Relation between observed locomotion traits and locomotion score in dairy cows

Andrés Schlageter-Tello,^{*1} Eddie A. M. Bokkers,[†] Peter W. G. Groot Koerkamp,[‡] Tom Van Hertem,[§] Stefano Viazzi,[§] Carlos E. B. Romanini,[§] Ilan Halachmi,[#] Claudia Bahr,[§] Daniël Berckmans,[§] and Kees Lokhorst^{*}

^{*}Wageningen UR Livestock Research, and

[†]Animal Production Systems Group, Wageningen University, PO Box 338, 6700 AH, Wageningen, the Netherlands

[‡]Farm Technology Group, Wageningen University, PO Box 36, 6700 AA, Wageningen, the Netherlands

[§]Division Measure, Model and Manage Bioresponses, Katholieke Universiteit Leuven, PO Box 2456, 3001 Heverlee, Belgium

[#]Institute of Agricultural Engineering—Agricultural Research Organization, PO Box 6, Bet-Dagan 50250, Israel

ABSTRACT

Lameness is still an important problem in modern dairy farming. Human observation of locomotion, by looking at different traits in one go, is used in practice to assess locomotion. The objectives of this article were to determine which individual locomotion traits are most related to locomotion scores in dairy cows, and whether experienced raters are capable of scoring these individual traits consistently. Locomotion and 5 individual locomotion traits (arched back, asymmetric gait, head bobbing, reluctance to bear weight, and tracking up) were scored independently on a 5-level scale for 58 videos of different cows. Videos were shown to 10 experienced raters in 2 different scoring sessions. Relations between locomotion score and traits were estimated by 3 logistic regression models aiming to calculate the size of the fixed effects on the probability of scoring a cow in 1 of the 5 levels of the scale (model 1) and the probability of classifying a cow as lame (locomotion score ≥ 3 ; model 2) or as severely lame (locomotion score ≥ 4 ; model 3). Fixed effects were rater, session, traits, and interactions among fixed effects. Odds ratios were calculated to estimate the relative probability to classify a cow as lame when an altered (trait score ≥ 3) or severely altered trait (trait score ≥ 4) was present. Overall intrarater and interrater reliability and agreement were calculated as weighted kappa coefficient (κ_w) and percentage of agreement, respectively. Specific intrarater and interrater agreement for individual levels within a 5-level scale were calculated. All traits were significantly related to the locomotion score when scored with a 5-level scale and when classified as (severely) lame or nonlame. Odds ratios for altered and severely altered traits were 10.8 and 14.5 for reluctance

to bear weight, 6.5 and 7.2 for asymmetric gait, and 4.8 and 3.2 for arched back, respectively. Raters showed substantial variation in reliability and agreement values when scoring traits. The acceptance threshold for overall intrarater reliability ($\kappa_w \geq 0.60$) was exceeded by locomotion scoring and all traits. Overall interrater reliability values ranged from $\kappa_w = 0.53$ for tracking up to $\kappa_w = 0.61$ for reluctance to bear weight. Intrarater and interrater agreement were below the acceptance threshold (percentage of agreement $< 75\%$). Most traits tended to have lower specific intrarater and interrater agreement in level 3 and 5 of the scale. In conclusion, raters had difficulties in scoring locomotion traits consistently, especially slight alterations were difficult to detect by experienced raters. Yet, the locomotion traits reluctance to bear weight, asymmetric gait, and arched back had the strongest relation with the locomotion score. These traits should have priority in locomotion-scoring-system guidelines and are the best to be used for the development of automated locomotion scoring systems.

Key words: agreement, lameness, locomotion, observer, reliability

INTRODUCTION

Locomotion scoring methods are procedures used to indicate the quality of the locomotion of cows and often used to classify them as lame or nonlame. Locomotion scoring methods are therefore used to create comparable records for lameness control and management (Whay, 2002; Flower and Weary, 2009). When assessing locomotion, raters focus their attention onto traits that are generally described in the protocol of the method and that are related to the quality of locomotion. Using these traits, raters assign an overall score to the locomotion of cows. In total 12 traits have been described in locomotion scoring methods (Schlageter-Tello et al., 2014b). Depending on which locomotion scoring meth-

Received November 3, 2014.

