors (1988). *Experimental techniques in Plant Disease Epidemiology*. Springer-Verlag, Heidelberg.

Krueger, L. E. (1972). Perceived numerosity. *Perception and Psychophysics*, 11:5–9.

Krueger, L. E. (1982). Single judgment of numerosity. *Perception and Psychophysics*, 31:175–182.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86.

Laming, D. (1995). Screening cervical smears. *British Journal of Psychology*, 86:507–516.

Lindley, D. V. (1956). On a measure of the information provided by the experiment. *Annals of Mathematical Statistics*, 27(4):986–1005.

Lockhead, G. R. and King, M. C. (1983). A memory of sequential effects in scaling tasks. *Journal of Experimental Psychology*, 9(3):461–473.

Mandel, J. (1971). A new analysis of variance model for non-additive data. *Technometrics*, 13:1–18.

McEwan, J. A., Hunter, E. A., van Gemert, L. J., and Leah, P. (2002). Proficiency testing for sensory profile panels: measuring panel performance. *Food Quality and Preference*, 13:181–190.

Mead, R. and Gay, C. (1995). Sequential design of sensory trials. *Food Quality and Preference*, 6:271–280.

Morris, R. B. and Rule, S. J. (1988). Sequential judgement effects in magnitude estimation. *Canadian Journal of Psychology*, 42:69–77.

Nabugoomu, F., Kempton, R. A., and Talbot, M. (1999). Analysis of series of trials where varieties differ in sensitivity to locations. *Journal of Agricultural, Biological and Environmental Statistics*, 4(3):310–325.

Naes, T. (1998). Detecting individual differences among assessors and differences among replicates in sensory profiling. *Food Quality and Preference*, 9(3):107–110.

Nutter, F. W., Gleason, M. L., Jenco, J. H., and Christians, N. C. (1993). Assessing the accuracy, intra-rater repeatability and inter-rater reliability of disease assessment systems. *Phytopathology*, 83(8):806–812.

Nutter, F. W. and Guan, J. (2002). Quantifying alfalfa yield losses caused by foliar disease in Iowa, Ohio, Wisconsin, and Vermont. *Plant Disease*, 86(3):269–277.

Nutter, F. W. and Schultz, P. M. (1995). Improving the accuracy and precision of disease assessments - selection of methods and use of computer-aided training programs. *Canadian Journal of Plant Pathology*, 17(2):174–184.

Oman, S. D. (1991). Multiplicative effects in mixed model analysis of variance. *Biometrika*, 78(4):729–739.

Osborne, C. (1991). Statistical calibration: A review. *International Statistical Review*, 59:309–336.

Parker, S. R., Shaw, M. W., and Royle, D. J. (1995). The reliability of visual estimates of disease severity on cereal leaves. *Plant Pathology*, 44:856–864.

Piepho, H. P. (1997). Analyzing genotype-environment data by mixed models with multiplicative terms. *Biometrics*, 53:761–766.

Piggot, J. R., editor (1988). *Sensory Analysis of Foods*. London: Elsevier, 2nd edition.

Piggot, J. R. and Hunter, E. A. (1999). Evaluation of assessor performance in sensory analysis. *Italian Journal of Food Science*, 11:289–303.

Roberts, H. V. (1965). Probabilistic prediction. *Journal of the American Statistical Society*, 60:50–62.

Rossi, F. (2001). Assessing sensory panelist performance using repeatability and reproducibility measures. *Food Quality and Preference*, 12:467–479.

Sampford, M. R. (1957). Methods of construction and analysis of serially balanced sequences. *Journal of the Royal Statistical Society B*, 19:286–304.

Sawyer, T. F. and Wesenstein, N. J. (1994). Anchoring effects on judgment, estimation and discrimination of numerosity. *Perceptual and Motor Skills*, 78:91–98.

Schifferstein, H. N. J. and Oudejans, I. M. (1996). Determinants of cumulative successive contrast in saltiness intensity judgments. *Perception and Psychophysics*, 58(5):713–724.

Shannon, C. E. and Weaver, W. (1963). *The Mathematical Theory of Communication*. The University of Illinois Press.

Shaw, M. B. and Royle, D. J. (1989). Estimation and validation of a function describing the rate at which mycosphaerella graminicola causes yield loss in winter wheat. *Annals of Applied Biology*, 115:425–442.

Sherwood, R. T., Berg, C. C., Hoover, M. R., and Zeiders, K. E. (1983). Illusions in visual assessment of stagonospora leaf spot of orchardgrass. *Phytopathology*, 73(2):173–177.

Smith, A., Cullis, B., Brockhoff, P., and Thompson, R. (2003). Multiplicative mixed models for the analysis of sensory evaluation data. *Food Quality and Preference*, 14:387–395.

Spezzaferri, F. (1985). A note on multivariate calibration experiments. *Biometrics*, 41:267–272.

Spiegelhalter, D. J., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistal Society B*, 64:583–640.

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Gilks, W. (1996). *WinBUGS: Bayesian Inference Using Gibbs Sampling*. MRC Biostatistics Unit, Cambridge, 1.4 edition.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, 64:153–181.

Stone, H. and Sidel, J. L. (1993). *Sensory Evaluation Practices*. Redwood City California: Academic Press.

Street, A. P. and Street, D. J. (1987). *Combinatorics of Experimental Design*. Oxford Science Publications.

Theobald, C. M. and Mallinson, J. R. (1978). Comparative calibration, linear structural relationships and congeneric measurements. *Biometrics*, 34:39–45.

Theobald, C. M., Talbot, M., and Nabugoomu, F. (2002). A bayesian approach to regional and local-area prediction from crop variety trials. *Journal of Agricultural, Biological and Environmental Statistics*, 7(3):403–419.

Tomerlin, J. R. and Howell, T. A. (1988). DISTRAIN: A computer program for training people to estimate disease severity on cereal leaves. *Plant Disease*, 72:455–459.

Williams, E. J. (1949). Environmental designs balanced for the estimation of residual effects of treatments. *Australian Journal of Scientific Research*, 2(2):149–168.

Williams, R. M. (1952). Experimental designs for serially correlated observations. *Biometrika*, 39:151–167.

Yates, F. and Cochran, W. (1938). The analysis of groups of experiments. *Journal of Agricultural Science*, 28:556–580.

Yeh, C., Bradley, R. A., and Notz, W. I. (1985). Nearly trend-free block designs. *Journal of the American Statistical Association*, 80(392):985–992.