J. Dairy Sci. 98:4560–4571
http://dx.doi.org/10.3168/jds.2014-9266

Experienced and inexperienced observers achieved relatively high within-observer agreement on video mobility scoring of dairy cows

E. Garcia,*†[1] K. König,* B. H. Allesen-Holm,‡ I. C. Klaas,* J. M. Amigo,† R. Bro,† and C. Enevoldsen*
*Centre for Herd-Oriented Education, Research and Development (HERD), Department of Large Animal Sciences, University of Copenhagen, Grønnegaardsvej 2, DK-1870, Frederiksberg C, Denmark
†Department of Food Science, Spectroscopy and Chemometrics, and
‡Department of Food Science, Sensory and Consumer Science, University of Copenhagen, Rolighedsvej 26, DK-1958, Frederiksberg C, Denmark

ABSTRACT

Assessment of lameness prevalence and severity requires visual evaluation of the locomotion of a cow. Welfare schemes including locomotion assessments are increasingly being adopted, and more farmers and their veterinarians might implement a locomotion-scoring routine together. However, high within-observer agreement is a prerequisite for obtaining valid mobility scorings, and within-observer agreement cannot be estimated in a barn, because the gait of cows is dynamic and may change between 2 occasions. The objective of this study was to estimate the within-observer agreement according to the observers' educational background and experience with cattle, based on video recordings with very diverse types of gait. Groups of farmers, bovine veterinarians, first- and fourth-year veterinary students, researchers, and cattle-inexperienced sensory assessors evaluated mobility using a 5-point mobility score system developed specifically for walking cows (n = 102 observers). The evaluation sessions were similar for all groups, lasted 75 min, and were organized as follows: introduction, test A, short training session, break, and test B. In total, video recordings of 22 cows were displayed twice in a random order (11 cows in each test × 2 replicates). Data were analyzed applying kappa coefficient, logistic regression, and testing for random effects of observers. The crude estimates of 95% confidence interval for weighted kappa in test A and B ranged, respectively, from 0.76 to 0.80 and 0.70 to 0.75. When adjusting for the fixed effects of video sample and gait scoring preferences, the probability of assigning the same mobility score twice to the same cow varied from 55% (sensory assessors) to 72% (fourth-year veterinary students). The random effect of the individual observers was negligible. That is, in general observers could categorize the mobility characteristics of cows quite well. Observers who preferred to assess the attributes back arch or the overall mobility score (based on uneven gait) had the highest agreement, respectively, 69 or 68%. The training session seemed insufficient to improve agreement. Nonetheless, even novice observers were able to achieve perfect agreement up to 60% of the 22 scorings with merely the experience obtained during the study (introduction and training session). The relatively small differences between groups, together with a high agreement, demonstrate that the new system is easy to follow compared with previously described scoring systems. The mobility score achieves sufficiently high within-observer repeatability to allow between-observer agreement estimates, which are reliable compared with other more-complex scoring systems. Consequently, the new scoring scale seems feasible for on-farm applications as a tool to monitor mobility within and between cows, for communication between farmers and veterinarians with diverse educational background, and for lameness benchmarking of herds.

Key words: mobility scoring, animal welfare, sensory assessor, within-observer agreement

INTRODUCTION

Mobility scoring is a tool used to monitor lameness prevalence and welfare, and to compare dairy cows within and across herds (Manske, 2002; Archer et al., 2010; Barker et al., 2010; Chapinal et al., 2013). If a farmer implements a mobility-scoring routine, he will often coordinate this task with his veterinarian, other herd advisors, and possibly the farm personnel. Yet, achieving common understanding between farmers and veterinarians has been a recurrent discussion theme among international experts, because farmers underestimate lameness prevalence and its economic effect (Whay et al., 2002; Amory et al., 2006; Borderas et al., 2008). Some studies suggest that most farmers do not assess lameness as logically, nor as consistently,

Received December 19, 2014.
Accepted March 26, 2015.
[1] Corresponding author: zjg203@ku.dk

as a trained researcher (Reader et al., 2011). Possible explanations for this could be insufficient training of the farm personnel to detect lame cows and absence of direct financial incentives or lack of motivation to reduce lameness prevalence. Recent review articles have questioned the validity of visual scoring systems (Nielsen et al., 2014; Schlageter-Tello et al., 2014b), because these assessments may be subjective. Obviously, the results of mobility scoring will be invalid for comparison of individual cows, and for groups of cows, if the same observer scores the same type of movement (mobility score) differently, for example, on 2 consecutive days (or at 2 different time points within day). As the between-observer agreement cannot logically be better than the within-observer agreement of 2 given observers, the first step in a validation process must be the assessment of the within-observer agreement (i.e., intrarater or intraassessor agreement). Specifically, do observers agree with themselves when scoring a sample replicate twice?

Few studies assess within-observer agreement compared with the number of studies available that assess between-observer agreement (Schlageter-Tello et al., 2014b). Typically, the percentage of exact agreement (**PA**) is used to estimate within-observer agreement. Using the original score, PA ranged between 30% using a 9-point scale and 56% using a 5-point scale (Manson and Leaver, 1988; O'Callaghan et al., 2003; Schlageter-Tello et al., 2014b), whereas a study based on unweighted kappa statistics (κ) reported values from 0.30 to 0.68 (Thomsen et al., 2008). However, these estimates seem not to account for the effect of the observer experience or other sources of variation (e.g., the cows present in the study) when evaluating the scoring system.

Logically, mobility-score systems will only be accepted if they have practical value to farmers and veterinarians in the long run, and practical value often comes from making appropriate decisions at the cow level (e.g., to treat or cull given certain criteria). If farmers and veterinarians disagree with themselves, they obviously cannot agree on a joint criterion for decision making. Observers with considerable cattle experience obtained 60 to 83% within-observer PA while gait scoring on video, yet some pairs of observers encountered a much lower between-observer agreement (Schlageter-Tello et al., 2014a). Lower between-observer agreement than within-observer agreement can be explained by simple probability calculations. If 2 observers both obtain 60% within-observer agreement, the chance to agree on a given score cannot be more than 0.6 × 0.6 = 0.36 (36%). If the observers use the scoring scale differently, between-observer agreement will be even lower.