Accepted July 12, 2015.

¹Corresponding author: andres_schlageter@yahoo.com

od is being used, raters have to evaluate up to 7 traits. Most locomotion scoring methods use between 3 and 5 traits (Schlageter-Tello et al., 2014b). Some of the most used traits in locomotion scores are asymmetric gait, reluctance to bear weight, arched back, head bobbing, and tracking up (Schlageter-Tello et al., 2014b).

Cows showing impaired locomotion, however, do not always express all traits described in the locomotion scoring method. Bach et al. (2007) and Thomsen et al. (2008), for example, reported that not all cows with impaired locomotion showed an arched back, and Chapinal et al. (2009) reported that few cows displayed head bobbing. The fact that cows express impaired locomotion in different ways implies that human raters have to weigh different traits and decide which of them is more important to base a locomotion score on. From a practical point of view, knowing the relation of different traits with the locomotion score would allow the giving of priority to these traits when raters have to assess locomotion.

Automatic locomotion scoring systems are an attempt to mimic locomotion scoring performed by human raters by measuring traits using different types of sensors (Van Nuffel et al., 2009; Schlageter-Tello et al., 2014b; Van Hertem et al., 2014). Automatic locomotion scoring systems, for example, measure forces exerted on the floor surface by the hoofs while cows walk (Scott, 1988; Rajkondawar et al., 2002; Rajkondawar et al., 2006) or measure the weight distribution of individual limbs (Neveux et al., 2006; Rushen et al., 2007; Pastell et al., 2010). Both types of automatic locomotion scoring systems measure variables that can be compared with reluctance to bear weight. A different approach measures time and distance of variables associated with limb movement such as tracking up (Song et al., 2008; Pluk et al., 2010); gait variables that can be associated with asymmetric gait, such as asymmetry of step length, asymmetry of step time, asymmetry of step width, stance time, and stride length (Maertens et al., 2011), and some specific posture characteristics such as back curvature (Viazzi et al., 2013; Van Hertem et al., 2014). Most of the current automatic locomotion scoring systems focus on the measurement and analysis of only one trait (Schlageter-Tello et al., 2014b). Because automatic locomotion scoring systems are based on the measurement of a single trait, it is important to know which trait is most related to locomotion scoring and would be helpful to create better automatic locomotion scoring systems for locomotion scoring and lameness detection.

Finally, it has been proposed that trait scoring could replace traditional locomotion scoring (based on the assessment a combination of traits), because it could be

easier and faster than scoring locomotion for a simple on-farm utilization (Chapinal et al., 2009; Thomsen, 2009). For using trait scoring instead of locomotion scoring, it is not only required to know which traits are most related to locomotion, but it is also important to know whether human raters can identify and score these traits consistently. Consistency is expressed as reliability and agreement within and between raters (Kottner et al., 2011). Reliability indicates the capability of raters to differentiate among levels within the scoring system, whereas agreement indicates the capability of raters to assign identical scores to the same cow (Kottner et al., 2011). Reliability and agreement can be calculated by comparing data of scores assigned to a cow by the same rater under similar conditions at different times (intrarater reliability and agreement) or by comparing scores from 2 or more raters assigned to the same cow under similar conditions (interrater reliability and agreement) (Martin and Bateson, 2007).

Taking the previously stated into consideration, the objectives of this study were to determine which individual traits are most related to locomotion scores in dairy cows, and whether experienced raters are capable of scoring these individual traits consistently.

MATERIALS AND METHODS

Experimental Design

Locomotion and 5 traits were scored using videos of cows walking through an alley. Video recording was performed at a dairy farm with 1,100 milking cows located in Israel, previously described by Van Hertem et al., (2013). Cows walking through an alley (1.5 m wide, 7 m long) on a concrete floor were recorded with a Nikon D7000 camera (Nikon Corporation, Tokyo, Japan). To obtain flank views of cows, the camera was positioned 4 m perpendicular to the progression line of the alley and 1.35 m above ground level.

