*Journal of Agricultural Science* (2003), **140**, 317–333. © 2003 *Cambridge University Press*
DOI: 10.1017/S0021859603002983 *Printed in the United Kingdom*


# Assessment of observer performance in a subjective scoring system: visual classification of the gait of cows

B. ENGEL[1]*, G. BRUIN[1], G. ANDRE[2] AND W. BUIST[1]

[1] *Institute for Animal Science and Health, ID-Lelystad, P.O. Box 65, 8200 AB Lelystad, The Netherlands*
[2] *Research Institute for Animal Husbandry, P.O. Box 2176, 8203 AD Lelystad, The Netherlands*

(*Revised MS received 29 November 2002*)

## SUMMARY

As with any measurement procedure, the performance of a subjective classification procedure must be evaluated. Observers have to be trained and their performance has to be assessed, preferably on a regular basis, to guarantee sufficient consistency and accuracy of classification results. The current paper is a study of observer performance where observers were asked to classify the gait of cows from video recordings. Gait was classified in nine ordered categories (ranging from 1 = normal gait to 9 = severely abnormal gait) and also as a continuous fraction by putting a mark on a paper strip (the left end corresponding to 0 = normal gait and the right end to 1 = severely abnormal gait). The use of statistical models and methodology for analysis of these visual scores is demonstrated and discussed. Observers were assessed by comparing their classification results with the results of an expert. Models and methodology take proper account of typical features of the data, i.e. the fact that data are discrete scores or continuous scores with an upper and lower bound, the variance heterogeneity and non-linearity of model terms that arises from this, and the dependence between repeated classifications of videos of the same cow. Results of the analyses are summarized in simple tables and plots. These are useful tools to indicate possible flaws in judgement of an observer, that may be corrected by further training. When a high standard is developed, which usually takes the form of the opinion of one or more experts, this methodology can be applied prior to any experiment where responses are ordered subjective scores.

## INTRODUCTION

The literature on gait, posture and locomotion of animals is largely concerned with the association between scores for gait and other traits, such as reproductive performance (Sprecher *et al*. 1997), or with the influence of treatments, such as different amounts of concentrates (Manson & Leaver 1988), on scores for gait. In the current paper, attention is focused on the reliability of the scoring system. As part of an animal welfare study of the Institute for Animal Science and Health and the Research Institute for Animal Husbandry in The Netherlands, the gait of cows was visually assessed at several farms. Subsequently, to evaluate observer performance, nine observers were asked to classify video recordings of the gait of 50 different cows. Assessment of the observers on the basis of their classification results of the video recordings is the subject of the current paper.

All video recordings were classified five times in five successive sessions by each observer. To form an impression of the effect of further training, after discussion with an expert observer, the videos were classified once more, again five times in five sessions. Visual assessment was performed in two ways: by classification into nine discrete ordered categories (from 1 = normal to 9 = severely abnormal) and as a continuous fraction (from 0 = normal to 1 = severely abnormal) by putting a mark on a paper strip (left end = 0 = normal and right end = 1 = severely abnormal). The aforementioned sequence of five sessions/discussion/five sessions was followed for the discrete scores (1, …, 9) first, and for the continuous scores (fractions) afterwards, employing the same video recordings. In all sessions the video recordings were offered in a (different) random order.

* To whom all correspondence should be addressed, at Biometris, P.O. Box 100, 6700 AC Wageningen, The Netherlands. Email: Bas.Engel@wur.nl

Although five repeat sessions to classify 50 videos per session puts a strain on observers, the resulting data set is moderately sized for current purposes and results based on model calculations are to be preferred over tables of raw means. In the choice of model the fact that the data are either discrete scores or fractions, with associated non-linearity of model terms and heterogeneity of variances, and that repeated observations of the same cow will be correlated, had to be accounted for. For each observer separate analyses of the data collected before and after further training were performed, i.e. discussion with the expert. Simultaneous analysis of data, for instance analysis of scores from several observers, would be feasible. However, such an analysis is more complex while little extra information is extracted from the data. Discrete scores (1, …, 9) and continuous scores (fractions) were analysed separately. The statistical analyses offered useful tools, in the form of tables and plots, to evaluate observer performance.

The expert's results were used as a 'gold standard', i.e. as a measure of the true score for a video recording. In a discussion between expert and observers, after classification was finished, there was general agreement on the expert's discrete scores in ordered categories for all cows except one. For the continuous fractions (paper strip scores), the mean of the expert was taken as the gold standard. This seemed a reasonable choice since the expert was far more experienced than the other observers. Moreover, to ensure that the expert mean was sufficiently stable, it was derived from 10 repeated classifications of each video recording. For scoring systems where no gold standard is available, attention generally focuses on measures of variation within (repeatability) and between (reproducibility) observers (Garner *et al*. 2002).

In the next section, first an overview of the data is presented. Subsequently the model for the discrete scores (1, …, 9) is introduced. Scores correspond to intervals defined by threshold values on an underlying continuous scale. This threshold model is a particular instance of a generalized linear mixed model (GLMM) and methodology developed by Keen & Engel (1997) for GLMMs for ordered categorical data is used. Next, the model for the continuous scores (fractions) is presented. This is also a GLMM and methodology developed by Engel & Keen (1994) is used. Plots and tests for model checking are discussed. Some results are briefly presented in passing in the Materials and Methods section to illustrate the methodology. Results are discussed in more detail in the Results section. However, since the main aim is to demonstrate that the analyses offer useful tools to assess the observers, a detailed discussion of all results is not presented. Finally, in the discussion some practical aspects, such as implementation in a monitoring system, are addressed. Most of the technical details are transferred to the Appendix. All the calculations were performed with the statistical package GenStat (GenStat Committee 2000).

## MATERIALS AND METHODS

### *The classification data*

The scoring system in nine ordered categories is described in the Appendix of Manson & Leaver (1988). The scores in Manson and Leaver are numbered 1·0, 1·5, 2·0, …, 4·5, 5·0 and were re-numbered 1, …, 9 in this paper. A summary of the discrete scores is presented in Table 1 *a* for all observers including the expert observer. Score 9 (a severely abnormal gait) did not occur. Therefore, scores 8 and 9 were pooled and the analyses were effectively reduced to scores 1, …, 8. A summary of the continuous scores (fractions) is presented in Table 1 *b*.

### *A threshold model for the discrete scores*

An analysis of variance (ANOVA) model with the expert scores as levels of an explanatory factor was adopted. However, an ANOVA model is appropriate for a continuous response variable but not for a discrete score $y$. Therefore an underlying continuous variable $z$, that follows an ANOVA model, and is connected to an observer's score $y$ by a threshold concept was introduced. A particular score is assigned when the underlying variable is in the corresponding interval defined by two successive threshold values:

$$y=k \quad \text{when } \theta_{k-1}<z\leqslant\theta_k,\ k=1,\ldots,K$$

Here, $\theta_0<\theta_1<\cdots<\theta_K$ are the threshold values. Threshold values $\theta_0$ and $\theta_K$ are equal to $-\infty$ and $+\infty$ respectively; they were introduced to simplify the notation. The number of scores $K=8$ (because score 9 did not occur in the data). It was assumed that $z$ is normally distributed with mean $\eta$, that depends on the expert score and the cow, and constant variance $\sigma^2$. The expert score will be denoted by $x$. For a cow with expert score $x=j$, $j=1,\ldots,K$:

$$\eta=\beta_j+u$$

Here, $\beta_j$ is the observer's mean for cows with expert score $x=j$ and $u$ is the departure from that mean for a particular cow on the underlying scale. The cow effects $u$ are assumed to be normally distributed with mean 0 and variance $\sigma_u^2$. They induce a (positive) correlation between scores of the same cow. Note that cows have to be representative for their expert scores $x$, but may be selected on the basis of $x$, for instance with over-sampling of the extremes.

The model is illustrated in Fig. 1, where the shaded area equals the probability for a score $y=3$ for a particular cow. With a different expert score or a different cow effect, the distribution in Fig. 1 may move

Table 1 *a*. *A summary of the discrete scores 1, …, 9* (*columns*) *per observer* (*rows*) *before* (*B*) *and after* (*A*) *further training. For observer 9 there was only one series of five repeats of 50 videos. The expert was quite familiar with the videos and performed no repeat classifications for the discrete scores*

| Observer | Before/ after | Percentages for discrete scores | | | | | | | | | Number of observations |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | |
| 1 | B | 5·2 | 14·0 | 26·4 | 13·6 | 12·0 | 6·0 | 11·2 | 11·6 | 0 | 250 |
| | A | 4·0 | 16·0 | 24·4 | 13·2 | 13·6 | 8·0 | 5·2 | 15·6 | 0 | 250 |
| 2 | B | 10·0 | 14·0 | 25·6 | 24·8 | 12·0 | 7·2 | 4·4 | 2·0 | 0 | 250 |
| | A | 8·8 | 7·2 | 19·6 | 22·8 | 15·2 | 8·4 | 11·6 | 6·4 | 0 | 250 |
| 3 | B | 2·0 | 8·0 | 30·8 | 19·2 | 14·0 | 5·6 | 9·2 | 11·2 | 0 | 250 |
| | A | 6·0 | 11·2 | 20·4 | 24·4 | 7·6 | 5·2 | 11·6 | 13·6 | 0 | 250 |
| 4 | B | 5·6 | 8·0 | 21·2 | 23·2 | 11·2 | 14·0 | 10·8 | 6·0 | 0 | 250 |
| | A | 4·4 | 13·2 | 21·6 | 15·6 | 16·0 | 6·4 | 14·4 | 8·4 | 0 | 250 |
| 5 | B | 9·2 | 17·6 | 16·0 | 11·2 | 14·4 | 13·6 | 10·0 | 8·0 | 0 | 250 |
| | A | 6·0 | 10·8 | 30·0 | 14·0 | 11·2 | 8·8 | 8·4 | 10·8 | 0 | 250 |
| 6 | B | 4·8 | 18·4 | 19·6 | 12·8 | 15·6 | 14·8 | 8·8 | 5·2 | 0 | 250 |
| | A | 0 | 11·6 | 29·6 | 13·2 | 14·8 | 6·8 | 9·6 | 14·4 | 0 | 250 |
| 7 | B | 3·6 | 20·8 | 17·2 | 15·2 | 14·0 | 4·8 | 16·0 | 8·4 | 0 | 250 |
| | A | 4·0 | 28·0 | 19·2 | 7·6 | 7·2 | 6·0 | 12·8 | 15·2 | 0 | 250 |
| 8 | B | 5·6 | 20·4 | 25·6 | 16·0 | 8·6 | 12·8 | 4·8 | 6·0 | 0 | 250 |
| | A | 6·0 | 14·8 | 25·2 | 24·8 | 9·2 | 6·0 | 6·8 | 7·2 | 0 | 250 |
| 9 | — | 5·2 | 11·6 | 19·2 | 20·4 | 16·8 | 6·0 | 15·6 | 5·2 | 0 | 250 |
| Expert | — | 6·0 | 16·0 | 18·0 | 16·0 | 16·0 | 8·0 | 12·0 | 8·0 | 0 | 50 |