Most studies evaluating mobility-scoring systems are conducted on-farm, but because the gait of a cow is dynamic and often influenced by turbulent conditions in the barn (e.g., other cows pushing), it is virtually impossible to obtain true replicates. Furthermore, often the number of observers who can assess a cow from exactly the same visual angle is limited because of space constraints, especially when mobility is assessed while herdmates are present. Thus, it is not possible to discriminate the variation due to the cow's gait from the observer variation, when scoring at 2 different time points within-day. The only practically relevant alternative is to show and assess the same video sample twice. To the best of our knowledge, the largest study assessing within-observer agreement with video included up to 10 experienced observers (Schlageter-Tello et al., 2014a). Therefore, little is known about the within-observer agreement among farmers, students, and veterinarians outside research environments and which factors may influence their performance. Hence, we designed an experiment of video-based mobility scoring using a newly developed 5-point scale with the purpose of assessing within-observer agreement of veterinary students, farmers, bovine veterinarians, researchers, and food sensory assessors based on true replicates. Including a randomized setup of different gait types, we hypothesized that the background education, experience, and gait-scoring preferences affect agreement level. The main objective of this study was to investigate the within-observer probability of perfect agreement in groups of observers with widely different educational background. The specific objectives were (1) to estimate the effect of individual observers; (2) to estimate the effect of widely different types of mobility on video based on a predefined 5-point level scale; (3) to explore the effects of background education, experience, and gait-scoring preferences on within-observer agreement; and (4) to validate a 5-point mobility-score scale for experienced and inexperienced observers.

MATERIALS AND METHODS

Development of Video Questionnaire

An online questionnaire with video samples was developed based on recordings of lactating dairy cows walking on farm. We collected samples from 4 dairy herds located in Zealand, Denmark, between October and December 2013, with the objective of maximizing variation in gait (or mobility). Table 1 describes the mobility score (named König-Garcia mobility score). In contrast to some other scoring systems that require assessment of both standing and walking cows, this scoring system was specifically developed to enable scoring while walking, because it is difficult to get an opportunity to see cows standing *and* walking under

Table 1. König-Garcia mobility score—developed specifically for the video questionnaire to enable scoring without seeing a cow standing[1]

Score	Description
1	**The cow does not have uneven gait at all** • No signs of uneven gait • Even weight bearing between legs • The back is flat while walking. • No signs of head bob.
2	**The cow has a slightly uneven gait** • The cow might walk almost normally. • The gait is often slightly uneven, and the cow is likely to take shorter strides. • But, no signs are evident of limping or uneven weight bearing. • The back might be arched. • No signs of head bob
3	**The cow has uneven gait** • Abnormal gait pattern • Walks with short strides on one or more legs • In most cases, a trained observer will be able to tell which leg is affected (mild signs of uneven weight bearing). • The back is usually arched. • But, signs of head bob may or may not exist.
4	**The cow has a very much uneven gait** • The cow is obviously limping on one or more legs. • An untrained observer will usually be able to detect the affected legs. • In most cases, the back is arched. • Head bob will be evident.
5	**The cow has an extremely uneven gait** • The cow is unable, unwilling, or very reluctant to bear weight on the affected legs. • The back is arched. • Head bob is evident and can be extreme.

[1] The scoring system is based on sensory science and previous lameness-scoring systems (Sprecher et al., 1997; Thomsen et al., 2008; Barker et al., 2010). The scores do not include treatment indications but simply focus on the ability of the cow to move.

practical conditions. For developing the scoring scale, we followed sensory science guidelines and revised previously reported scoring systems (Sprecher et al., 1997; Thomsen et al., 2008; Barker et al., 2010) to obtain uniform distances between all score values. The definitions of the 5 mobility scores excluded treatment recommendations and diagnoses, to avoid inclusion of personal attitudes and traditions for use of medical treatment. Focused on gait attributes and back arch, we gradually created different categories of mobility to obtain a truly ordinal score (Table 1). The gait of a cow was recorded using a Sony HDR-CX740 video camera (1,440 × 1,080 pixels, 25 frames per second) or a Samsung Galaxy GT-I9300 smartphone (1,280 × 7,20 pixels, 29 frames per second), while one person, walking from behind, gently encouraged each cow to walk along the alleys. The herds had 108 to 303 lactating cows and used robotic milking systems or rotary parlor. Cows were kept in free-stalls with slatted floors (one farm with rubber mats on top, which seemed more slippery). The videos, clipped into short sequences (5 to 12 s) using VLC media player (v. 2.1.3, VideoLAN, Paris, France), were critically selected based on the overall film quality (time length, clear space, good background contrast, and no visual obstructions) and to achieve similar patterns of mobility in both tests when allocating them to tests A and B. Both tests were composed of 11 unique video sequences (samples) shown in replicates. Each video sample was assigned a mobility score as assessed by the authors. Scores 1, 2, 3, 4, and 5 were shown with a sample size of 1, 3, 3, 2, and 2 cows, respectively, in each test. A test was composed of 22 samples, replicate 1 and 2 of the 11 original samples. Hence, 11 unique videos were shown twice yet with random order within each replicate set of test A and test B, being the same order for all observers. The farm origin, cow number and history, and authors' score were blinded.

Selection of Observers

We organized 7 independent sessions in February and March 2014, where answers from 108 observers were registered (Table 2). To maximize diversity of observers, we selected the following groups: bovine veterinarians undergoing a cattle continuing-education course, farmers undergoing a cattle specialization in 2 agricultural schools, veterinary students in the first and fourth year of their curriculum at the University of Copenhagen (**UCPH**), researchers from the Department of Large Animal Sciences at UCPH, and sensory food assessors from the Department of Food Science (Sensory and Consumer Science section) at UCPH. The inclusion criterion for the sensory panel was that they should not have any experience with dairy cows and should be familiar with participation in sensory tests. The educational background was labeled as follows: farmer 1 and

Table 2. Overview of the video-questionnaire sessions and study observers described by sample size, age, sex, continuing-education frequency, and experience

Session	Sample size	Age range, yr	Male/female, %	Continuing education more than once a year, no. (%)	Inexperienced observers, no. (%)
1. Bovine veterinarians	8	31 to 41	25/75	6 (75)	0 (0)
2. Yr-4 veterinary students	30	21 to 35	13/87	0 (0)	14 (47)
3. Farmers 1	9	19 to 23	55/44	2 (22)	0 (0)
4. Yr-1 veterinary students	12	19 to 28	100/0	1 (8)	10 (83)
5. Farmers 2	22	19 to 27	73/27	7 (32)	0 (0)
6. Sensory assessors	9	23 to 42	11/89	0 (0)	9 (100)
7. Researchers	12	25 to 43	17/83	3 (25)	1 (8)
All sessions	102	19 to 43	28/72	19 (19)	34 (33)

farmer 2 (session 3 and 5, respectively), sensory assessors (session 6), researchers (session 7), yr-1 and yr-4 students (sessions 4 and 2, respectively), and bovine veterinarians (session 1). The experience of the observers regarding work with dairy cattle was categorized as follows: low (little or no experience), medium (1 to 10 yr), or high (more than 10 yr).

Instructions and Experimental Design

The second author was the panel leader and presented the guidelines for the test by reading from a script to ensure that all observers in all sessions received the same information in exactly the same way. The whole session was conducted in Danish. Figure 1 shows a flowchart of the session.