Video records of individual cows in the herd were stored in a video data set. Each video record was scored for locomotion according to a 5-level scale (described later) by one experienced rater who did not participate in the experiment. From the video data set 58 video records from 58 different cows were selected. A video record was included in the experiment only if the cow made at least 4 steps and if there was enough contrast between the cow and the background. If a video record did not meet the quality criteria, a new video record was selected randomly from the video data set until the predetermined number of 12 video records per level was reached. For level 5, only 8 video records were available that met the criteria. Two extra video records

were taken for level 3 because this level appeared to be the most difficult to assess consistently in previous studies (Schlageter-Tello et al., 2014a). The number of video records selected for the present experiment was determined using other articles reporting reliability and agreement for locomotion scoring in dairy cows as reference (Flower and Weary, 2006; Thomsen et al., 2008; Channon et al., 2009) and taking into account that a large number of video records would exhaust raters, which would negatively affect the outcomes.

Locomotion scoring was performed using a 5-level ordinal scale. Locomotion scoring was based on the judgment of 5 traits as described by Flower and Weary (2006). Cows with a locomotion score of level 1 had a smooth and fluid movement, and cows with a locomotion score of level 5 could hardly move. The 5 traits used to evaluate locomotion were (1) asymmetric gait defined as differences of distance or time in the imprints between 2 consecutive strides; (2) arched back, defined as the convex back line formed by the spine between the withers and tailbone; (3) reluctance to bear weight, defined as the inability of cows to bear weight in the affected limbs; (4) tracking up, defined as the distance between the position of the front foot and hind foot on the same body side on the floor in the subsequent step; and (5) head bobbing, defined as exaggerated movement of the head when an affected limb is lifted from the ground. The 5 traits were scored separately from locomotion using a similar 5-level ordinal scale with level 1 indicating a not-altered trait and level 5 indicating an extremely altered trait. Further description of locomotion and traits scoring during the experiment can be found in Table 1.

Locomotion and traits scoring was performed by 10 experienced raters with different backgrounds and without further training together. They had originally been trained using different locomotion scores. A detailed description of the raters can be found in Table 2. Raters were not informed about the objectives of the study, the number of different videos used, and the randomizations performed during the experiment. The 58 video records were shown to the 10 raters in 2 scoring sessions in a random order. Each scoring session was split in 6 parts in which raters scored the 58 video records each time again for either locomotion or 1 of the 5 traits separately. At the start of each part in both scoring sessions, the raters received 15 min of instruction on scoring locomotion and traits. The instruction consisted of a description of the locomotion or trait to be scored while showing 2 videos per level of the 5-level scale. Videos used for instruction were not included in the experiment. The instruction was the only moment in which raters were allowed to discuss scoring.

Table 1. Description of locomotion and traits score (based on Flower and Weary, 2006)

Level	Locomotion	Arched back	Asymmetric gait	Head bobbing	Reluctance to bear weight	Tracking up
1	Smooth and fluid movement	Flat back	Long and confident stride or step	Steady head carriage	All legs bear weight equally	Hind hooves land on or in front of fore hooves
2	Imperfect locomotion but ability to move freely not diminished	Mildly arched back	Slightly asymmetric gait	Slight head bobbing	Slight limp can be discerned	Hind hooves and front hooves do not track up perfectly (approximately one hoof distance between track of front and hind hooves)
3	Capable locomotion but ability to move freely is compromised	Arched back	Short strides or step	Head bobbing (associated with release of the affected limb from the floor)	Limp can be discerned	Hind hooves do not track up (approximately 2 hooves distance between track of front and hind hooves)
4	Ability to move freely is obvious diminished	Obvious arched back	Short and hesitant strides or step	Obvious head bobbing (associated with release of the affected limb from the floor)	Clearly reluctant to bear weight on at least one limb but uses that limb in locomotion	Hind hooves do not track up (approximately 3 hooves distance between track of front and hind hooves)
5	Ability to move is severely restricted and must be vigorously encouraged to move	Exaggerated arched back	Very short, hesitant, and deliberate strides or step	Exaggerated head bobbing (associated with release of the affected limb from the floor)	Inability to bear weight on the limb or more than one limb clearly affected	Poor tracking up with short strides (approximately 4 or more hooves distance between track of front and hind hooves)

A detailed explanation of the experimental design is described by Schlageter-Tello et al. (2014a).

