The Old School, Brewhouse Hill, Wheathampstead,
Hertfordshire AL4 8AN, UK

Animal Welfare 2007, 16: 131-133
ISSN 0962-7286

# Effect of training on the inter-observer reliability of lameness scoring in dairy cattle

S March*†, J Brinkmann† and C Winkler‡

† Research Centre for Animal Production and Technology, Georg-August-University of Goettingen, Driverstraße 22, 49377 Vechta, Germany
‡ Division of Livestock Sciences, Department of Sustainable Agricultural Systems, University of Natural Resources and Applied Life Sciences, Gregor-Mendel-Straße 33, 1180 Vienna, Austria
* Contact for correspondence and requests for reprints: solveig.march@agr.uni-goettingen.de

## Abstract

*In the present study, the effect of training on inter-observer reliability was studied for a 5-category lameness scoring system used for routine on-farm surveys of welfare in dairy cattle. The inter-observer agreement between an experienced and an initially inexperienced observer was determined during an initial training phase and at specific time points in the course of data collection in 46 herds. During the training phase on three farms, inter-observer reliability increased to an acceptable level for both the 5-category gait scoring system and the distinction between lame and non-lame cows.*

*The 4th testing after 17 on-farm visits revealed a considerable increase in inter-observer reliability which was further improved in the course of the on-farm visits.*

*In conclusion, acceptable inter-observer agreement for differentiating between non-lame and lame cows was achieved after only a brief introduction. In order to obtain high inter-observer repeatability with the 5-category gait scoring system used in this study, (more) intensive training procedures are required.*

**Keywords:** *animal welfare, gait-scoring, inter-observer reliability, lameness, observer training, PABAK*

## Introduction

Various lameness scoring systems based on numerical rating scales are used for routine surveys of welfare in dairy cattle (Winckler & Willen 2001; De Rosa *et al* 2003; Winckler *et al* 2003). As they do not require any equipment, these methods can be easily applied in on-farm research. Their validity with regard to claw lesions and/or other behavioural traits has been shown in several studies (Winckler & Willen 2001; O'Callaghan *et al* 2002). However, information on inter-observer reliability is rather scarce and sometimes contradictory (eg Baadsgaard & Enevoldsen 1997). In the present study, we investigated the effect of training on the inter-observer reliability for a 5-category gait scoring system in order to evaluate the importance of a learning phase and the amount of training necessary for reliable results.

## Materials and methods

The inter-observer agreement between an experienced and an initially inexperienced observer was determined in loose-housed Holstein Friesian herds. First, lameness scoring was carried out on three farms during a gait scoring training, ie after a theoretical introduction to the method including videotapes for demonstration and limited live observations. Subsequently, inter-observer reliability testing took place another six times in the course of data collection in 46 herds (two to six months later). These farms were visited by both observers, but inter-observer reliability testing only took place at specific time points. Locomotion was assessed using the 5-category gait scoring system described by Winckler and Willen (2001; Table 1). Animals were observed while walking in the feed or walking alleys (slatted and solid floors) without forceful driving.

The prevalence-adjusted bias-adjusted Kappa ([PABAK] described by Gunnarsson 2000; Keppler *et al* 2004; Petersen *et al* 2004), the weighted Kappa coefficient, the Spearman rank correlation coefficient ($r_s$) and the proportion agreement were calculated as parameters of inter-observer reliability. Calculations were either carried out with the original scores (5-category) or after transformation into lame/non-lame scores ie 1-2: non-lame, 3-5: lame. According to Byrt *et al* (1993), the Kappa coefficient measures the agreement beyond what would be expected by chance. The weighted Kappa coefficient also takes into account that larger disagreement is more important than near disagreement. Finally, the prevalence-adjusted bias-adjusted Kappa (PABAK = $[(k \times p)-1]/(k-1)$ where k is the number of categories and p the proportion of matchings) is based on the unweighted Cohen's kappa test and it is the value that kappa would take if, in addition, the prevalence of each category was equal (Gunnarsson 2000). Matchings are only counted, if both observers give exactly the same score.

All coefficients may range between 0 and 1 meaning no agreement between observations if the coefficient is equal to 0 and perfect agreement if the value is equal to 1.



https://doi.org/10.1017/S096272860003116X Published online by Cambridge University Press

**Table 1 Lameness scoring system (modified after Winckler & Willen 2001).**

| Lameness score | Definition |
|---|---|
| 1 | Normal gait |
| 2 | Uneven gait: stiff, very careful |
| 3 | Lame: Short striding gait with one limb (even if just noticeable). |
| 4 | Lame: Short striding gait with more than one limb or strong reluctance to bear weight on one limb. |
| 5 | Lame: Does not support on one limb or strong reluctance to put weight on limb in two or more limbs; holding a limb up whenever possible. |