The session started with a short introduction regarding the study purpose and describing the process. The observers were asked not to discuss, nor to look at each other's answers during the tests, stating the importance of registering their very own perception of the score. Also, we took the following steps to ensure everyone got acquainted with the procedures, including the web-survey platform (named "Absalon," UCPH) where answers were registered. For example, the observers were encouraged to ask questions if they were in doubt about the instructions. The video samples were projected on a large screen, while the panel leader led the observers through the session and indicated when they had to move to the next question on the web survey, to ensure that all were on the same question and had it answered. Before test A, the panel leader introduced verbally and visually the König-Garcia mobility score (translated Danish version) for approximately 3 min with video examples (from FirstStep; Zinpro Corp., 2009) and handed out paper copies of the mobility-scoring scale, which could be consulted freely by the observers during the tests. Before test A, flexible time was included for 3 extra video samples meant for adaptation and to be excluded from data analysis (mobility score 2, 4, and 3 according to authors' assessments). However, this information was blinded simply to allow observers to be comfortable with the questionnaire and the web platform without stressing them with time constraints (called "dummy samples" in sensory science). Then, the 22 samples followed, and the assessment time was roughly 30 s per sample (each video sample played twice + 10 s for answering). After test A, all the observers were supposed to be acquainted with the context and the diversity of gaits. Therefore, we conducted a training session, which stressed gait traits that had been associated with lameness in previous studies and can be assessed in a short time. Five individual attributes met these criteria: back arch, head bobbing, gait asymmetry, stride length, and pain. All these could be scored from 1 to 5, where score 1 corresponded to a normal mobility and score 5 corresponded to extremely abnormal. The training scheme was handed out in paper after test A. The observers were asked to score 3 cows selected from test A, ordered upon the authors' score: score 1, score 3, and score 5, so that they could gradually experience the increasing lameness degree (or decreasing mobility). The observers again had the opportunity to see each of the samples twice per attribute, including a final mobility score (sixth attribute) after scoring the individual attributes. At the end of the training session, the observers were asked to reflect upon the attributes used in the training and to indicate personal gait-scoring preferences: easiest and hardest of the 6 attributes they had scored. After registering

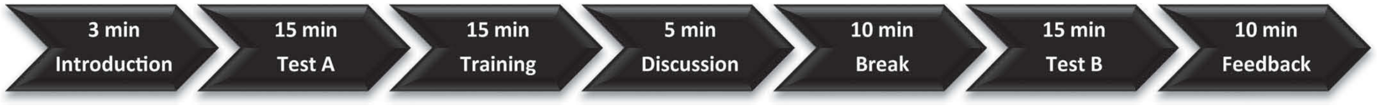

Figure 1. Flowchart of the session outline and respective approximate duration of each part (minutes). The whole session lasted roughly 75 min.

their scores for the training samples, the observers were asked to share and discuss their answers on the mobility score and attributes. The training could then vary according to the observers' engagement. Because of time constraints, if the discussion would get too long, the panel leader would wrap-up the comments trying to obtain observer consensus. To consolidate the learning process, the panel leader would end the discussion by explaining how a cow might not have the same score in all attributes including mobility score, but rather that there would be a mobility score that most likely generates agreement. Then, observers were asked to take a 10-min break away from the screen. Test B followed similar to test A, including new samples not shown in test A but this time without dummy samples. Finally, the observers were asked to give written and oral feedback on the test experience. Demographic data were collected as well as open-end questions related to lameness and observer experience with dairy cattle and frequency of continuing education. All sessions were held before lunch, took place in classroom-like facilities, and lasted approximately 75 min including the break. Image mirroring was introduced on both tests A and B for the replicate 2 set from session 3 to 7, to minimize the risk that observers would remember cows previously scored. Finally, it is noteworthy that this experimental design was a compromise between different practical and time constraints: (1) sensory studies need to include breaks because observers get tired; (2) observers were volunteers so the amount of time and samples had to be limited because (3) our purpose was to study within-observer agreement, which called for replicates; (4) maximizing the diversity of the mobility patterns was considered more important than including a larger sample size in regard to videos; however, (5) more than one sample per score level was included in the abnormal mobility levels and (6) the most challenging levels (score 2 and 3) had the highest number of samples.

Data Editing

Missing data on some answers occurred in 11 out of 108 observers (10%). The farmers' teacher was excluded from analysis. Two observers were excluded because we detected that their responses were misaligned with the question number of the questionnaire. Finally, 3 observers were excluded because their answers regarding observer characterization were missing (preferences or education), resulting in a data set of 102 observers × 22 video samples. As all samples had 2 replicates, we expected 2,244 pairs of answers ($102 \times 22 \times 2 = 4{,}488$), but because 8 mobility-score answers from 6 observers were missing, this resulted in a total of 2,236 pairs of answers available for analysis ($2{,}244 - 8 = 2{,}236$). The replicate scores were cross-classified. Because the major interest was to describe perfect within-observer agreement and factors with major influence on the agreement, we dichotomized the cross-classifications into perfect agreement (1) and deviation between replicate assessments (0). Because the frequency of more than one unit difference was rather small (3%), we used the degree of deviation for estimation of crude kappa values only (see below).

Data Analysis

We used a multilevel logistic regression model to estimate the probability of perfect agreement (1 vs. 0), with random effect of observers, which we tested against a null model. Because little evidence existed for the effect of observers, as well as in models with additional explanatory variables, the random observer effect was removed in the final model. Then, we tested the fixed effects of video sample and session given by design, in addition to effects related to the observers (preferred and not preferred attributes, experience, frequency of continuing education, age, and sex). The model fit was evaluated using the Akaike information criterion. We tested 2-way interactions, but these did not decrease the Akaike information criterion compared with the simple model. The models were run via maximum likelihood with the Laplace approximation using the glmer function of the lme4 R package (Bates et al., 2014) or the glm R function, and statistical significance of main effects was assessed using drop1 and anova R functions. Estimates of least squares means (**LSM**) were computed using lsmeans R package (Lenth, 2014). Also, for comparison with previously reported scoring systems, we provide estimates for the range and median of unweighted and weighted kappa values for within-observer agreement, κ and κw, respectively, in addition to overall kappa values for test A and test B with 95% CI. Kappa estimates were calculated using the R package vcd (Meyer et al., 2014). The crude percentage of agreement and 95% CI for the overall results of test A and B were calculated with the exactci function from R package PropCIs (Scherer, 2014). Significance level was set at 0.05. Data analysis was done using R software 3.0.2 (R Core Team, 2013) and RStudio (v. 0.98.501, 2009–2013 RStudio Inc.).