Statistical Analysis

The distribution of locomotion and trait scores over the 5 levels was calculated considering the scores of 10 raters of both sessions and was expressed as percentage of the 58 scored cows.

To establish the relation of traits with locomotion, 3 generalized linear mixed models on a logistic scale were developed. The first model calculated probability of scoring a cow in each of the 5 levels of the locomotion score. This model comprised the fixed effects of session, traits scored on a 5-level scale (arched back, asymmetric gait, head bobbing, reluctance to bear weight, and tracking up), and interactions between rater × session, rater × traits, and traits × session. Effects that were not significant were deleted from the model. The final model included the effects of traits (arched back, asymmetric gait, head bobbing, reluctance to bear weight, and tracking up) and the interactions between rater × tracking up and rater × session.

The second and third model calculated the probability of classifying a cow as lame and severely lame, respectively. These 2 models included the same fixed effects as the first model. In the second and third model, traits were transformed into a binary scale. The threshold to classify a cow with an altered trait was a score ≥3 (for the second model), whereas for an extremely altered trait, a score ≥4 was used (for the third model). The final model included the effects of traits (arched back, asymmetric gait, head bobbing, reluctance to bear weight, and tracking up) and the interaction between rater × tracking up. For the second and third model, odds ratios and their respective confidence intervals were calculated to estimate the relative probability of classifying a cow as lame or severely

lame when cows showed an altered or extremely altered trait when compared with a nonaltered trait. Rater and cow were included as random effect in all models. The relation between locomotion and traits was established by estimating the level of significance ($P < 0.05$) and the size of the relative size of the fixed effect (F -test). Logistic regression was performed using the Glimmix procedure in SAS 9.2 (SAS Institute Inc., Cary, NC).

Intrarater and interrater reliability and agreement were calculated for the 5-level scales for locomotion and traits. Intrarater reliability and agreement were calculated comparing the scores from the same cow in 2 different sessions for 10 raters. Interrater reliability and agreement were calculated comparing the scores of the same cow assigned for pairwise comparisons for 2 different raters. Interrater reliability and agreement were calculated for the 2 scoring sessions.

Overall intrarater reliability and agreement were calculated by creating a cross table that included all comparisons for the same rater. Overall interrater reliability and agreement were calculated with cross tables including all pairwise comparisons for raters and sessions.

Intrarater and interrater reliability was calculated as weighted kappa (κ_w ; Cohen, 1968) using linear weighing (Cicchetti and Allison, 1971). The κ_w corrects the percentage of agreement by the expected agreement by chance. Additionally, κ_w introduces different weighing according to the magnitude of disagreement of the raters. Intrarater and interrater agreement was expressed as percentage of agreement (PA). The acceptance threshold for reliability values was stated at $\kappa_w \geq 0.6$ (Landis and Koch, 1977). The commonly accepted threshold for agreement is PA ≥75% (Burn and Weir, 2011).

The percentage of specific intrarater and interrater agreement was calculated for locomotion and trait scores. Percentage of specific agreement is based on the

Table 2. Background and experience of the 10 raters participating in the study

Rater	Background	Trained in locomotion score described by	Last scoring ¹
1	Researcher	Sprecher et al., 1997	Less than 1 yr ago
2	Researcher	Manson and Leaver, 1988	2 yr ago
3	Researcher	Manson and Leaver, 1988	2 yr ago
4	Farmer and researcher	Manson and Leaver, 1988	2 yr ago
5	Veterinarian	Sprecher et al., 1997	Currently
6	Technician	Winckler and Willen, 2001	4 yr ago
7	Veterinarian	Welfare Quality, 2009	Currently
8	Veterinarian	Welfare Quality, 2009	Currently
9	Researcher	Winckler and Willen, 2001	4 yr ago
10	Researcher	Sprecher et al., 1997	Currently

¹Indicates how long ago the raters regularly performed locomotion scoring in relation to the start of the experiment. “Currently” indicates that raters performed locomotion scoring regularly before the experiment.