Table 1 *b*. *A summary of the paper strip scores per observer* (*rows*) *before* (*B*) *and after* (*A*) *further training. A few scores are missing*

| Observer | Before/after | Percentile points | | | Number of observations (median) |
|---|---|---|---|---|---|
| | | 10 % | 50 % | 90 % | |
| 1 | B | 0·14 | 0·50 | 0·88 | 247 |
| | A | 0·20 | 0·50 | 0·90 | 250 |
| 2 | B | 0·03 | 0·54 | 0·78 | 247 |
| | A | 0·09 | 0·50 | 0·83 | 250 |
| 3 | B | 0·08 | 0·49 | 0·94 | 247 |
| | A | 0·05 | 0·52 | 0·95 | 250 |
| 4 | B | 0·24 | 0·54 | 0·86 | 247 |
| | A | 0·19 | 0·51 | 0·89 | 250 |
| 5 | B | 0·12 | 0·42 | 0·83 | 250 |
| | A | 0·12 | 0·46 | 0·87 | 250 |
| 6 | B | 0·16 | 0·53 | 0·90 | 250 |
| | A | 0·09 | 0·47 | 0·92 | 250 |
| 7 | — | 0·07 | 0·35 | 0·80 | 250 |
| 8 | B | 0·10 | 0·40 | 0·75 | 250 |
| | A | 0·04 | 0·45 | 0·75 | 250 |
| 9 | B | 0·21 | 0·44 | 0·78 | 250 |
| | A | 0·15 | 0·44 | 0·80 | 250 |
| Expert | — | 0·12 | 0·37 | 0·79 | 500 |

either to the left, which increases the probabilities for low scores, or to the right, which increases the probabilities for high scores.

Scores $y$ do not change when $z$ and $\theta_1, \ldots, \theta_{K-1}$ are divided by an arbitrary non-zero constant, nor will they change when an arbitrary constant is added. Therefore, without loss of generality, but merely to pin down the location and scale for $z$, it was assumed that the underlying residual variance $\sigma^2 = 1$ and threshold $\theta_1 = 0$. The other thresholds $\theta_2, \ldots, \theta_{K-1}$ together with $\beta_1, \ldots, \beta_K$ and $\sigma_u^2$ will have to be estimated from the data.

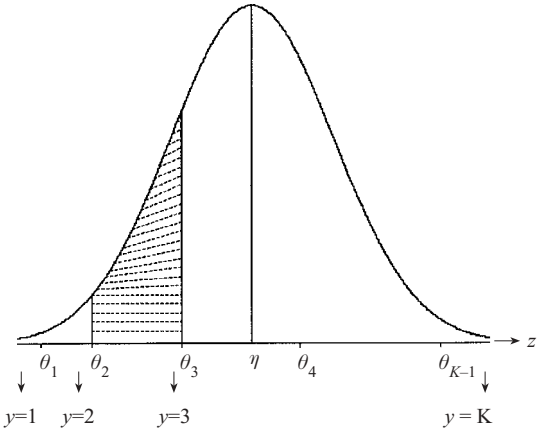


Fig. 1. The threshold model.

The probability for a score equal or below $k$ for a cow with expert score $x=j$ is

$$\gamma_{jk} = P(y \leqslant k | x=j) = P(z \leqslant \theta_k | x=j) = \Phi\left(\frac{\theta_k - \beta_j}{\sqrt{(\sigma_u^2+1)}}\right) \quad (1)$$

where $\Phi$ denotes the cumulative standard normal distribution function (the standard normal probability integral) (some details are in Appendix A1). Note that $\Phi$ is available in most statistical packages, including the GenStat package used in the current paper. The probability for a score equal to $k$ is

$$\Pi_{jk} = P(y=k|x=j) = P(\theta_{k-1} < z \leqslant \theta_k | x=j) = \gamma_{jk} - \gamma_{jk-1} \quad (2)$$

Estimates and associated confidence intervals for the probabilities $\Pi_{jk}$ were used to evaluate observer performance. The model is a particular instance of a generalized linear mixed model; see Keen & Engel (1997) for additional details and references.

With an underlying logistic distribution, and without random cow effects, the threshold model is also known as the proportional odds model; some details are in Chapter 5 of McCullagh & Nelder (1989). An underlying normal distribution was preferred, because in conjunction with normally distributed random cow effects this yields the relatively simple expression (1) for cumulative probabilities $\gamma_{jk}$. For an underlying logistic distribution, expressions for $\gamma_{jk}$ are only available in an approximate form (Aitchison & Shen 1980; Engel *et al.* 1995).

### *Inference with the threshold model*

Threshold values $\theta_2, \ldots, \theta_{K-1}$, means $\beta_1, \ldots, \beta_K$ for the underlying variable $z$ and variance component $\sigma_u^2$ for variation between cows were estimated by the method described in Keen & Engel (1997), utilizing procedure IRCLASS (Keen 2001), which is written in GenStat. With these parameter estimates, estimated probabilities $\hat{\Pi}_{jk}$ were derived from expressions (1) and (2).

Standard errors and 95% confidence intervals for probabilities $\Pi_{jk}$ were calculated by a parametric bootstrap method (Efron & Tibshirani 1993). With the estimated parameter values, 500 new data sets of 250 scores each were generated by simulation and analysed. This resulted in 500 estimated values $\hat{\Pi}_{jkl}$, $l=1, \ldots, 500$, for each of the probabilities $\Pi_{jk}$. The standard deviations of these 500 simulated estimates provided the estimated standard errors of the original estimates $\hat{\Pi}_{jk}$. Some details about the calculations of the confidence intervals are given in Appendix A1. As an illustration, estimated probabilities and associated standard errors and intervals are shown in Table 2 for observer 4 before and after further training, i.e. before and after a discussion with the expert. Rows correspond to expert scores and columns to the scores of observer 4. For instance, before training there was only a probability $\hat{\Pi}_{55}=0{\cdot}34$ for observer 4 to classify a cow with expert score 5 in that same category. There was a probability of $\hat{\Pi}_{54}+\hat{\Pi}_{56}=0{\cdot}26+0{\cdot}32=0{\cdot}58$ that the cow was classified in a neighbouring category. Even when the lower bounds of the confidence intervals are used, offering an optimistic view of the differences with the expert, the latter probability was still sizeable.

Simple summary statistics, such as a weighted mean of the diagonal elements

$$\Sigma_j w_j \hat{\Pi}_{jj}$$

can be calculated from the estimated probabilities, with standard errors and confidence intervals derived from the bootstrap sample. Weights may be chosen equal to the population probabilities $P(x=j)$ for the true scores or by weighing the consequences of misclassification. Since neither the probabilities $P(x=j)$, nor the full consequences of misclassification were known, in this paper equal weights $w_j=1/K$ were employed.

### *A model for the continuous fractions*

A regression model for the observed fraction $y$ of an observer on the true fraction $x$ was adopted, i.e. the mean fraction $x$ of the expert, with some extra features because the scores are between 0 and 1. In contrast to ordinary linear regression it could not simply be assumed that the mean $\mu$ of $y$ is a linear function of $x$, because that may produce fitted values for $\mu$ below 0 or above 1. Therefore, the means $\mu$ were 'stretched' by a transformation to values ranging from $-\infty$ to $+\infty$. A popular transformation to