RESULTS

Descriptive Statistics

Table 2 describes the sessions where data of 102 observers was kept for analysis. A total of 34 observers

Table 3. Crude cross-classification of scores obtained from answers of 102 observers on a mobility-scoring video questionnaire using 22 video samples (n = 2,236 because of 8 missing answers)[1]

	Score of replicate 2						
Item	1	2	3	4	5	Subtotal	Frequency, %
Test A[2]							
Score of replicate 1							
1	**226**	72	3	0	0	301	27
2	71	**170**	35	0	0	276	25
3	5	49	**84**	40	0	178	16
4	1	4	29	**168**	29	231	21
5	0	0	0	25	**110**	135	12
Subtotal	303	295	151	233	139	1,121	
Frequency, %	27	26	13	21	12		
Test B[3]							
Score of replicate 1							
1	**177**	86	16	2	0	281	25
2	55	**203**	72	11	1	342	31
3	12	65	**111**	31	2	221	20
4	0	1	14	**80**	26	121	11
5	0	0	1	15	**134**	150	13
Subtotal	244	355	214	139	163	1,115	
Frequency, %	22	32	19	12	15		
Total						2,236	

[1] Test A and B each had 11 samples shown in replicates and randomized. Test B was done after a short training session. Agreement estimates kappa (κ), weighted kappa (κw), and percentage of perfect agreement (PA) are given in footnotes below with the corresponding 95% CI.

[2] κ = 0.59 (95% CI: 0.55 to 0.62), κw = 0.78 (95% CI: 0.76 to 0.80), and PA = 68% (95% CI: 65 to 70%).

[3] κ = 0.53 (95% CI: 0.49 to 0.56), κw = 0.72 (95% CI: 0.70 to 0.75), and PA = 63% (95% CI: 60 to 66%).

(33%) had no experience with cattle, mostly among the students and sensory assessors. A total of 2,236 replicate assessments were available and are presented in Table 3. Perfect agreement (same score on replicate 1 and 2) was obtained in 65% of the answers (n = 1,463), one unit deviation occurred in 32% (n = 714), and 3% had more than one unit deviation (n = 59, <5% at session level). During the final comments, most observers mentioned there were some repeated cows. The difference between the authors' score (consensus agreement) and the overall mean score for all observers at each replicate ranged from −0.2 to 1.7 score units, where the 10th and 90th percentile were, respectively, −0.10 and 0.90 score units. In other words, in less than 10% of the answers the observers disagreed with the authors' score by more than one score unit deviation. Regarding within-observer agreement estimated with kappa values, pooling the answers from both tests for each observer, returned a median κ = 0.57 (range: 0.17 to 0.94) and a median κw = 0.76 (range: 0.46 to 0.97). Alternatively, as shown in Table 3, the pooled within-observer agreement estimates of the whole test A and test B were, respectively, κ = 0.59 (95% CI: 0.55 to 0.62) and κ = 0.53 (95% CI: 0.49 to 0.56), κw = 0.78 (95% CI: 0.76 to 0.80) and κw = 0.72 (95% CI: 0.70 to 0.75), PA = 68% (95% CI: 65 to 70%) and 63% (95% CI: 60 to 66%).

Factors Affecting Within-Observer Agreement Level

Session, video sample, and preferred attribute statistically significantly influenced within-observer agreement, as described in the final model (Table 4). Year-4 students achieved the numerically highest average probability of perfect agreement (72%) and differed statistically significantly from the averages of the sessions for sensory assessors (55%) and yr-1 students (60%). However, the random effect of individual observers was close to zero and thus removed. We did not find in the final model a statistically significant effect of the hardest attribute (gait-scoring preferences), experience, frequency of continuing education, age, or sex.

Figure 2 shows the LSM and 95% confidence limits for agreement by session. Figure 3 shows a much more pronounced effect of video sample (cow) on mobility score than the effects of session and gait-scoring preferences. The 5 video samples with the highest PA were at the extremes of the scale, 2 video samples close to score 1 and 3 video samples close to score 5. These differed statistically significantly from the 4 samples in the lower end, mostly samples in which the mean score was close to score 2 and 3. In test A, the LSM probability of perfect agreement for the 11 samples ranged from 54 to 90%, compared with 45 to 84% in test B.

For an average video sample (~score 3) and average observer, the highest probability of perfect agreement

Table 4. Model summary with LSM probabilities of perfect agreement (logistic regression), accounting for fixed effects of session, video sample (not shown), and preferred attribute[1]

Variable	Probability of perfect agreement (LSM)	95% CI	LSM significance
Session			
6. Sensory assessors	0.55	0.48 to 0.63	a
4. Yr-1 veterinary students	0.60	0.53 to 0.67	a
7. Researchers	0.62	0.55 to 0.68	ab
3. Farmers 1	0.63	0.55 to 0.70	ab
1. Bovine veterinarians	0.65	0.57 to 0.72	ab
5. Farmers 2	0.66	0.61 to 0.72	ab
2. Yr-4 veterinary students	0.72	0.67 to 0.76	b
Preferred attribute			
Pain	0.55	0.42 to 0.67	ab
Gait asymmetry	0.57	0.51 to 0.63	a
Head bobbing	0.65	0.61 to 0.69	ab
Stride length	0.66	0.55 to 0.76	ab
Mobility score	0.68	0.64 to 0.72	b
Back arch	0.69	0.64 to 0.74	b

[a,b]Different letters correspond to significant differences ($P < 0.05$).

[1]The effect of sample is presented in Figure 3.

(LSM) was achieved by observers who assessed the overall mobility score (68%) or the back arch (69%) as easiest attributes, whereas the lowest was achieved by those preferring gait asymmetry (57%) or pain (55%). Table 4 and Figure 4 show all attributes with confidence limits.

DISCUSSION

The aim of this study was to estimate within-observer agreement among a highly diverse population of observers using video samples and an on-line survey platform to score mobility of dairy cows. To the best