Table 3. Percentage of cows ($n = 58$) scored in each level of a 5-level scale for locomotion (LS) and 5 traits: arched back (AB), asymmetric gait (AG), head bobbing (HB), reluctance to bear weight (RB), and tracking up (TU)¹

Item	Level 1	Level 2	Level 3	Level 4	Level 5
LS	18.7 (5.5–30.2)	31.7 (25.0–40.1)	24.9 (13.8–33.9)	20.0 (10.3–28.4)	4.8 (0.0–14.7)
AB	15.6 (5.2–24.1)	34.9 (25.5–43.1)	28.1 (20.7–39.7)	15.9 (10.3–19.8)	5.4 (0.9–8.7)
AG	25.5 (15.0–42.2)	32.4 (23.3–44.8)	22.4 (15.5–30.1)	14.9 (10.3–20.9)	4.7 (0–10.3)
HB	26.8 (12.1–53.5)	38.1 (17.5–47.4)	19.7 (11.4–32.8)	12.4 (8.6–18.3)	3.0 (0.9–7.0)
RB	25.7 (14.7–40.5)	27.5 (18.6–34.8)	22.1 (15.5–31.0)	20.2 (16.5–23.0)	4.4 (0–12.1)
TU	11.0 (0.0–21.6)	19.1 (6.0–29.3)	27.4 (18.3–35.3)	25.7 (14.7–43.1)	16.8 (10.3–23.3)

¹Values were based on the scoring of 10 raters in 2 sessions. Ranges of individual raters are given in parentheses.

concept of positive and negative agreement (Cicchetti and Feinstein, 1990). The specific agreement indicates the agreements of raters on average in each specific level of the 5-level scale in 2 sessions. In this regard, the PA can be considered a weighed sum of the specific agreements of each level (Cicchetti and Feinstein, 1990; Warrens, 2013). The confidence intervals for the specific agreement were calculated with the delta method as proposed by Graham and Bull (1998). Because no acceptance threshold for specific intra- and interrater agreement is known, it was set at $\geq 75\%$ as was done for PA. Reliability, agreement, and specific agreement were calculated using the Frequency procedure in SAS 9.2 (SAS Institute Inc., Cary, NC).

RESULTS

Distribution of Scores

Relative distributions of scores assigned to locomotion and traits on a 5-level scale are presented in Table 3. Distribution for each of the 5 levels showed high variation for different raters (Table 3). On average, locomotion and most traits (except tracking up) had a similar number of cows in levels 1, 2, 3, and 4 of the scale and few cows in level 5. On average, tracking up

had fewer cows in level 1 of the scale when compared with other levels in the scale (Table 3).

Relation Between Traits and Locomotion

The results of the 3 models for estimating (a) the probability of scoring a cow in 1 of the 5 levels of the scale, (b) the probability of classifying a cow as lame, or (c) the probability of classifying a cow as severely lame show that all 5 traits had a significant effect for predicting the dependent variable in the 3 models (Table 4). Significant interactions rater $\times$ tracking up and rater $\times$ session were found when the probability of scoring a cow along the 5-level locomotion score was used as dependent variable (Table 4). When the probability of classifying a cow as lame or severely lame was used as dependent variable, only the interaction rater $\times$ tracking up was significant.

The trait reluctance to bear weight showed a higher odd ratio for the probability of classifying a cow as lame (10.8) or as severely lame (14.5), as compared with the other traits (Table 4). The odds ratio for asymmetric gait, head bobbing, and reluctance to bear weight increased when the threshold score was changed from ≥ 3 to ≥ 4 for locomotion and traits, whereas the odds ratio for arched back and tracking up decreased (Table 4).