Table 2. *The estimated probabilities for observer 4, before and after further training. Standard errors are in parentheses and 95% confidence intervals in square brackets. Row numbers are expert scores and column numbers are scores for observer 4*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Before further training | | | | | | | | |
| 1 | 0·39 (0·08) | 0·35 (0·06) | 0·23 (0·06) | 0·03 (0·02) | 0·00 (0·00) | 0·00 (0·00) | 0·00 (0·00) | 0·00 (0·00) |
| | [0·25, 0·55] | [0·24, 0·47] | [0·14, 0·36] | [0·01, 0·10] | [0·00, 0·00] | [0·00, 0·00] | [0·00, 0·00] | [0·00, 0·00] |
| 2 | 0·17 (0·04) | 0·31 (0·05) | 0·39 (0·05) | 0·12 (0·03) | 0·01 (0·00) | 0·00 (0·00) | 0·00 (0·00) | 0·00 (0·00) |
| | [0·10, 0·28] | [0·22, 0·43] | [0·31, 0·48] | [0·07, 0·21] | [0·00, 0·02] | [0·00, 0·00] | [0·00, 0·00] | [0·00, 0·00] |
| 3 | 0·02 (0·01) | 0·09 (0·02) | 0·37 (0·04) | 0·43 (0·04) | 0·08 (0·02) | 0·01 (0·01) | 0·00 (0·00) | 0·00 (0·00) |
| | [0·00, 0·05] | [0·05, 0·16] | [0·29, 0·46] | [0·34, 0·51] | [0·04, 0·16] | [0·00, 0·05] | [0·00, 0·00] | [0·00, 0·00] |
| 4 | 0·01 (0·00) | 0·06 (0·02) | 0·32 (0·04) | 0·47 (0·05) | 0·11 (0·03) | 0·02 (0·01) | 0·00 (0·00) | 0·00 (0·00) |
| | [0·00, 0·03] | [0·03, 0·13] | [0·24, 0·41] | [0·37, 0·57] | [0·06, 0·19] | [0·01, 0·08] | [0·00, 0·00] | [0·00, 0·00] |
| 5 | 0·00 (0·00) | 0·00 (0·00) | 0·03 (0·01) | 0·26 (0·04) | 0·34 (0·06) | 0·32 (0·05) | 0·05 (0·02) | 0·00 (0·00) |
| | [0·00, 0·00] | [0·00, 0·00] | [0·01, 0·08] | [0·18, 0·36] | [0·24, 0·45] | [0·23, 0·42] | [0·02, 0·12] | [0·00, 0·02] |
| 6 | 0·00 (0·00) | 0·00 (0·00) | 0·00 (0·00) | 0·04 (0·02) | 0·15 (0·04) | 0·46 (0·06) | 0·27 (0·05) | 0·08 (0·03) |
| | [0·00, 0·00] | [0·00, 0·00] | [0·00, 0·01] | [0·01, 0·11] | [0·09, 0·26] | [0·35, 0·57] | [0·18, 0·38] | [0·03, 0·19] |
| 7 | 0·00 (0·00) | 0·00 (0·00) | 0·00 (0·00) | 0·01 (0·01) | 0·08 (0·03) | 0·38 (0·05) | 0·35 (0·06) | 0·17 (0·05) |
| | [0·00, 0·00] | [0·00, 0·00] | [0·00, 0·00] | [0·00, 0·05] | [0·04, 0·16] | [0·28, 0·49] | [0·25, 0·47] | [0·09, 0·29] |
| 8 | 0·00 (0·00) | 0·00 (0·00) | 0·00 (0·00) | 0·00 (0·00) | 0·02 (0·01) | 0·19 (0·05) | 0·37 (0·06) | 0·43 (0·07) |
| | [0·00, 0·00] | [0·00, 0·00] | [0·00, 0·00] | [0·00, 0·01] | [0·00, 0·06] | [0·11, 0·31] | [0·26, 0·49] | [0·30, 0·56] |
| After further training | | | | | | | | |
| 1 | 0·24 (0·05) | 0·54 (0·05) | 0·22 (0·04) | 0·01 (0·00) | 0·00 (0·00) | 0·00 (0·00) | 0·00 (0·00) | 0·00 (0·00) |
| | [0·16, 0·34] | [0·44, 0·62] | [0·15, 0·31] | [0·00, 0·02] | [0·00, 0·00] | [0·00, 0·00] | [0·00, 0·00] | [0·00, 0·00] |
| 2 | 0·16 (0·03) | 0·52 (0·04) | 0·31 (0·03) | 0·02 (0·01) | 0·00 (0·00) | 0·00 (0·00) | 0·00 (0·00) | 0·00 (0·00) |
| | [0·10, 0·22] | [0·43, 0·60] | [0·25, 0·38] | [0·01, 0·04] | [0·00, 0·01] | [0·00, 0·00] | [0·00, 0·00] | [0·00. 0·00] |
| 3 | 0·00 (0·00) | 0·09 (0·02) | 0·52 (0·04) | 0·24 (0·04) | 0·13 (0·02) | 0·02 (0·01) | 0·00 (0·00) | 0·00 (0·00) |
| | [0·00, 0·01] | [0·05, 0·14] | [0·45, 0·59] | [0·18, 0·31] | [0·08, 0·19] | [0·01, 0·05] | [0·00, 0·02] | [0·00, 0·00] |
| 4 | 0·00 (0·00) | 0·05 (0·02) | 0·45 (0·03) | 0·28 (0·04) | 0·17 (0·03) | 0·03 (0·01) | 0·01 (0·00) | 0·00 (0·00) |
| | [0·00, 0·00] | [0·03, 0·10] | [0·39, 0·52] | [0·21, 0·35] | [0·12, 0·24] | [0·01, 0·08] | [0·00, 0·03] | [0·00, 0·00] |
| 5 | 0·00 (0·00) | 0·00 (0·00) | 0·08 (0·02) | 0·18 (0·03) | 0·36 (0·06) | 0·19 (0·04) | 0·17 (0·03) | 0·02 (0·01) |
| | [0·00, 0·00] | [0·00, 0·00] | [0·04, 0·14] | [0·12, 0·25] | [0·26, 0·47] | [0·13, 0·28] | [0·11, 0·26] | [0·01, 0·05] |
| 6 | 0·00 (0·00) | 0·00 (0·00) | 0·01 (0·00) | 0·04 (0·01) | 0·18 (0·04) | 0·20 (0·04) | 0·42 (0·05) | 0·15 (0·04) |
| | [0·00, 0·00] | [0·00, 0·00] | [0·00, 0·02] | [0·01, 0·09] | [0·12, 0·26] | [0·13, 0·30] | [0·33, 0·52] | [0·09, 0·26] |
| 7 | 0·00 (0·00) | 0·00 (0·00) | 0·00 (0·00) | 0·02 (0·01) | 0·14 (0·03) | 0·18 (0·04) | 0·45 (0·05) | 0·20 (0·04) |
| | [0·00, 0·00] | [0·00, 0·00] | [0·00, 0·01] | [0·01, 0·06] | [0·09, 0·21] | [0·12, 0·27] | [0·35, 0·54] | [0·13, 0·29] |
| 8 | 0·00 (0·00) | 0·00 (0·00) | 0·00 (0·00) | 0·00 (0·00) | 0·02 (0·01) | 0·05 (0·02) | 0·32 (0·04) | 0·61 (0·05) |
| | [0·00, 0·00] | [0·00, 0·00] | [0·00, 0·00] | [0·00, 0·01] | [0·01, 0·05] | [0·02, 0·10] | [0·25, 0·41] | [0·50, 0·70] |

achieve this is the logit transformation: $\text{logit}(\mu) = \log(\mu/(1-\mu))$. A transformation that usually produces similar results and in this case is mathematically more convenient (Appendix A2) is the probit: $\text{probit}(\mu) = \Phi^{-1}(\mu)$, where $\Phi^{-1}$ denotes the inverse of the cumulative standard normal distribution function:

$$\text{probit}(\mu) = \alpha + \beta\,\text{probit}(x) + u$$

or

$$\mu = \Phi(\alpha + \beta\,\text{probit}(x) + u) \qquad (3)$$

Here, probit($x$) is the probit transform of the true fraction $x$ and $u$ is a normally distributed cow effect with mean 0 and variance $\sigma_u^2$. Some of the notation is similar to the threshold model, but (the smaller number of) parameters have a different interpretation. Unlike in ordinary linear regression, it could not be assumed that the variance of $y$ is constant and independent of the mean $\mu$. For more extreme $\mu$, observations $y$ will be closer to 0 or 1, and the variation will be relatively smaller. So, a large variation in the middle of the interval (0, 1) and small variation at the extremes 0 and 1 was expected. This is illustrated in Fig. 2, where the raw variance estimates are plotted against the means per cow for the data of observer 4 after further training. Despite the 'noise' in the variance estimates, because each variance is based on five repeats only, the expected pattern is clearly there. It was assumed that

$$\text{Var}(y) = \sigma^2\mu(1-\mu)$$

where $\sigma^2$, to be referred to as the residual variance component, will be estimated from the data. Figure 2 also illustrates that there is no gain in examining a more intricate variance-mean relationship: it would not be possible to see the difference with the simple variance function assumed above.

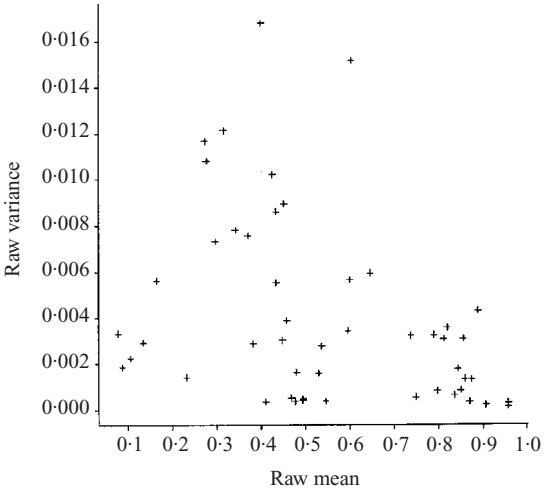


Fig. 2. Raw variances against raw means per cow for continuous scores (fractions). Observer 4 after training.

The population average over all cows with expert score $x$, i.e. the expectation of $y$ for a given expert score $x$, is equal to

$$E(y|x)=\Phi\left(\frac{\alpha+\beta\,\mathrm{probit}(x)}{\sqrt{(\sigma_u^2+1)}}\right)$$
$$=\Phi(a+b\,\mathrm{probit}(x)) \qquad (4)$$

Note that compared with Eqn (3), averaging over the cows (integration over $u$ in mathematical terms) involves shrinkage of the intercept and slope:

$$a=\alpha/\sqrt{(\sigma_u^2+1)} \quad \text{and} \quad b=\beta/\sqrt{(\sigma_u^2+1)}$$

There are some details in Appendix A2.

*Inference with the model for fractions*

The intercept $\alpha$, slope $\beta$, and variance components $\sigma_u^2$ and $\sigma^2$ for variation between and within cows were estimated by a quasi-likelihood method as described in Engel & Keen (1994), utilizing GenStat procedure IRREML (Keen & Engel 2001). Confidence bounds for expected fractions from expression (4) were derived from a bootstrap sample. To simulate data, it was necessary to be more specific about the distribution of fractions $y$ given a particular expert score $x$ and cow effect $u$. A beta distribution was assumed. Some details of the simulation are given in Appendix A2. By way of an illustration, in Fig. 3 the difference between the expected scores of observer 4 and the true score $x$ (the expert's mean score), before and after training, are plotted against $x$. Ideally the curves should be close to 0 for all $x$. The 95% confidence bounds in Fig. 3 were constructed in the same way as described in Appendix A1 for probabilities in the threshold model. They offer a more discriminating