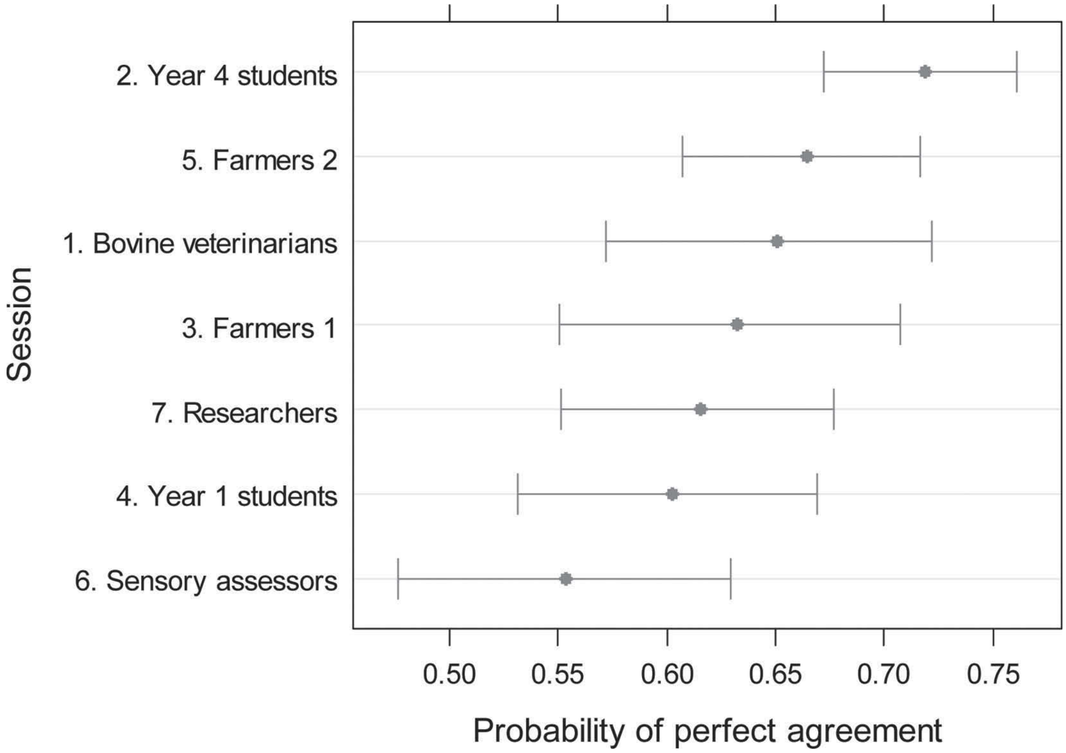

Figure 2. Predicted probability of perfect agreement with 95% confidence limits ranked according to each session for the average video sample and average gait-attribute preferences (LSM, logistic regression). Year-4 veterinary students differ significantly from sensory assessors and yr-1 veterinary students.

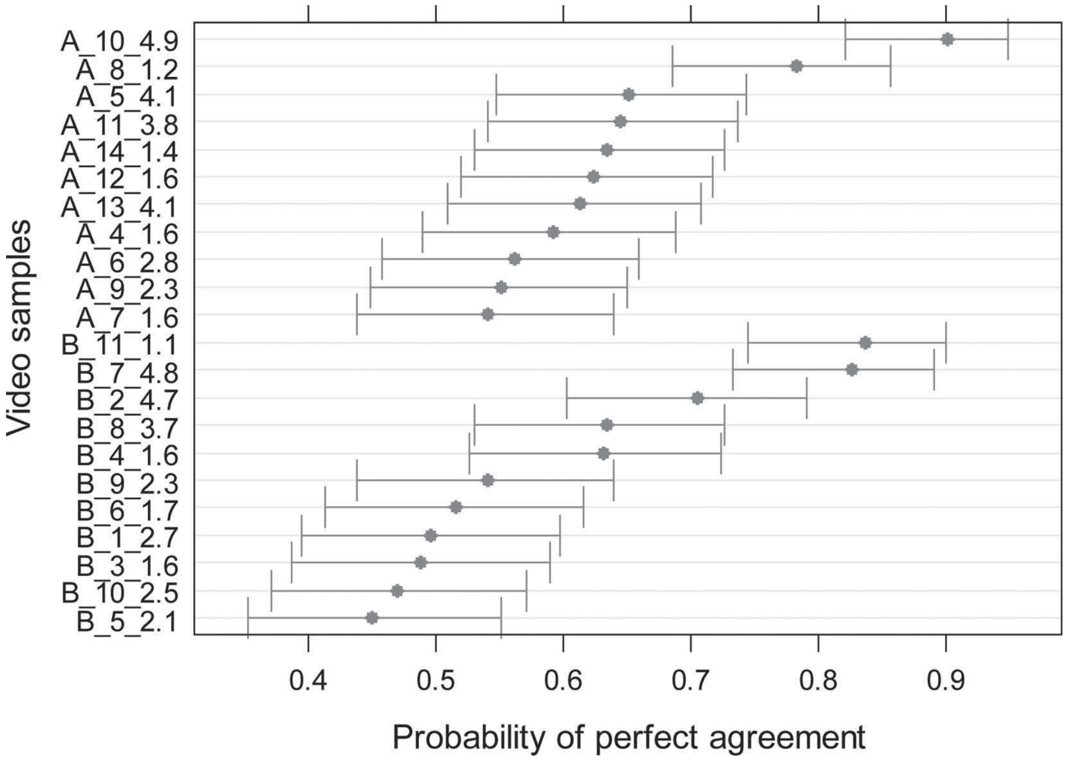

Figure 3. Predicted probability of perfect agreement with 95% confidence limits ranked by test and video sample for the average observer and average gait-attribute preferences (LSM, logistic regression). The video sample label contains information about the test group and test order of replicate 1, followed by the overall mean score calculated based on all observers. Video samples close to score 1 and 5 have the highest probability of perfect agreement, and several videos differ significantly, such as samples with a mean score close to score 2 and 3.

of our knowledge, this is the largest within-observer-agreement study published. Using a fixed experimental design across 7 sessions, our results indicate differences on within-observer agreement due to background education and preferred attribute of the gait-scoring scale, yet the cow (or type of mobility score) is the most important factor influencing the probability of perfect agreement. As seen in Figure 3 and depending on the sample, the LSM probability of perfect agreement varied roughly between 40 and 90%.

Effect of Background Education

Despite only a brief introduction before test A, most observers achieved a probability of perfect agreement with a new 5-point mobility score, which is relatively high compared with previous studies: for an average observer the probability ranged from 54 to 90% (LSM). As expected, the observers with most cattle-experience—yr-4 students, bovine veterinarians, and farmers—achieved the highest probability of perfect agreement. Remarkably, the occurrence of a 2-unit or more disagreement was negligible (<3%). Regarding the relatively good results of yr-4 students compared with experienced bovine veterinarians, we propose 3 explanations: (1) these students were very much used to the Absalon system and to performing online assessments (e.g., homework, exams); (2) they had prior training in clinical exams and had been doing farm visits on the same week (including lameness-scoring tasks); and (3) they had intrinsic motivation to perform well on evaluations within academia. This combination of factors might be specific of this group, which could have helped them to disregard distracting factors and simply focus on scoring. The results of farmers are surprisingly good, because previous studies claimed farmers usually underestimate lameness prevalence about 3 times (Whay et al., 2003; Espejo et al., 2006; Alawneh et al., 2012). This might be associated with their age range in our study (younger generations might be more critical towards animal welfare) but also with the fact that they had both farm experience and a background education in agriculture.