**Table 2 Development of coefficients of inter-observer reliability between an experienced and an initially inexperienced observer on 9 occasions.**

| Test number | Previous experience with gate scoring (cows/farms) | Number of cows | PABAK (lame/non-lame) | 5-category | Weighted Kappa 5-category | Spearman rank correlation coefficient[1] | Proportion agreement 5-category |
|---|---|---|---|---|---|---|---|
| 1 | -/- | 68 | 0.53 | 0.32 | 0.41 | 0.55 | 0.46 |
| 2 | 68/1 | 21 | 0.71 | 0.40 | 0.54 | 0.73 | 0.52 |
| 3 | 89/2 | 50 | 0.52 | 0.40 | 0.52 | 0.67 | 0.52 |
| 4 | 623/17 | 40 | 0.75 | 0.66 | 0.69 | 0.82 | 0.73 |
| 5 | 1,099/30 | 40 | 0.85 | 0.88 | 0.86 | 0.87 | 0.90 |
| 6 | 1,311/35 | 42 | 0.95 | 0.94 | 0.66 | 0.59 | 0.95 |
| 7 | 1,665/44 | 35 | 0.89 | 0.86 | 0.83 | 0.89 | 0.89 |
| 8 | 1,768/47 | 58 | 0.86 | 0.81 | 0.64 | 0.73 | 0.84 |
| 9 | 1,859/49 | 50 | 0.88 | 0.68 | 0.75 | 0.87 | 0.74 |

Test 1: without; tests 2 and 3: after short practical training; tests 4 to 9; in the course of an on-farm research project in 47 herds.
[1] $P < 0.01$.

## Results

With regard to the identification of lame and non-lame cows, respectively, the $PABAK_{lame/non-lame}$ reached 0.53 (n = 68) after only a brief theoretical introduction. It was improved, however not consistently, during the initial training period (Table 2). The coefficients of inter-observer reliability for the 5-category gait scoring system ranged initially between $PABAK_{5\text{-}category} = 0.32$, $r_s = 0.55$ and Kappa weighted$_{5\text{-}category} = 0.41$. Further experience in applying the scoring system during the training period resulted in slightly higher values ($PABAK_{5\text{-category}} = 0.40$, $r_s = 0.73$ and weighted $Kappa_{5\text{-}category} = 0.54$; n = 21). This was confirmed on the third test, ($PABAK_{5\text{-}category} = 0.40$, $r_s = 0.67$ and weighted $Kappa_{5\text{-}category} = 0.52$; n = 50).

After 623 cows on 17 farms had been assessed, the inter-observer reliability in test 4 substantially increased for both the distinction between lame and non-lame cows ($PABAK_{lame/non\text{-}lame} = 0.75$; n = 40) and the 5-category system ($PABAK_{5\text{-}category} = 0.66$, $r_s = 0.82$ and weighted $Kappa_{5category} = 0.69$). A further improvement in most of the inter-observer reliability coefficients was achieved in the comparisons thereafter (tests 5 to 9; Table 2).

## Discussion

The inter-observer agreement achieved during the initial training period (tests 1 to 3) is well within the range of the (little) information given in the literature (Winckler & Willen 2001; O'Callaghan *et al* 2002; Engel *et al* 2003). However, it has to be taken into account that different gait scoring systems use different numbers of categories which is likely to affect the level of agreement. With increasing number of categories, the use of discrete scores decreases the chance of agreement.

Further training and experience with the gait scoring system substantially increased the level of agreement between the formerly inexperienced and the expert observer. Based on video recordings, Engel *et al* (2003) also found a training effect, which was different for the individual observers and in some cases even paradoxical. When a difference of one class within nine categories was accepted, the agreement was in the order of 80% thus confirming the results of the present study.

With regard to the acceptability of the level of agreement, Holzhauer *et al* (2004) defined Kappa values between 0.4 and 0.5 as moderate, values between 0.5 and 0.6 as sufficient and values between 0.6 and 0.8 as good. Habison *et al* 2002 interpreted Kappa coefficients lower than 0.4 as



https://doi.org/10.1017/S096272860003116X Published online by Cambridge University Press

an indicator of low agreement, values between 0.4 and 0.6 as moderate and values equal to or greater than 0.6 as high agreement. Accordingly, PABAK values lower than 0.4 are rated as unsatisfactory; values above 0.4 as acceptable, above 0.6 as good/satisfactory and above 0.8 as very good (Keppler *et al* 2004). Spearman rank correlation coefficients ($r_s$) equal to or higher than 0.7 have also been regarded as indicators of good inter-observer reliability (Keppler *et al* 2004; Rousing & Waiblinger 2004).

Based on these definitions, acceptable/moderate up to satisfactory/good inter-observer reliability for differentiating between non-lame and lame cows and acceptable levels for the 5-category gait scoring system were already achieved after a rather short theoretical and practical introduction. The experience gained in the course of data collection in 46 dairy herds increased all parameters of inter-observer reliability to an at least good/satisfactory level. However, the decision how well an observer should perform for example in a welfare monitoring system will depend finally on the accuracy that is demanded, eg the discrimination between farms. Feasible instruction schemes for on-farm welfare assessors will probably provide much less training than in the present study (more than 1,800 jointly assessed cows). It is likely, therefore, that for monitoring systems with frequent observer changes and little time and resources available for training, a lower level of reliability/agreement will have to be accepted (Engel *et al* 2003).