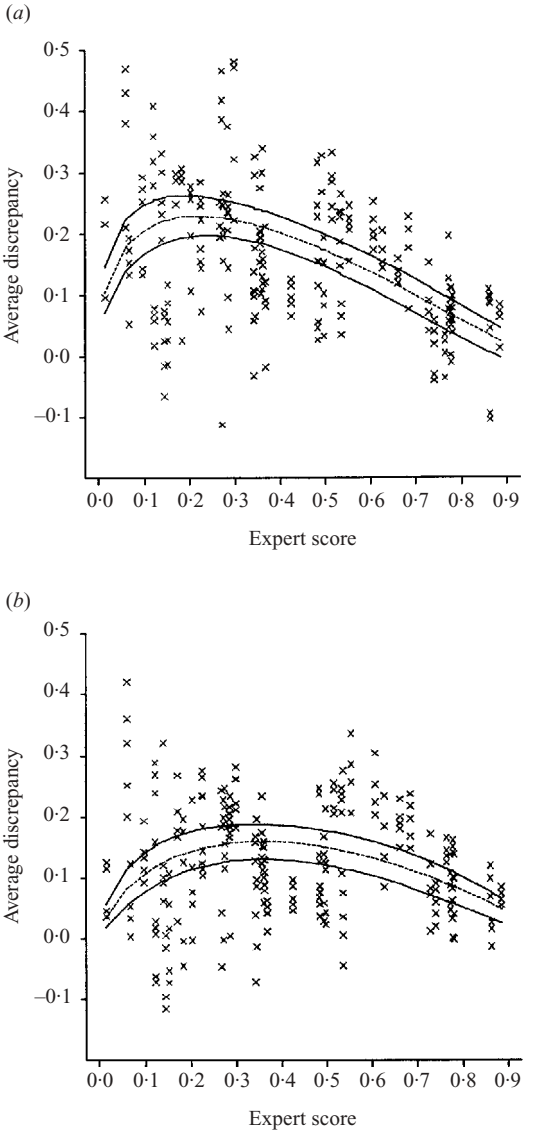


Fig. 3. The average discrepancy between observer 4 and the expert plotted against the expert score with 95% confidence limits for continuous scores (fractions). Curves (*a*) before and (*b*) after further training.

view of the differences between expert and observer than mere point estimates. Systematic differences between observer 4 and the expert are smaller after further training. However, both before and after training observer 4 scored significantly higher than the expert.

A simple summary statistic to compare the expert and an observer is the root of the expected mean squared error:

$$\mathrm{RMSE}(x)=\sqrt{E(y-x)^2}$$

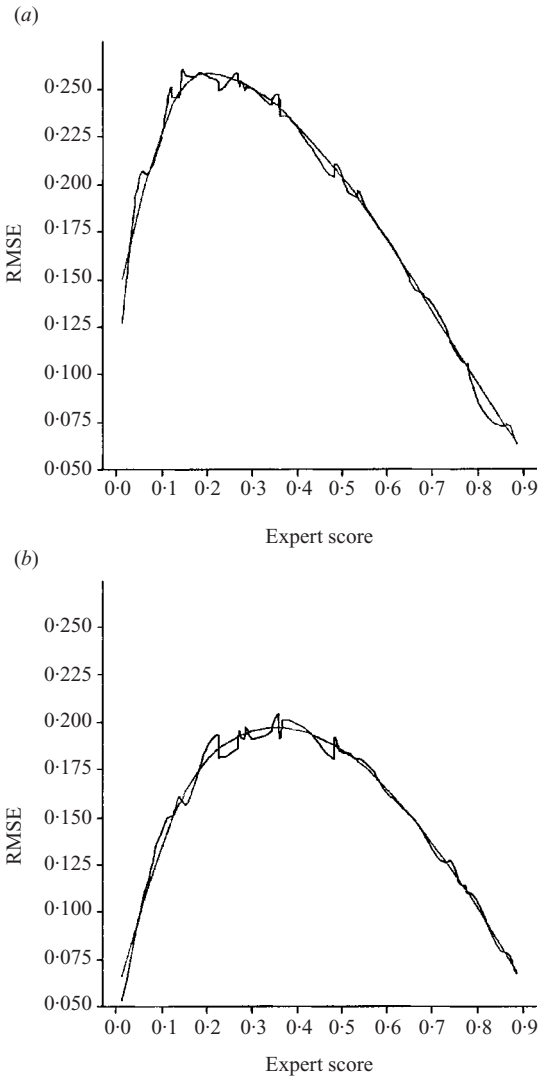


Fig. 4. RMSE for observer 4, (*a*) before and (*b*) after further training, for continuous scores (fractions).

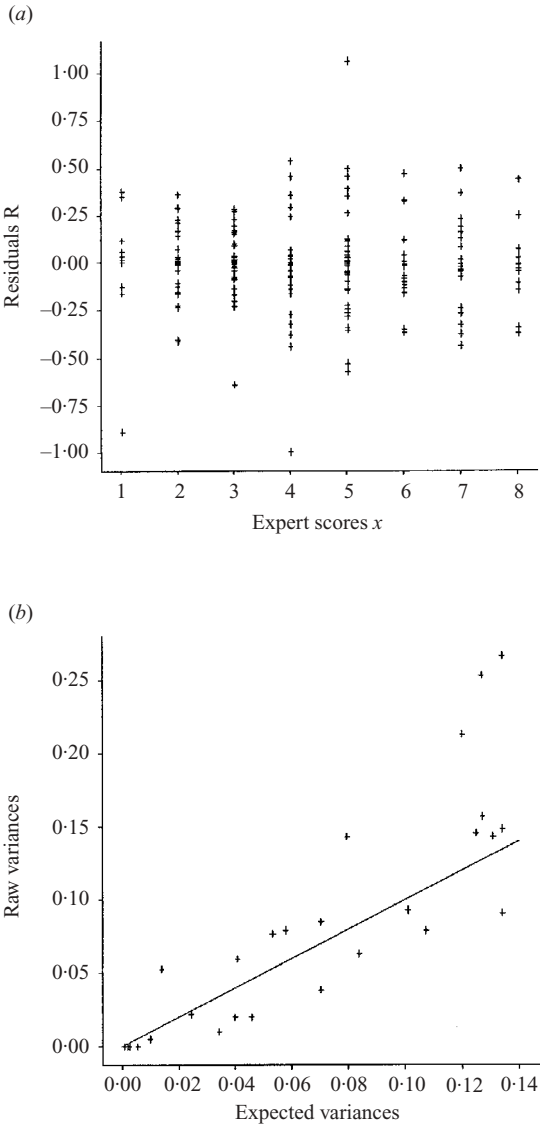


Fig. 5. (*a*) Pearson residuals for cumulative probabilities plotted against expert scores for discrete scores. Observer 4 before training. (*b*) Raw variances plotted against expected variances for discrete scores. Observer 4 before training.

This function can be evaluated from the bootstrap sample and is plotted in Fig. 4 for observer 4. The plot is smoothed by fitting a spline. The RMSE combines the systematic differences with the expert (bias) with the variability of the observer.

### *Checking the fit of the models*

Residuals are inspected routinely after the fit of a model. For the threshold model Pearson residuals were examined, which are scaled differences between proportions of scores equal or below $k$ or proportions equal to $k$, $k=1, \ldots, K$, and their values as predicted from the model. For a cow with expert score $x=j$

$$R_{jk}=(N_k/n-\hat{P}_{jk})/\sqrt{(\hat{P}_{jk}(1-\hat{P}_{jk}))}$$

and

$$r_{jk}=(n_k/n-\hat{p}_{jk})/\sqrt{(\hat{p}_{jk}(1-\hat{p}_{jk}))}$$

where $N_k$ is the number of scores equal or below $k$, $n_k$ is the number of scores equal to $k$, $n$ is the total

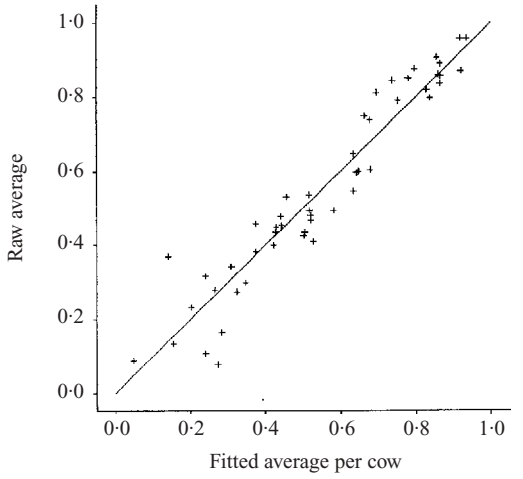


Fig. 6. Raw averages per cow against fitted values for continuous scores (fractions). Observer 4 after training.

number of scores for the particular cow (here $n=5$), $\hat{P}_{jk}=\Phi(\hat{\delta}_k-\hat{\beta}_j-\hat{u})$ and $\hat{P}_{jk}=\hat{P}_{jk}-\hat{P}_{jk-1}$ are the fitted probabilities for a score equal or below $k$ and for a score equal to $k$ respectively, and $\hat{u}$ is a prediction for the cow's random effect. Details regarding predictions $\hat{u}$ are given in Engel & Keen (1994) and Keen & Engel (1997). Patterns in the sign or size of residuals may indicate the need for a different underlying distribution or a lack of homogeneity of variances on the underlying scale. For observer 4 before further training, the residuals $R_{jk}$ are plotted against the corresponding expert scores in Fig. 5*a*. There is no indication for a marked departure from homogeneity of variances on the underlying scale: the ranges of the residuals for the different values of $x$ are fairly similar. There are some extreme residuals, which deserve close inspection (they were retained in the analysis).

To check on the (co)variance structure induced by the random cow effects, the raw variances of proportions $N_k/n$ for all cows in the sample with the same expert score $x$, say $V_k(x)$ were calculated. These variances $V_k(x)$ were plotted against their expected values (Appendix A1) as derived from the analysis. Figure 5*b* offers an example for observer 4 before further training. Considering that the distribution of a raw variance estimator will be skewed to the right, i.e. many estimates around the true value and an occasional estimate which is much larger, the plot looks quite satisfactory.