Agreement is assessed in different ways in previous studies, which makes it challenging to compare findings. We focused on the PA because it can be derived from most studies, for comparisons with our estimates, and has a simple and practical interpretation. In addition, for management decisions in a herd, available records usually must be dichotomized (e.g., as criterion

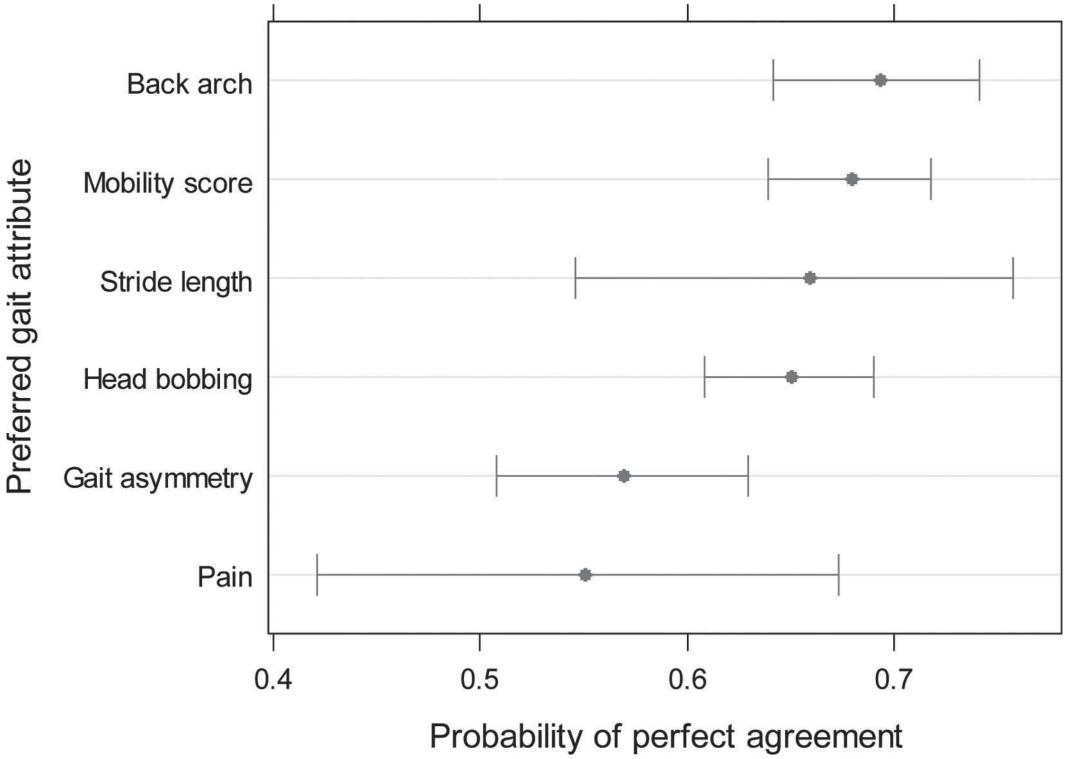

Figure 4. Predicted probability of perfect agreement with 95% confidence limits ranked according to preferred attribute for the average video sample and average observer (LSM, logistic regression). Mobility score (uneven gait) and back arch differ significantly from gait asymmetry.

for medical treatment or culling) and this calls for use of PA to assess the validity of information, whereas kappa or similar indices of concordance cannot be used directly as decision criteria. Manson and Leaver (1988) reported an on-farm within-observer PA of 30% using the original 9-point score, and O'Callaghan et al. (2003) reported 56% using a 5-point score (Schlageter-Tello et al., 2014b). Regardless of the background education, the LSM probability of perfect within-observer agreement in our study was consistently above 50%, and up to 72% in the case of yr-4 students. Consequently, our new scale may be superior to the ones used in those 2 studies. In addition, our estimates were adjusted for significant fixed effects of video sample and preferred scoring attribute. Notably, one-third of the observers in our study considered themselves inexperienced about cattle (Table 2).

Overall, our results are in the range of observers experienced in locomotion scoring (Schlageter-Tello et al., 2014a), who achieved between 60 and 82% of perfect agreement. Alternatively, kappa statistic could also be used for comparisons because it takes into account the agreement by chance for the answers on replicate 1 and 2 (Cohen, 1968). The weighted kappa range in our data was from 0.46 to 0.97, which seems higher than the range described by Thomsen et al. (2008), 0.38 to 0.78 with a similar 5-point scale, and overlaps with the range found by Schlageter-Tello et al. (2014a), between 0.63 and 0.86 using another 5-point scale.

Effect of Years of Experience

Surprisingly, contrary to current belief that training is crucial to achieve high agreement levels, the highest level of experience was not associated with a higher chance of perfect agreement, particularly if taken into account that observers with limited to no cattle experience were present. Possibly, few observers were present with more than 10 yr of experience to show this effect. Or, the introduction before test A might have brought naïve observers to the same agreement level of experienced observers. Overall, it seems to require very limited effort to obtain reasonably good within-observer agreement on the applied mobility score. However, the study also shows that repeated perfect within-observer agreement is highly unlikely.

Effect of Specific Attributes of the Gait-Scoring Scale

The specific gait-scoring attributes had an effect, because the observers who judged the back arch or the mobility score as the easiest attributes had the highest chance of perfect agreement (~70%), as seen in Figure 4. This association could be an indirect indicator of experience or an indicator of the observer's inherited ability to detect abnormal mobility. However, this finding could also mean that the observers who paid attention to those attributes had an advantage, because the arched back has been extensively documented as a characteristic associated with lame cows. Moreover, findings from a study on body condition scoring show that strict adherence to a decision tree for classifying different levels might cause inexact classifications compared with the reference method (ultrasound), as opposed to using the overall impression of the cow and a decision tree (Isensee et al., 2014). Our results seem to support a similar interpretation. The new mobility-scoring system developed for this study had several attributes describing each level in a rather flexible way, but the user would also be guided by the main description of each level regarding the overall impression of uneven gait.

Effect of Video Sample and Experimental Design

The results for test B indicate a systematic minor decrease in agreement after the training. The video recordings for test A and B were critically selected to achieve similar patterns of mobility in both tests. Because test A and B represent different cows, we cannot directly use a comparison of results for estimating training effect. However, to justify the claim that we had a pronounced training effect, we should have observed a marked increase in within-observer agreement in test B, despite minor differences due to gait types. Consequently, because we observed a systematic (minor) decrease in test B, we cannot claim that the applied training efforts improved agreement. The increase in variation in test B could indicate that the observers were more confused by the new information given during training, thus leading to worse agreement. They may have focused more on details and less on the overall impression of uneven gait. On the other hand, the observers, or at least some of them, might have become more confident in using some of the scoring levels, which may be a long-term advantage. Exploration of the raw data showed that observers more often gave score 2 and less often score 1 in test B, in comparison to test A. Eventually, the interpretation of the definition of the 5 levels has become clearer to the observers after the training session, even if the proportion of random variation increased (±1 point). This hypothesis is difficult to test, because it could alternatively be argued that this effect is simply due to different cows in test B. Even though the average video sample length was the same in both tests (~8 s), test B had 2 samples of 5-s length, whereas the minimum length in test A was 7 s.

