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

Animal Welfare 2007, 16: 127-129
ISSN 0962-7286

Reliability of a subjective lameness scoring system for dairy cows

C Brenninkmeyer*†, S Dippel‡, S March§, J Brinkmann§, C Winckler‡ and U Knierim†

† Department of Farm Animal Behaviour and Husbandry, University of Kassel, Nordbahnhofstraße 1a, 37213 Witzenhausen, 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
§ Research Centre for Animal Production and Technology, Georg-August-University of Goettingen, Driverstrasse 22, 49377 Vechta, Germany
* Contact for correspondence and requests for reprints: brenninkmeyer@wiz.uni-kassel.de

Abstract

Four observers were trained in lameness assessment using a subjective scoring system with five categories, and observer agreement was investigated four times at different stages of training and experience. Inter-observer reliability increased with time and reached acceptable levels in the last session. Retrospectively simplified versions of the scoring system were satisfactorily reliable already at a fairly low training level. For experienced raters, the original scoring system with five categories is suitable in terms of reliability for on-farm welfare assessment.

Keywords: *animal welfare, dairy cattle, lameness scoring, observer agreement, reliability*

Introduction

As lameness is highly prevalent in dairy cattle (eg Mülleder & Waiblinger 2004) and considered to be associated with pain and impaired behaviour (eg Vermunt 2004), it should be one of the major foci of on-farm welfare assessment. Therefore, it is necessary to provide a feasible and reliable lameness assessment tool for researchers, veterinarians and farmers. Subjective scoring systems, as they are also used in most epidemiological lameness studies, appear to be particularly suitable for on-farm assessment and monitoring of impaired walking ability as they do not require any equipment. However, it is necessary to test for inter-observer reliability (IOR) and train observers until IOR is acceptable. In association with an epidemiological study on lameness in dairy cattle we tested IOR for a subjective assessment system with five categories.

Materials and methods

Four observers, three of them inexperienced, were introduced to a slightly modified version of the lameness scoring system of Winckler and Willen (2001; Table 1). The on-farm lameness scoring was carried out at six to eight different dairy farms in Austria and Germany. Four farms were visited during three assessment sessions before the beginning of the epidemiological lameness study. The fourth session took place after the study on two (with four observers) and four (with only two observers) farms. A minimum of 183 cows were scored by each observer pair; one observer pair scored 298 cows.

Between sessions, observers practised locomotion scoring with both live cows and video clips. In between the 3rd and 4th session, observers visited 35 to 49 farms, scoring at least 25 cows on each as part of the epidemiological study. Within the study, two observers scored gaits together, the other two worked separately. All farms included in the study had loose housing systems. Gait scoring was performed within the cows walking areas (alleys and outside runs where provided). Cows were encouraged to walk by one observer walking behind them and calling and/or arm waving where necessary. During IOR sessions observers took turns in walking cows. Each cow was walked until all observers were sure about her gait score, while observers were free to watch the cow from any perspective.

IOR was calculated for all observer pairings after each scoring session, using the Prevalence Adjusted Bias Adjusted Kappa (PABAK)-statistic (Byrt *et al* 1993; Abramson 2004):

PABAK = ([k × p]-1)/(k-1). Where k is the number of categories and p the proportion of matchings between observers.

PABAK can reach values up to 1. Values above 0 show a positive correlation between observer's ratings. Matchings are only counted, if both observers give exactly the same score. As for Kappa, PABAK values were interpreted as fair to good agreement if larger then 0.4 and excellent agreement if equal or larger than 0.75 (Fleiss *et al* 2003 cf Woodward 2005).

Spearman correlations (SPSS 12.0) between the number of session and the PABAK values of all observer pairs at these

Table 1 Modified version of a gait scoring system by Winckler and Willen (2001) with five categories, and two retrospectively simplified versions, with four and two categories only.

	Gait description by Winckler and Willen 2001 (slightly modified)	Non-lame categories merged	Non-lame and lame categories merged
no. of categories	5	4	2
1	normal gait	non-lame	non-lame
2	uneven gait (stiff/very careful/swinging of legs around the udder/swaying of trunk and/or hindquarters)		
3	short striding gait with one limb (even if just noticeable)	short striding gait with one limb (even if just noticeable)	lame
4	short striding gait with more than one limb or strong reluctance to bear weight on one limb	short striding gait with more than one limb or strong reluctance to bear weight on one limb	
5	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	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 Mean-, maximum- and minimum-PABAK-values and % of PABAK-values larger than 0.4 over four sessions among four observers (resulting in 6 pairs). The scoring system originally had five categories and was retrospectively simplified into a four and two category systems (see Table 1).

		Session (number of cows; number of farms)			
No. of categories in system		1 (68-90; 2)	2 (21; 1)	3 (42-52; 1)	4 (50-144; 2/4)
5	Mean	0.37	0.41	0.38	0.53
	Maximum	0.43	0.52	0.46	0.68
	Minimum	0.32	0.29	0.25	0.43
	% above 0.4	33	50	33	100
4	Mean	0.60	0.65	0.39	0.52
	Maximum	0.74	0.87	0.47	0.65
	Minimum	0.52	0.56	0.26	0.41
	% above 0.4	100	100	67	100
2	Mean	0.59	0.60	0.62	0.70
	Maximum	0.72	0.90	0.81	0.88
	Minimum	0.44	0.43	0.35	0.44
	% above 0.4	100	100	83	100

four sessions were calculated to test whether observer agreement increased with experience.