For the continuous scores (fractions), the linear relationship between probit($\mu$) and probit($x$) in Eqn (3) was initially checked. Formally, this can be done by adding terms to Eqn (3) that are expected to improve the fit and test whether these extra terms are needed. Common choices are quadratic and cubic terms of probit($x$). A more flexible alternative, used

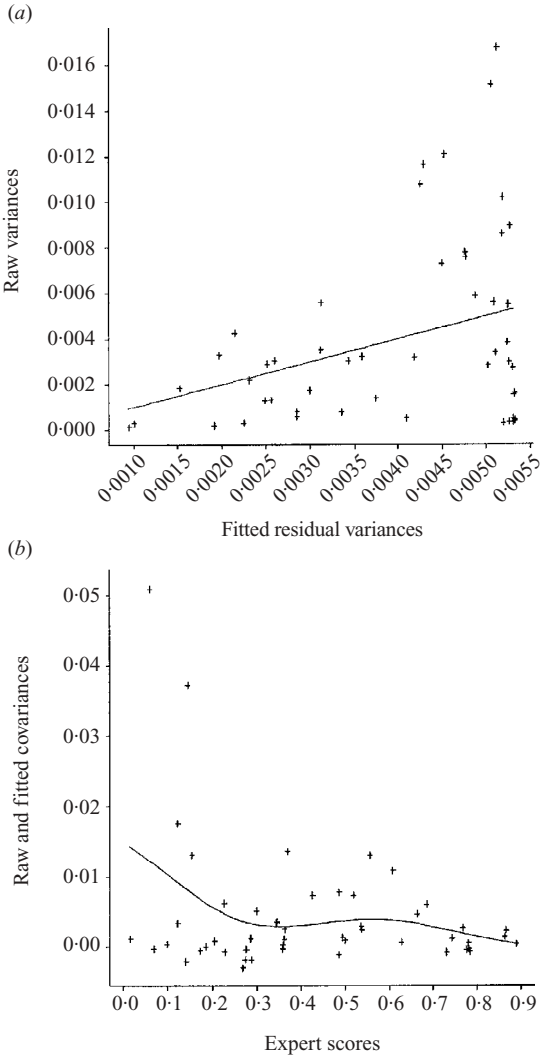


Fig. 7. (*a*) The raw variances per cow against the fitted residual variances for continuous scores (fractions). Observer 4 after further training. (*b*) The raw and fitted (solid line) covariance estimates against the expert scores for continuous scores (fractions). Observer 4 after further training.

here, is to add a natural cubic spline (NCS) (Green & Silverman 1994). A NCS may be fitted as if it were a particular set of random effects (Verbyla *et al.* 1999), and is part of the standard mixed model features in GenStat. With an approximate restricted likelihood ratio test (Appendix A2), the need for the additional spline term was assessed. In addition, a more informal visual check was obtained by plotting the averages per cow against the expected values derived from expression (4). This plot is shown in Fig. 6 for observer 4 after further training.

Table 3. *The proportion agreement* (*average of diagonal elements of e.g. Table 2*) *with the expert for the discrete scores and the proportion agreement up to a difference of one class* (*diagonals and off diagonals*), *per observer before* (*B*) *and after* (*A*) *further training. 95% confidence intervals are shown in parentheses. Observer 9 did not repeat the five sessions*

| Observer | Before/after | Diagonals | Diagonals and 1st off diagonals |
|---|---|---|---|
| 1 | B | 0·38 (0·31; 0·46) | 0·76 (0·58; 0·99) |
| | A | 0·39 (0·31; 0·50) | 0·84 (0·62; 1·00) |
| 2 | B | 0·25 (0·17; 0·34) | 0·64 (0·46; 0·88) |
| | A | 0·41 (0·31; 0·51) | 0·85 (0·63; 1·00) |
| 3 | B | 0·45 (0·34; 0·56) | 0·92 (0·67; 1·00) |
| | A | 0·47 (0·37; 0·56) | 0·94 (0·73; 1·00) |
| 4 | B | 0·39 (0·28; 0·51) | 0·86 (0·61; 1·00) |
| | A | 0·40 (0·31; 0·49) | 0·85 (0·65; 1·00) |
| 5 | B | 0·38 (0·30; 0·47) | 0·78 (0·59; 1·00) |
| | A | 0·45 (0·34; 0·56) | 0·87 (0·64; 1·00) |
| 6 | B | 0·34 (0·25; 0·45) | 0·80 (0·58; 1·00) |
| | A | 0·43 (0·37; 0·49) | 0·89 (0·73; 1·00) |
| 7 | B | 0·34 (0·25; 0·44) | 0·81 (0·59; 1·00) |
| | A | 0·42 (0·35; 0·51) | 0·85 (0·66; 1·00) |
| 8 | B | 0·25 (0·17; 0·36) | 0·61 (0·41; 0·87) |
| | A | 0·46 (0·36; 0·56) | 0·86 (0·63; 1·00) |
| 9 | — | 0·39 (0·29; 0·50) | 0·86 (0·63; 1·00) |

Subsequently the variance function was checked by plotting raw variances per cow against expected residual variances. Such a plot is presented in Fig. 7*a* for observer 4 after further training. Again, there was confirmation that there is no need for more intricate variance functions.

Finally, the covariance structure as induced by the random cow effects was studied. Raw and fitted covariance estimates are plotted against the expert scores in Fig. 7*b* for observer 4 after further training (some details are in Appendix A2). The fitted covariances were smoothed by using a spline. There is no indication that the covariance structure induced by the model is inadequate, although there are some large values for the raw covariance estimates.

## RESULTS

### *Results for discrete ordered scores*

Table 2 presents a typical summary for the discrete scores. Before further training, observer 4 regularly classified cows into the immediately neighbouring classes of the expert score. The probability of a difference of two classes or more was substantial for the two lowest expert scores. Further training clearly affected the results, but changes were not always for the best. Results for expert score 2 have improved, but for expert score 1 there is an even stronger probability for a higher score than before. The observer also showed a marked tendency to underestimate cows with expert score 4 and overestimate cows with expert score 6. Results for the extreme scores 7 and 8 have improved.

The average probability of agreement for observer 4, which is the average of the diagonal elements in Table 2, was about the same before and after training: 0·39 and 0·40 respectively. These probabilities are shown in Table 3 for all observers. Clearly, standards with respect to agreement between observer and expert should not be put too high for individual observers, since none of them agreed for more than 47% with the expert. When a difference of 1 class was allowed for, i.e. examining the average of the diagonal and adjacent subdiagonal elements, percentages agreement were in the order of 80%. Note that observers 2 and 8 who performed relatively poorly, showed substantial improvement after further training.

Fitting the threshold model may offer some problems when the number of observations for a particular score is very small. In that case some scores will have to be pooled. Observer 6 did not assign score 1 and in the analysis (results of which are not shown) scores 1 and 2 were pooled.

### *Results for continuous scores* (*fractions*)

Figures 3 and 4 present typical summaries of the results for the analysis of fractions. It had already been noted that there is some improvement after further training, but observer 4 still scored markedly and

significantly higher than the expert. The discrepancies for observer 4 were quite large compared with some of the other observers as can be seen in Fig. 8. Observers 5, 7 and 8 were generally quite close to the expert.

Observer 4 had the largest root mean square error (RMSE) from all observers, as can be seen in Fig. 9. Observer 5 had a relatively small RMSE and the difference with observer 4 was substantial. Note that the RMSE of observer 8 was relatively large for expert scores near 1. Observer 9 showed an odd curve for the RMSE before training because the systematic difference with the expert (Fig. 8) was positive for the first half of the range of $x$ and negative for the second half. After training observer 9 had the smallest RMSE among all observers.

The estimated intercept $a$ and slope $b$ from expression (4) and components of variance per observer are given in Table 4. Values for $a$ and $b$ were indeed close to 0 and 1 for observers 5, 7 and 8. Observer 2 had a relatively large cow component that was reduced after further training. This was not merely a difference of opinion with the expert about the importance of certain aspects of an animal's gait, since this was largely reflected by the linear part $\alpha+\beta\,\text{probit}(x)$ of the model. It suggested that some of the considerations of this observer were not shared with the expert at all. The residual variance component was often distinctly reduced after further training, implying a higher repeatability.

The lack of fit test based on an additional spline term was not significant ($P=0{\cdot}08$) for observer 4 before training (Table 4). This was in agreement with Fig. 6, which suggests an adequate fit for this observer. Only for observer 8, after further training, was there significant lack of fit. This slightly inflated the estimated component for cows after further training. The spline (not shown) meandered closely around the curve shown in Fig. 8 and the lack of fit was of little practical importance. Figures 7*a* and 7*b* show that only marked departures from the assumed (co)variance structures can be detected.

In Fig. 10 the paper scores are plotted against the discrete scores for the expert. The paper score means corresponding to the discrete scores were nearly equidistant, with the first mean at 0·05 and successive means about 0·11 apart. The fitted means from an analysis of the expert's paper strip scores, with the expert's discrete scores as the levels of an explanatory factor, were nearly the same. It might be hypothesized that the mean for score 9 (which does not occur in our data) is about 0·95. Possibly, the expert, who was used to the system with discrete scores $1, \ldots, 9$, divided the paper strip in about equal parts, with some aversion for the extremes. The expert's estimated cow and residual variance components, with values 0·01 (0·003) and 0·008 (0·0005) respectively, were reassuringly small.

*Additional information from the bootstrap samples*

The bootstrap samples offered an opportunity to study some of the properties of the statistical methods used. For the discrete scores, from results in the literature for binary data (which is the special case $K=2$ of two scores) some underestimation of the component of variance for cows $\sigma_u^2$ may be expected (Breslow & Clayton 1993; Engel *et al.* 1995; Engel 1998; Engel & Buist 1998). Indeed, the averages of the bootstrap samples were about 14% below the estimated components. It is possible to correct for this bias (Kuk 1995) with a more extensive bootstrap procedure. However, the need to do so was not pressing, particularly since the averages of the bootstrap probabilities $\hat{\mathbf{\Pi}}_{jkl}$, $l=1, \ldots, 500$, and the estimates $\hat{\mathbf{\Pi}}_{jk}$ were quite similar, showing that the estimation procedure was fairly unbiased in this respect. For the paper strip scores the estimated cow and residual components were practically unbiased.