One way of exploring whether test B was more difficult than test A and led to a decrease in agreement could be to run the same experiment while swapping the tests: test B first, followed by the training session and test A. Nonetheless, several observers from all sessions gave oral or written feedback that demonstrated the instructions and scoring scale were feasible: e.g., "Good to have summarized the facts we need to keep an eye on the farms," farmer; "I felt it was instructive for veterinary students," yr-4 student. Instead, point-light investigation as used in human gait studies could be an alternative to hide the cow identity (Gunns et al., 2002), facilitating the focus on the overall mobility pattern.

Test Experience and Mobility-Scoring System

Even though some observers remarked during the final comments that there were some repeated cows, they still maintained a substantial level of disagreement, evidenced mostly by a 1-point deviation. This indicates the observers have probably memorized some cows but not necessarily the mobility score they previously gave. The use of a video-based questionnaire was effective to guarantee all observers could see the cows from the same angle, and that the study included true sample replicates. This design is optimum for assessing true within-observer error, compared with an on-farm study, because standardized conditions were ensured. Therefore, we can use such estimates to infer upon the theoretical between-observer agreement (cf. probability calculation example in the introduction). However, in a herd-management context 2 observers may face several constraints (e.g., different angle, assess the cow at a different time point, and so on), and then the corresponding within- and between-observer agreement estimates on farm are likely to be lower compared with the results presented herein.

Mobility can be seen as a latent continuous variable, which could call for the use of visual analog scales. However, the use of visual analog scales is problematic without anchoring points (A. Vieira, Centre for Interdisciplinary Research in Animal Health, Lisbon, Portugal, personal communication), which suggests some sort of ordinal scoring system with well-defined levels is essential to guide the observer to distinguish between different scores, e.g., visual analog scales together with multiple anchors (Tuyttens et al., 2009). Our scoring system is ordinal but the descriptions of the 5 levels describe a continuous trait, which might explain the relatively good results achieved. That is, the observers could directly identify in the scale the overall mobility pattern they were seeing. Our results showed that observers preferred to use different gait attributes in the training and that these preferences are associated with the probability of perfect agreement. This indicates a need for dialog about how different people perceive mobility; yet it also reveals an opportunity for improving within-observer agreement with attribute-specific training.

Interpretation Under Different Scenarios

Often diagnostic test evaluation is based on comparison with a reference test (gold standard). Despite highly standardized assessment conditions, none of the numerous observers in this study obtained perfect within-observer agreement. Consequently, a gold standard for mobility scoring based on visual assessment is unlikely to exist in practice. Therefore, a veterinarian and a herd manager will disagree to some level whenever they score the same cow (between-observer agreement). However, it is questionable to assess between-observer agreement if the individual observers agree poorly with themselves when scoring a true replicate twice. If for example, the herd manager implements a strategy to provide a treatment and soft bedding for all cows that score $\geq$4, then, assuming his within-observer PA was 80%, we would expect 20% of his scorings to be incorrect. That is, 20% of scorings, which could be deviating from his average assessment for a given mobility pattern, could represent animals left untreated or animals that the manager did not register the correct state. Therefore, he may not know exactly whether the cow is recovering or getting worse, compared with his previous assessment.

On the other hand, if 2 veterinarians from a given bovine practice perform lameness prevalence assessment in their farms interchangeably for benchmarking purposes, and assuming they have both a within-observer agreement of 80% and use the scale in the same way, then we should expect a between-observer agreement of 64% (0.8×0.8). In other words, 36% of the practice scorings will most likely have a ±1 error, which could be compensated by increasing the sample size.

Finally, let us assume a scenario where 2 technicians from veterinary authorities with the same within-observer agreement use the scale for applying penalties in the presence of untreated score-5 cows in a farm. Yet, they use the score differently: they could be 100% repeatable within observer, thus having excellent kappa values, but one of them could score 3 a mobility pattern that the second technician scores 5. This systematic measurement error (bias) would obviously lead to unfair assessments. Considering these and other specific scenarios of practical implementation, where the within- and between-observer errors may have different effects, with the agreement levels described in this study, we have provided a benchmark for different dairy professionals, which could be useful for comparisons in future studies of within- or between-observer agreement.

Final Remarks

Our study with a large panel was useful to define a mean score for each sample and to identify group differences, where comparisons of agreement can be made between experienced and inexperienced observers. As reported in a recent study (Schlageter-Tello et al., 2014a), we also found higher agreement at low and high ends of the score, particularly, in score 1, 4, and 5, whereas scores 2 and 3 showed lowest agreement. Although it is common practice to merge levels to achieve higher agreement, we claim that this might not be optimal, if for example the score is meant to develop automatic lameness detection systems. The reason is that, when we clearly find a higher agreement level at score 1 than score 2, building a model that uses score-2 cows as a reference for not being lame will lead to a model that considers normal mobility a gait pattern that is not completely sound. Hence, when reporting lameness indicator figures in a herd, it would be relevant to show not only the proportion of very lame cows (score 4 and 5) but also the proportion of very sound cows (score 1). Accordingly, Tadich et al. (2013) demonstrated that an acute-phase protein—used as inflammation and welfare indicator (haptoglobin)—was increased in cows with locomotion score >1, compared with cows with score 1 that had a normal level of plasma concentration, thus concluding it was a sensitive measure of pain due to lameness. However, assessing pain visually is challenging, which was highlighted by the agreement results of observers who preferred this attribute, and more research is needed to understand how objective pain assessments correlate with visual observations.

Further studies of within-observer agreement are essential to estimate how high we should aim in relation to perfect agreement. Additionally, on-farm testing with similar diversity of observers will help to define how suitable scoring systems are for implementation and communication across stakeholders.

CONCLUSIONS

We developed and presented a mobility score that does not require assessment of cows standing. Among groups of observers with diverse educational background and cattle experience, the adjusted probability of assigning the same score twice to the same dairy cow ranged from 45 to 90%, mostly depending on the video sample. The session of yr-4 veterinary students and the preferred attributes back arch and overall mobility score were associated with the highest probability of perfect agreement.

Even observers without prior cattle experience were able to achieve perfect within-observer agreement up to 60% of the time, with merely a 3-min introduction. Reporting of score-1 and score-5 cows can be recommended (highest agreement). Our new mobility score proved to achieve relatively high within-observer agreement and seems feasible for on-farm implementation as a tool to monitor mobility within and between cows and herds and for lameness benchmarking.