We also calculated PABAK-values for differentiating between lame and non-lame cows only, by merging non-lame (1, 2) and lame (3–5) categories (see Table 1).

As the focus of our study was on lameness, we also tested the effect of only merging the two categories for non-lame cows, resulting in a system with 4 categories and three degrees of lameness.

Results

During the first scoring session, only two observer pairs reached a PABAK above 0.4 based on the five-category system. PABAKs increased significantly with time (*rho* = 0.516, *P* = 0.01), yet the aim of all PABAKs being above 0.4 was reached only in session 4. Moreover, the highest PABAK (0.68) was found in session 4 between those observers who went on-farm together during the epidemiological study. The highest average PABAK of all six pairings also occurred in the 4th and last session, with a value of 0.53.

However, PABAKs for differentiating between non-lame and lame cows were much higher: only one out of all 24 PABAKs (4.2%) was smaller than 0.4 with a value of 0.35. Although not significant, lame/non-lame-PABAKs seemingly increased with time. As with the five-category

https://doi.org/10.1017/S0962728600031158 Published online by Cambridge University Press

system, the highest averaged PABAK was reached in the fourth session, with a value of 0.70 (Table 2). The highest PABAK of one observer pair occurred in the second session with a value of 0.90.

With the four-category system, only two out of 24 PABAKs (8.3%) lay below 0.4, the lowest being 0.26. Both of them occurred in the third session and both with the same observer. The maximum PABAK reached in the four category system was 0.87 in session two; averaged PABAKs were between 0.39 and 0.65. There was no increase of PABAKs with time in the four category system.

Discussion

The best IOR was reached when differentiating between non-lame and lame cows only. In the original five-category system some of the insufficient PABAK-values during the first three sessions were due to disagreements within non-lame cows. Merging the two non-lame categories (score 1 and 2) resulted in a system with better PABAK-values. Using the scoring systems with two or four categories led to acceptable IOR after only a relatively small amount of training. Thus, if the aim of the lameness assessment is mainly the detection of lame cows it seems reasonable to favour the four-category system.

The gait described by score 2 ('stiff, very careful gait') typically occurs in healthy animals on a very slippery floor, or if cows have a very voluminous udder or slight leg deformations ('swinging around udder, swaying with body'). In these situations, locomotor behaviour is impaired, even though the animal is not lame. Thus, distinguishing between non-lame gait characteristics provides information about the existing floor quality from the perspective of a cow or the influence of (genetically predisposed) body conformation traits, which can be useful depending on the focus of a study.

However, for the five-category system a fair-to-good agreement of all observer pairs was only reached after considerable practice. As the outstanding IOR result of the two observers who scored cows on 45 farms together suggests, the opportunity of rating and discussing cow gait on-farm will benefit agreement. We thus recommend a combination of video clips with a considerable amount of joint live scoring. Taking the growing amount of international lameness studies into account, a publicly available pool of cow gait clips might increase comparability of research. Further research is needed to clarify which training methods are most efficient.

Our results highlight that care should be taken when comparing values for lameness prevalence from different studies based on subjective scoring systems, as measures for IOR between studies usually are not available.

Conclusions

In this study, inter-observer agreement increased over four sessions. A fair-to-good agreement of all four raters was reached in the fourth and last session after intensive training and practice.

A retrospective simplification of the five category system to a four (one category for non-lame cows and three degrees of lameness) or two (non-lame, lame) category system resulted in sufficient IOR at a relatively low training level.

Acknowledgments

We'd like to thank all farmers and cows who participated in this study.

The present study is part of the Welfare Quality® research project which has been co-financed by the European Commission, within the 6th Framework Programme, contract no. FOOD-CT-2004-506508. The text represents the authors' views and does not necessarily represent a position of the Commission who will not be liable for the use made of such information.

References

Abramson JH 2004. WINPEPI (PEPI-for-Windows) computer programs for epidemiologists. *Epidemiologic Perspectives & Innovations 2004:* 1-6 www.epi-perspectives.com/content/1/1/6

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

Fleiss JL, Levin B and Paik MC 2003 *Statistical Methods for Rates and Proportions, 3rd edition.* John Wiley & Sons: New York, USA

Mülleder C and Waiblinger S 2004 Analyse der Einflussfaktoren auf Tiergerechtheit, Tiergesundheit und Leistung von Milchkühen im Boxenlaufstall auf konventionellen und biologischen Betrieben unter besonderer Berücksichtigung der Mensch-Tier-Beziehung. *Endbericht zum Forschungsprojekt 1267*, Eigenverlag, Wien. [Title translation: Analysis of factors affecting welfare, health and production parameters of dairy cows in cubicle loose housing systems on conventional and organic farms with a special focus on the human-animal relationship]

Vermunt J 2004 Herd lameness – a review, major causal factors, and guidelines for prevention and control. In: *Proceedings of the 13th International Symposium and fifth Conference on Lameness in Ruminants* pp 3-18. Maribor, Slovenia

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 30:* 103-107

Woodward M 2005 *Epidemiology – Study Design and Data Analysis, 2nd edition.* Chapman & Hall/CRC: Boca Raton, USA