## DISCUSSION

It was shown how visual scores in the form of ordered categorical data or continuous fractions can be analysed by models that account for the type of data, the associated non-linearity and heterogeneity of variance, and the dependence between observations of the same cow. The results of the analyses, as summarized in tables and plots, offer useful tools to assess and safeguard the quality of subjective observations. These tools can be implemented in a system to monitor observer performance. Many questions can be raised, such as: how well can we expect an observer to perform, how well should an observer perform, how often should observer performance be assessed, what amount of (regular) training is needed. The answers depend on the characteristic of interest, the degree of accuracy that is demanded and practical experience with the monitoring system. Results in Table 3 and Figs 8*a*, *b* and 9*a*, *b*, for instance, suggest what degree of agreement between observer and expert can be reasonably expected for classification of gait with the present amount of training. When higher observer performance is required, analyses of classification results after further intensified training, possibly for a larger number of potential observers, will have to show whether this can be achieved. In a system where observers participate over a long period, a high expertise can be built up under observer selection. In a system where observers regularly change, possibly a lower performance level will have to be accepted. Rules for selection of observers can be based on analyses of results of several rounds of classification of videos. In this paper we analysed 250 repeated scores assigned to 50 videos. When narrower confidence intervals are required to judge the observers, more videos will have to be classified, possibly with less

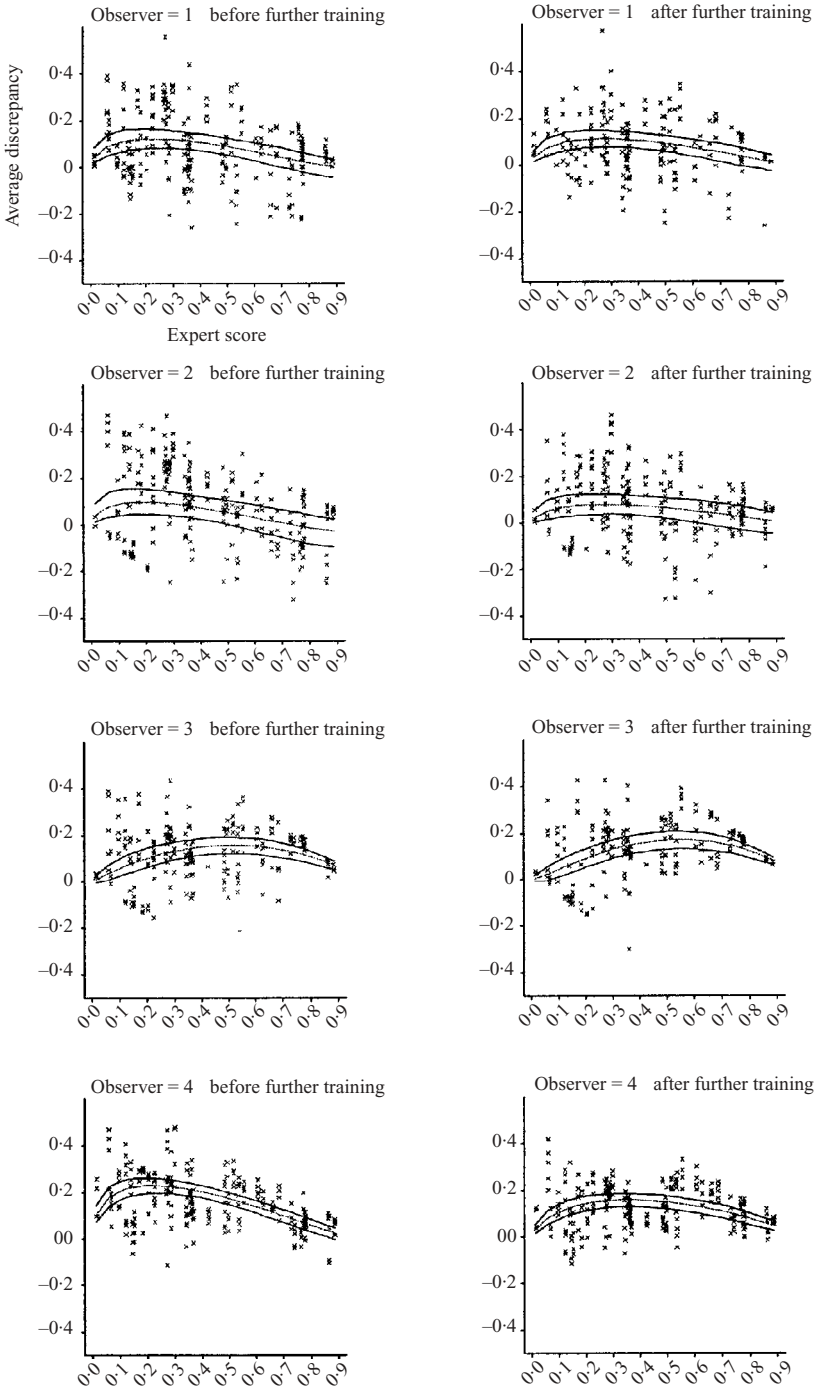


Fig. 8 *a*. The average discrepancy between observer and the expert plotted against the expert score with 95 % confidence limits for continuous scores (fractions). Curves before and after training. The original data are included as well.

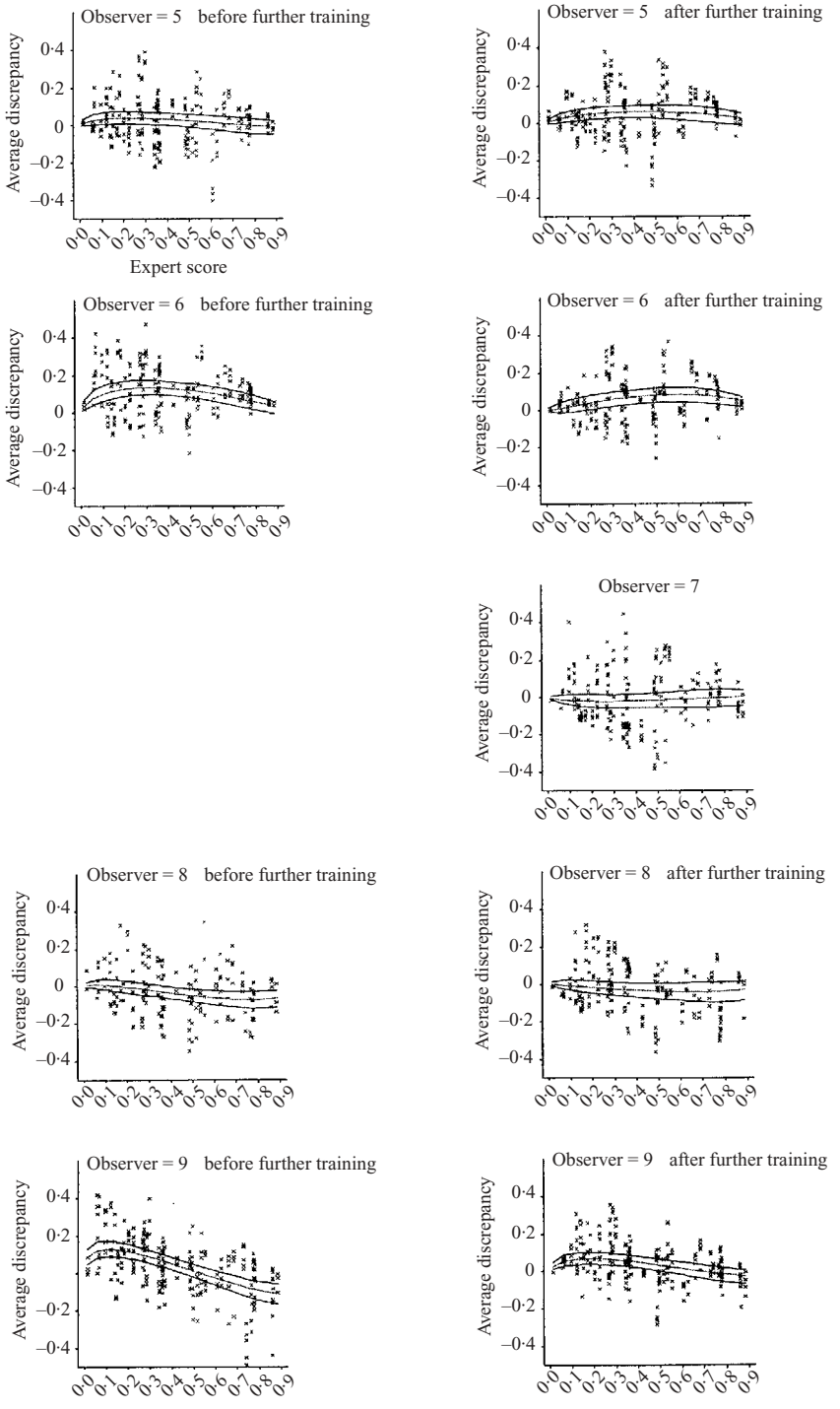


Fig. 8*b*. The average discrepancy between observer and the expert plotted against the expert score with 95 % confidence limits for continuous scores (fractions). Curves before and after training. The original data are included as well.