ACKNOWLEDGMENTS

We thank the University of Copenhagen (Frederiksberg C, Denmark), namely the Chemometrics Analysis Center (CHANCE) and Centre for Herd-oriented Education, Research and Development (HERD), and the Karla and Svend O. Kochs Foundation for funding this study. We also thank the farmers who welcomed us to collect the video materials and the observers for their collaboration in the questionnaire. The feedback on the study design from Frans W. J. van den Berg and Karina Gleerup (University of Copenhagen) was highly appreciated.

REFERENCES

Alawneh, J. I., R. A. Laven, and M. A. Stevenson. 2012. Interval between detection of lameness by locomotion scoring and treatment for lameness: A survival analysis. Vet. J. 193:622–625.

Amory, J. R., P. Kloosterman, Z. E. Barker, J. L. Wright, R. W. Blowey, and L. E. Green. 2006. Risk factors for reduced locomotion in dairy cattle on nineteen farms in the Netherlands. J. Dairy Sci. 89:1509–1515.

Archer, S. C., M. J. Green, and J. N. Huxley. 2010. Association between milk yield and serial locomotion score assessments in UK dairy cows. J. Dairy Sci. 93:4045–4053.

Barker, Z. E., K. A. Leach, H. R. Whay, N. J. Bell, and D. C. J. Main. 2010. Assessment of lameness prevalence and associated risk factors in dairy herds in England and wales. J. Dairy Sci. 93:932–941.

Bates, D., M. Maechler, B. Bolker and S. Walker. 2014. lme4: Linear Mixed-Effects Models Using Eigen and S4. R package version 1.1–7. R Foundation for Statistical Computing, Vienna, Austria. http://CRAN.R-project.org/package=lme4.

Borderas, T. F., A. Fournier, J. Rushen, and A. M. B. de Passillé. 2008. Effect of lameness on dairy cows' visits to automatic milking systems. Can. J. Anim. Sci. 88:1–8.

Chapinal, N., A. K. Barrientos, M. A. G. von Keyserlingk, E. Galo, and D. M. Weary. 2013. Herd-level risk factors for lameness in freestall farms in the northeastern United States and California. J. Dairy Sci. 96:318–328.

Cohen, J. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychol. Bull. 70:213–220.

Espejo, L., M. Endres, and J. Salfer. 2006. Prevalence of lameness in high-producing Holstein cows housed in freestall barns in Minnesota. J. Dairy Sci. 89:3052–3058.

Gunns, R., L. Johnston, and S. Hudson. 2002. Victim selection and kinematics: A point-light investigation of vulnerability to attack. J. Nonverbal Behav. 26:129–158.

Isensee, A., F. Leiber, A. Bieber, A. Spengler, S. Ivemeyer, V. Maurer, and P. Klocke. 2014. Comparison of a classical with a highly formularized body condition scoring system for dairy cattle. Animal 8:1971–1977.

Lenth, R. V. 2014. Lsmeans: Least Squares Means. R package version 2.12. http://www.r-project.org/.

Manske, T. 2002. Hoof Lesions and Lameness in Swedish Dairy Cattle: Prevalence, Risk Factors, Effects of Claw Trimming and Consequences for Productivity. T. Manske, Uppsala, Sweden.

Manson, F. J., and J. D. Leaver. 1988. The influence of concentrate amount on locomotion and clinical lameness in dairy cattle. Anim. Sci. 47:185–190.

Meyer, D., A. Zeileis and K. Hornik. 2014. Vcd: Visualizing categorical data. R package version 1.3–2. http://www.r-project.org/.

Nielsen, B. H., A. Angelucci, A. Scalvenzi, B. Forkman, F. Fusi, F. Tuyttens, H. Houe, H. Blokhuis, J. T. Sørensen, J. Rothmann, L. Matthews, L. Mounier, L. Bertocchi, M. Richard, M. Donati, P. P. Nielsen, R. Salini, S. de Graaf, S. Hild, S. Messori, S. S. Nielsen, V. Lorenzi, X. Boivin and P. T. Thomsen. 2014. Use of Animal Based Measures for the Assessment of Dairy Cow Welfare ANIBAM. EFSA-Q-2012–00724. Eur. Food Safe. Authority, Parma, Italy.

O'Callaghan, K. A., P. J. Cripps, D. Y. Downham, and R. D. Murray. 2003. Subjective and objective assessment of pain and discomfort due to lameness in dairy cattle. Anim. Welf. 12:605–610.

R Core Team. 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 3.0.2.

Reader, J. D., M. J. Green, J. Kaler, S. A. Mason, and L. E. Green. 2011. Effect of mobility score on milk yield and activity in dairy cattle. J. Dairy Sci. 94:5045–5052.

Scherer, R. 2014. PropCIs: Various Confidence Interval Methods for Proportions. R package version 0.2–5. http://www.r-project.org/.

Schlageter-Tello, A., E. A. M. Bokkers, P. W. G. Groot Koerkamp, T. Van Hertem, S. Viazzi, C. E. B. Romanini, I. Halachmi, C. Bahr, D. Berckmans, and K. Lokhorst. 2014a. Effect of merging levels of locomotion scores for dairy cows on intra- and interrater reliability and agreement. J. Dairy Sci. 97:5533–5542.

Schlageter-Tello, A., E. A. M. Bokkers, P. W. G. G. Koerkamp, T. Van Hertem, S. Viazzi, C. E. B. Romanini, I. Halachmi, C. Bahr, D. Berckmans, and K. Lokhorst. 2014b. Manual and automatic locomotion scoring systems in dairy cows: A review. Prev. Vet. Med. 116:12–25.

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

Tadich, N., C. Tejeda, S. Bastias, C. Rosenfeld, and L. E. Green. 2013. Nociceptive threshold, blood constituents and physiological values in 213 cows with locomotion scores ranging from normal to severely lame. Vet. J. 197:401–405.

Thomsen, P. T., L. Munksgaard, and F. A. Tøgersen. 2008. Evaluation of a lameness scoring system for dairy cows. J. Dairy Sci. 91:119–126.

Tuyttens, F. A. M., M. Sprenger, A. Van Nuffel, W. Maertens, and S. Van Dongen. 2009. Reliability of categorical versus continuous scoring of welfare indicators: Lameness in cows as a case study. Anim. Welf. 18:399–405.

Whay, H., D. Main, L. Green, and A. Webster. 2003. Assessment of the welfare of dairy cattle using animal-based measurements: Direct observations and investigation of farm records. Vet. Rec. 153:197–202.

Whay, H. R., D. C. J. Main, L. E. Green, and A. J. F. Webster. 2002. Farmer perception of lameness prevalence. Page 355–358 in 12th Int. Symp. Lameness Rumin., Orlando, FL.

Zinpro Corp, 2009. Locomotion scoring of dairy cattle, FirstStep. DVD. Zinpro, Eden Prairie, MN.