Depending on the prevalence distribution of the discrete scores, not all coefficients describe the inter-observer reliability correctly. This is true for the rank correlation coefficient as well as weighted Kappa as regards test 6 (Table 2). Both coefficients tend to be inaccurate when most of the data have the same values and/or show a skewed data distribution (Rousing & Waiblinger 2004). This underlines the usefulness of PABAK or other simple measures such as proportion agreement which should be provided additionally.

## Conclusions

Acceptable inter-observer agreement for differentiating between non-lame and lame cows could already be achieved after a short introduction to the method. Acceptable reliability of the 5-category gait scoring system used in this study which aims at a more detailed distinction between gait types and degrees of lameness can also be reached when limited practical experience is included in the training. However, in order to obtain further improvements in the inter-observer repeatability, intensive training procedures with live animals and an experienced observer are required. The number of animals required may amount to between 200 and 300. The integration of the 5-category gait scoring system into on-farm welfare assessment protocols seems to be justified, if such adequate practical learning phase is assured.

## Acknowledgements

The German Federal Agency for Agriculture and Food (BLE) is gratefully acknowledged for financial support within the Federal Organic Farming Scheme. We also would like to thank the organic farming associations for help with organising the field study and the farmers for allowing us to carry out the on-farm investigations.

## References

**Baadsgaard NP and Enevoldsen C** 1997 A potential approach to support animal welfare promotion in a Danish veterinary practice context. *Proceedings of the Society of Veterinary Epidemiology and Preventive Medicine*. 8-11 July 1997. Paris, France

**Byrt T, Bishop J and Carlin JB** 1993 Bias, prevalence and kappa. *Journal of Clinical Epidemiology 46:* 423-429

**De Rosa G, Tripaldi C, Napolitano F, Saltalamacchia F, Grasso F, Bisegna V and Bordi A** 2003 Repeatability of some animal-related variables in dairy cows and buffaloes. *Animal Welfare 12:* 625-629

**Engel B, Bruin G, Andre G and Buist W** 2003 Assessment of observer performance in a subjective scoring system: visual classification of the gait of cows. *Journal of Agricultural Science 140:* 317-333

**Gunnarson S** 2000 *Laying hens in loose housing systems.* Clinical, ethological and epidemiological aspects. *Acta Universitatis Agriculturae Sueciae Veterinaria 73:* 44

**Habison JL, Slater MR and Howe LM** 2002 Repeatability and prediction from a telephone questionnaire measuring diet and activity level in cats. *Preventive Veterinary Medicine 55:* 79-94

**Holzhauer M, Middelesch H, Bartels C and Frankena K** 2004 Evaluation of a Dutch claw health scoring system in dairy cattle. *Proceedings of the 13th international symposium and 5th conference on lameness in ruminants*. 11-15 February 2004. Maribor, Slovenia

**Keppler C, Schubbert A and Knierim U** 2004 Welche Methoden sind zur Beurteilung von Hühnern im Hinblick auf Federpicken und Kannibalismus geeignet? Erste Untersuchungen zum Vergleich verschiedener Methoden im Hinblick auf Durchführbarkeit, Aussagekraft und Wiederholbarkeit. *11. Freilandtagung/ 17. IGN-Tagung*: 71-74. [Title translation: Which methods are adequate to assess laying hens in terms of feather picking and cannibalism? First comparative investigations of methods with regard to feasibility, validity and repeatability]

**O'Callaghan KA, Murray RD and Cripps PJ** 2002 Behavioural indicators of pain associated with lameness in dairy cattle. *Proceedings of the 12th International Symposium on Lameness in Ruminants*. 9-13 January 2002. Orlando, Florida, USA

**Petersen HH, Enøe C and Nielsen CO** 2004 Observer agreement on pen level prevalence of clinical signs in finishing pigs. *Preventive Veterinary Medicine 64:* 147-156

**Rousing T and Waiblinger S** 2004 Evaluation of on-farm methods for testing the human-animal relationship in dairy herds with cubicle loose housing systems: Test-retest and inter-observer reliability and consistency to familiarity of test person. *Applied Animal Behaviour Science 85:* 215-231

**Winckler C and Willen S** 2001 The reliability and repeatability of a lameness scoring system for use as an indicator of welfare in dairy cattle. *Acta Agriculturae Scandinavica, Section A, Animal Science, Supplement 30:* 103-107

**Winckler C, Capdeville J, Gebresenbet G, Hörning B, Roiha U, Tosi M and Waiblinger S** 2003 Selection of parameters for on-farm welfare-assessment protocols in cattle and buffalo. *Animal Welfare 12:* 619-624