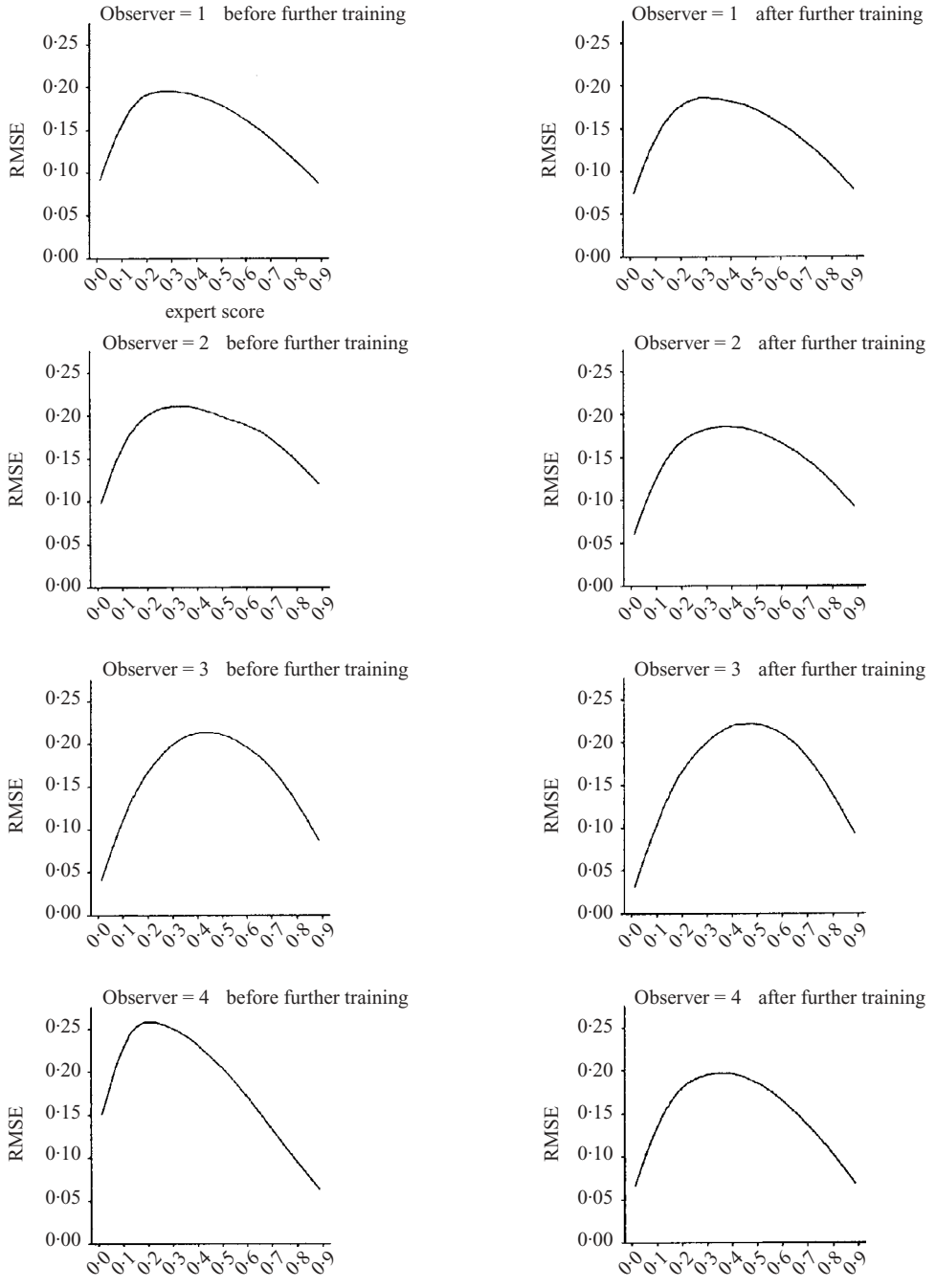


Fig. 9*a*. RMSE against the expert score, before and after further training, for continuous scores (fractions).

than five repeats per video. These issues are important and may be tackled analytically or by simulation. They were not the subject of this paper. Here we focused on the models and the methodology. They are the building bricks of a monitoring system. How that system should be built is still to be decided and depends on the requirements, practical restrictions and available means.

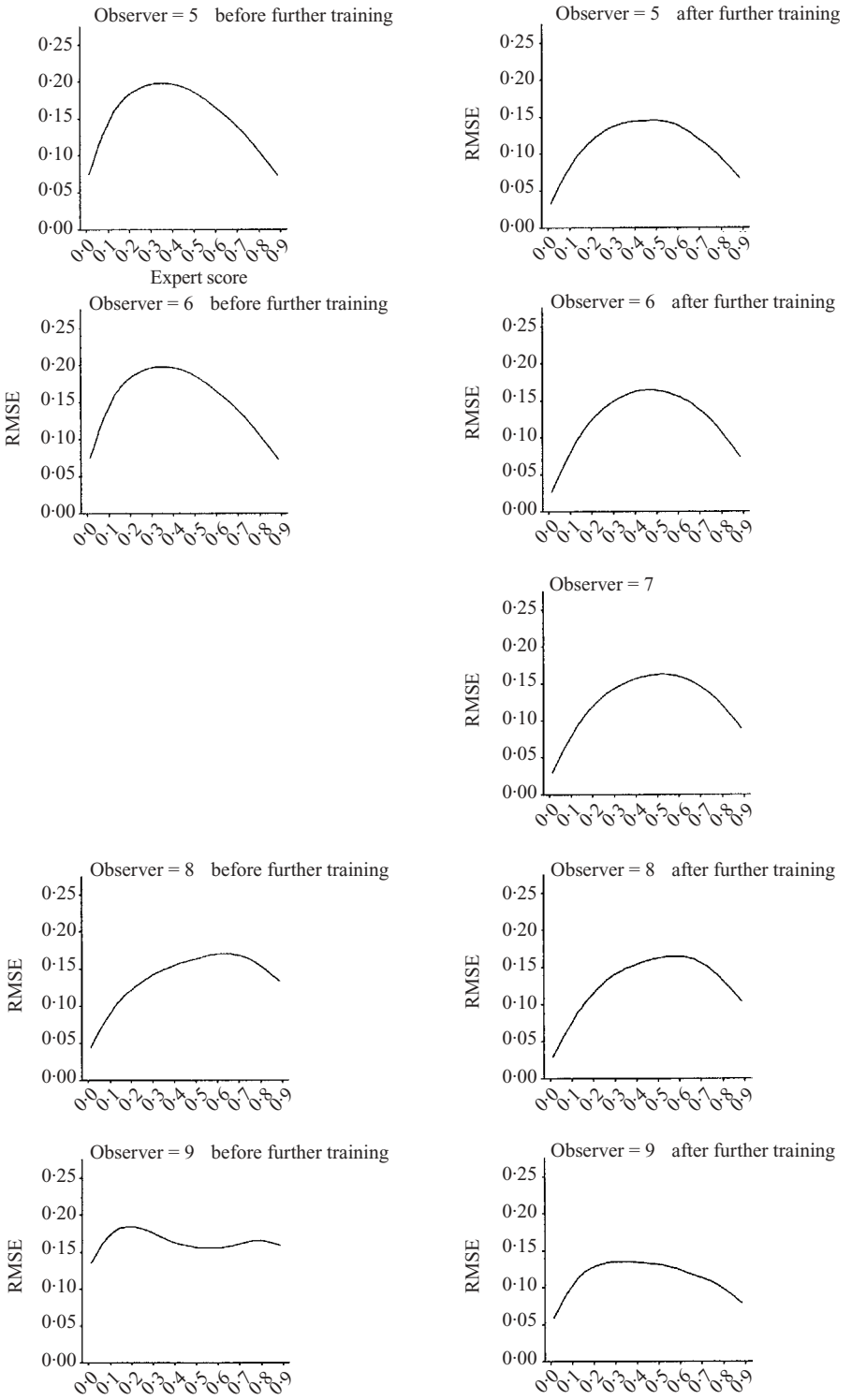


Fig. 9*b*. RMSE against the expert score, before and after further training, for continuous scores (fractions).

Table 4. *Summary of estimated intercept* a, *slope* b, *cow and residual variance components,* P-*value for the lack of fit test based on the additional spline term per observer, before* (*B*) *and after* (*A*) *further training, for the continuous fractions* (*paper strip scores*). *Standard errors are in parentheses. Observer 7 did not repeat the five sessions*

| Observer | Before/ after | *a* | *b* | $\sigma_u^2$ | $\sigma^2$ | *P* (spline) |
|---|---|---|---|---|---|---|
| 1 | B | 0·23 (0·05) | 0·82 (0·06) | 0·09 (0·02) | 0·047 (0·006) | 0·50 |
| | A | 0·23 (0·05) | 0·87 (0·06) | 0·09 (0·02) | 0·035 (0·004) | 0·20 |
| 2 | B | 0·13 (0·06) | 0·79 (0·08) | 0·21 (0·05) | 0·048 (0·005) | 0·40 |
| | A | 0·16 (0·06) | 0·89 (0·07) | 0·15 (0·03) | 0·036 (0·004) | 0·50 |
| 3 | B | 0·41 (0·05) | 1·10 (0·06) | 0·10 (0·02) | 0·032 (0·003) | 0·10 |
| | A | 0·44 (0·05) | 1·18 (0·07) | 0·13 (0·03) | 0·026 (0·003) | 0·16 |
| 4 | B | 0·45 (0·04) | 0·75 (0·05) | 0·05 (0·01) | 0·025 (0·003) | 0·08 |
| | A | 0·38 (0·04) | 0·93 (0·05) | 0·06 (0·01) | 0·021 (0·002) | 0·08 |
| 5 | B | 0·06 (0·05) | 0·91 (0·06) | 0·08 (0·02) | 0·034 (0·003) | 0·50 |
| | A | 0·16 (0·05) | 0·99 (0·06) | 0·09 (0·02) | 0·016 (0·002) | 0·50 |
| 6 | B | 0·30 (0·05) | 0·90 (0·07) | 0·11 (0·02) | 0·027 (0·003) | 0·39 |
| | A | 0·20 (0·05) | 1·08 (0·07) | 0·12 (0·03) | 0·016 (0·002) | 0·50 |
| 7 | — | −0·04 (0·05) | 1·04 (0·07) | 0·10 (0·03) | 0·050 (0·005) | 0·50 |
| 8 | B | −0·12 (0·04) | 0·87 (0·06) | 0·07 (0·02) | 0·064 (0·007) | 0·47 |
| | A | −0·09 (0·05) | 0·96 (0·07) | 0·13 (0·03) | 0·029 (0·003) | 0·02 |
| 9 | B | 0·03 (0·04) | 0·61 (0·06) | 0·08 (0·02) | 0·050 (0·005) | 0·50 |
| | A | 0·08 (0·04) | 0·83 (0·05) | 0·08 (0·02) | 0·021 (0·002) | 0·50 |

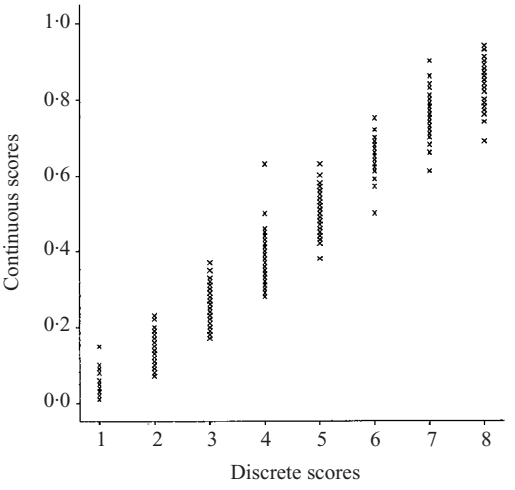


Fig. 10. Continuous scores (fractions) against discrete scores for the expert.

With the discrete scores, cows that are inbetween two categories may be classified on one side of a threshold by the observer and on the other side by the expert. So, although observer and expert may be quite close in their opinion, they will differ by one class. With the continuous scores this problem is solved but at the cost of additional variability inherent to a continuous score. Obviously, when discrete scores correspond to clearly separated different states of an individual or object, i.e. states of development of an embryo, there is no need for continuous fractions. Otherwise, which of the two types of score is more favourable will depend on how well observers are able to discriminate between e.g. gaits of cows. The paper score method offers a more concise summary of classification results than the discrete scores. However, the ultimate decision will largely have to depend on the opinion of classifiers and experts about the two scoring systems. Since classifiers and expert were used to the discrete score system, while the paper score method was something of a novelty, it was too early to decide which type of score is to be preferred for gait.

Care should be taken in the choice of a gold standard. In the current study the expert has a high repeatability for the discrete scores and there was almost full agreement with the observers in a discussion afterwards. For the paper scores there is more variability, as illustrated by Fig. 10. The mean of 10 repeated classifications from the expert was used. Generally, a gold standard based on several experts is to be preferred. It is not advisable to apply methodology to situations where no proper gold standard is available. For instance, routine use of the average of all observers as a substitute for a gold standard is not recommended.

Although the cows are viewed only briefly, some elements in the videos, perhaps in the background, may be recognized by an observer, possibly subconsciously. Videos are carefully selected, but some underestimation of the residual variance seems unavoidable. Memory effects may be reduced by regularly changing the videos or by manipulation of their background. Time effects are always an issue in

any scoring system. There is a variety of causes. For instance, an observer may happen to start with a number of videos with quite low scores and persist in scoring too low for a while. These effects may occur both in practice and during the scoring of the videos. The videos were offered in a random order. In practice, of course, scores may be considerably clustered. In scoring systems where a high precision has to be attained, effects of clustering may be investigated by offering the videos in a specific order. In practice, some standardization may be attained by showing some typical videos before the actual scoring starts.

We are grateful to Bonne Beerda, Harry Blokhuis, Marc Bracke and Joop de Bree for helpful comments on different versions of the manuscript and to Gidi Smolders for use of his expertise on scoring of a cow's gait.

## REFERENCES


Aitchison, J. & Shen, S. M. (1980). Logistic-normal distributions: some properties and uses. *Biometrika* **67**, 261–272.

Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

Efron, B. & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Engel, B. (1998). A simple illustration of the failure of PQL, IRREML and APHL as approximate ML methods for mixed models for binary data. *Biometrical Journal* **40**, 141–154.

Engel, B. & Buist, W. G. (1998). Bias reduction of approximate maximum likelihood estimates for heritability in threshold models. *Biometrics* **54**, 1155–1164.

Engel, B. & Keen, A. (1994). A simple approach for the analysis of generalized linear mixed models. *Statististica Neerlandica* **48**, 1–22.

Engel, B., Buist, W. G. & Visscher, A. (1995). Inference for threshold models with variance components from the generalized linear mixed model perspective. *Genetics Selection Evolution* **27**, 15–32.

Garner, J. P., Falcone, C., Wakenell, P., Martin, M. & Mench, J. A. (2002). Reliability and validity of a modified gait scoring system and its use in assessing tibial dyschondroplasia in broilers. *British Poultry Science* **43**, 355–363.

Genstat Committee (2000). *The Guide to GenStat Release 4.2*. (Ed. R. W. Payne). VSN International Ltd.

Green, P. J. & Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. London: Chapman & Hall.

Keen, A. (2001). Procedure IRCLASS. In *Biometris GenStat Procedure Library Manual* (Eds P. W. Goedhart & J. T. N. M. Thissen) (freely available from http://www.biometris.nl/genstat).

Keen, A. & Engel, B. (1997). Analysis of a mixed model for ordinal data by iterative re-weighted REML. *Statistica Neerlandica* **51**, 129–144.

Keen, A. & Engel, B. (2001). Procedure IRREML. In *Biometris GenStat Procedure Library Manual* (Eds P. W. Goedhart & J. T. N. M. Thissen) (freely available from: http://www.biometris.nl/genstat).

Kuk, A. Y. C. (1995). Asymptotically unbiased estimation in generalized linear models with random effects. *Journal of the Royal Statistical Society*, series B **57**, 395–407.

Manson, F. J. & Leaver, J. D. (1988). The influence of concentrate amount on locomotion and clinical lameness in dairy-cattle. *Animal Production* **47**, 185–190.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman & Hall.

Morell, C. H. (1998). Likelihood ratio testing of variance components in the linear mixed-effects model using restricted maximum likelihood. *Biometrics* **54**, 1560–1568.

Sprecher, D. J., Hostetler, D. E. & Kaneene, J. B. (1997). A lameness scoring system that uses posture and gait to predict dairy cattle reproductive performance. *Theriogenology* **47**, 1179–1187.

Verbyla, A. P., Cullis, B. R., Kenward, M. G. & Welham, S. J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines (with discussion). *Applied Statistics* **48**, 269–311.


## APPENDIX

### *A1: Some details of the analysis of ordered discrete scores*

To derive expression (1) note that $z$ is normally distributed with mean $\beta_j$ and variance $\sigma_u^2+1$ for expert score $x=j$.

For the threshold model the variance of a cumulative proportion $N_k/n$ for a random cow with expert score $x=j$ is equal to (Engel *et al.* (1995), expressions (4) and (5))

$$\gamma_{jk}(1-\gamma_{jk})[1+(n-1)\times\{\Phi_2(c_{jk},\, c_{jk};\rho)-\gamma_{jk}^2\}/\{\gamma_{jk}(1-\gamma_{jk})\}]/n$$

where $\Phi_2$ is the cumulative density of the standard bivariate normal distribution (two-dimensional normal probability integral) with correlation $\rho=\sigma_u^2/(\sigma_u^2+1)$, arguments $c_{jk}=(\theta_k-\beta_j)\sqrt{(\sigma_u^2+1)}$ and $\gamma_{jk}$ from expression (1). This is the expected value of the the raw variance $V_k(x)$. Function $\Phi_2$ is available in many statistical packages, including GenStat.

95% confidence intervals were derived from a normal approximation for logit transformed estimated probabilities $\hat{L}_{jk}=\text{logit}(\hat{\Pi}_{jk})=\log(\hat{\Pi}_{jk}/(1-\hat{\Pi}_{jk}))$. The logit transformation, which 'stretches' the probabilities from values between 0 and 1 to values between $-\infty$ and $+\infty$, was applied to improve the normal

approximation. For the transformed probabilities the familiar intervals ($\hat{L}_{jk} \pm 1{\cdot}96 \times$ standard error of $\hat{L}_{jk}$), say ($\mathrm{Low}_{jk}$, $\mathrm{Up}_{jk}$), were calculated. The standard error of $\hat{L}_{jk}$ was estimated by the standard deviation of the logit transformed bootstrap probabilities $\mathrm{logit}(\hat{\Pi}_{jkl})$, $l = 1, \ldots, 500$. The intervals were transformed back into intervals $(1/(1+\exp(-\mathrm{Low}_{jk}))$, $1/(1+\exp(-\mathrm{Up}_{jk}))$ for the probabilities $\Pi_{jk}$. More refined bootstrap confidence intervals are discussed in Chapter 14 of Efron & Tibshirani (1993). These intervals require a nested series of bootstrap samples, which is more computer intensive. For the present purpose, we consider the intervals obtained through the logit transformation to be sufficiently accurate.

### *A2: Some details of the analysis of fractions*

To derive expression (4) we introduced $z = \alpha + \beta\,\mathrm{probit}(x) + u + e$, where $u$ and $e$ are independently normally distributed with mean 0 and variances $\sigma_u^2$ and 1 respectively. Now, $P(z>0|x,u) = \Phi(\alpha+\beta\,\mathrm{probit}(x)+u) = \mu$, and $E(y|x) = E(\mu|x) = P(z>0|x) = \Phi((\alpha+\beta\,\mathrm{probit}(x))/\sqrt{(\sigma_u^2+1)})$, because $z$ is normally distributed with mean $\alpha+\beta\,\mathrm{probit}(x)$ and variance $\sigma_u^2+1$. The probit link was preferred over the logit link because of this exact result.

In the bootstrap simulation the beta distribution was fitted by equating the mean and variance of fraction $y$, given the expert score $x$ and cow effect $u$, to the mean and variance of a beta distribution with parameters $A$ and $B$, i.e. $\mu = A/(A+B)$ and $\sigma^2\mu(1-\mu) = \{A/(A+B)\}\ \{B/(A+B)\}/\{A+B+1\}$. For each simulation, new cow effects $u$ were sampled from a normal distribution with mean 0 and variance $\hat{\sigma}_u^2$ (the estimated value from the analysis). Employing the estimates $\hat{\alpha}$ and $\hat{\beta}$, the means $\mu$ were derived from (3). For each cow, $A$ and $B$ were calculated from $\mu$ and the estimate $\hat{\sigma}^2$ and fractions $y$ were sampled from the corresponding beta distribution.

The lack of fit test for an extra spline term was obtained from the last iteration of the iterative reweighted restricted maximum likelihood (IRREML) algorithm. In each step of this iteration process, a linear mixed model was fitted to an artificial dependent variable. Details are presented in Engel & Keen (1994). From the last iteration, this artificial dependent variable, and associated iterative weights, were saved. One extra iteration was performed with and without the additional spline term in the model. The test was derived from the difference between the two values of the $-2 \times \log$(restricted maximum likelihood) in the approximate linear mixed models. The null-hypothesis that the component of variance for the random part of the spline is zero is on the boundary of the parameter space, and the $P$-value was derived from a mixture of a probability point mass of 0·5 at value 0 and probability 0·5 for a chi-square distribution with one degree of freedom (Morell 1998).

The covariance between any two observed fractions on the same cow is $\mathrm{Var}(\mu|x) = \Phi_2(c, c; \rho) - \Phi(c)^2$, where $\Phi_2$ is the cumulative density of the standard bivariate normal distribution with correlation $\rho = \sigma_u^2/(\sigma_u^2+1)$ and arguments $c = a + b\,\mathrm{probit}(x)$ (Engel *et al.* 1995, expression (4)). The fitted covariances were obtained by substitution of the parameter estimates. The raw covariance estimate per cow was defined as $(nM_1^2 - M_2)/(n-1)$, where $n$ is the number of repeated classifications per cow, $M_1$ is the mean of the differences $(y - E(y|x))$ and $M_2$ is the mean of the squared differences $(y - E(y|x))^2$. The expectation $E(y|x)$ was evaluated for the parameter estimates, and $\hat{\sigma}_u^2$ from (4). A non-parametric alternative for $E(y|x)$ is a NCS in $x$ fitted to the cow averages. The expected value of this raw covariance estimate approximates the covariance between any two observations on the same cow with expert score $x$.