Table 4. Size of the fixed effect (F -value) and odds ratio for 5 traits [arched back (AB), asymmetric gait (AG), head bobbing (HB), reluctance to bear weight (RB), and tracking up (TU)] on the probability of classifying a cow as lame¹

Effect	5 Levels		Lame or nonlame			Severely lame or nonlame		
	F -value	P -value	F -value	P -value	Odd ratio ² (95% CI)	F -value	P -value	Odd ratio ³ (95% CI)
AB	23.5	<0.001	43.3	<0.001	4.8 (3.0–7.7)	14.8	<0.001	3.2 (1.7–5.8)
AG	15.7	<0.001	49.8	<0.001	6.5 (3.9–10.9)	28.6	<0.001	7.2 (3.5–15.1)
HB	15.6	<0.001	26.6	<0.001	4.8 (2.6–8.7)	25.8	<0.001	7.3 (3.9–15.9)
RB	30.4	<0.001	77.8	<0.001	10.8 (6.3–18.2)	63.4	<0.001	14.5 (7.5–28.2)
TU	17.1	<0.001	3.9	0.05	8.5 (1.0–71.9)	15.2	<0.001	5.8 (2.4–14.2)
Rater $\times$ TU	1.9	0.03	2.4	0.01	—	2.4	0.02	—
Rater $\times$ session	2.6	0.04	—	—	—	—	—	—

¹Level of significance was established at $P < 0.05$.

²Indicates the relative probability of classifying a cow as lame when a trait is indicated to be altered (lame and altered trait indicated when score was ≥ 3 on a 5-level scale).

³Indicates the relative probability of classifying a cow as severely lame when a trait is indicated to be severely altered (severely lame and severely altered trait indicated when trait score was ≥ 4 on a 5-level scale).

Overall Reliability and Agreement

Overall intra- and interrater reliability and agreement for locomotion and the 5 traits are shown in Tables 5 and 6. Overall intrarater reliability for locomotion scoring, tracking up, and head bobbing exceeded the acceptance threshold for κ w (Table 5). The acceptance threshold for intrarater reliability of individual raters was exceeded by all raters only for locomotion and head-bobbing scoring. Out of the 10 raters, 9, 6, 7, and 6 raters exceeded the acceptance threshold for κ w scoring arched back, asymmetric gait, reluctance to bear weight, and tracking up, respectively. On the other hand, raters did not exceed the acceptance threshold for overall intrarater agreement, for locomotion score, or for any of the individual traits (Table 5).

The acceptance threshold for overall interrater reliability was exceeded only for locomotion and the traits head bobbing and reluctance to bear weight (Table 6). For locomotion, 75% of 90 pairwise comparisons among raters were above the acceptance threshold for reliability. The acceptance threshold for overall interrater agreement was not exceeded by any of the 5 traits, with PA values ranging from 45.7% (tracking up) to 57% (locomotion) (Table 6).

Specific Agreement

Specific intrarater agreement and specific interrater agreement for individual levels of the 5-level scale for locomotion and the 5 traits are shown in Tables 5 and 6. The acceptance threshold for specific intrarater agreement was exceeded by locomotion in levels 1, 4, and 5 only (Table 5); all others values were lower than 75% (Tables 5 and 6).

Specific intrarater and interrater agreement for specific levels of the scale resulted in different patterns for locomotion and the 5 traits assessed during the experiment. The pattern for specific intrarater agreement for locomotion, asymmetric gait, reluctance to bear weight, and tracking up showed the highest values in levels 1, 4, or 5, whereas the lowest value was in level 3. The pattern for arched back showed higher values for specific intrarater agreement in levels 1 and 2 than in levels 3, 4, and 5. Head bobbing showed similar values of specific intrarater agreement in the 5 levels of the scale (Table 5). Specific interrater agreement for locomotion, arched back, asymmetric gait, head bobbing, and reluctance to bear weight had highest values in level 1, whereas lowest values were in levels 3 or 5. Tracking up had the highest specific interrater agreement in level 5 and the lowest in value in level 3 of the scale (Table 6).

Table 5. Overall intrarater reliability [expressed as weighted kappa (κ w)] and agreement [expressed as percentage of agreement (PA) and percentage of specific agreement] for locomotion score (LS), arched back (AB), asymmetric gait (AG), head bobbing (HB), and tracking up (TU) scored with a 5-level scale by 10 raters in 2 sessions

Trait	Reliability					Agreement				
	κ w	(95% CI)	PA, %	(95% CI)	Level 1, %	(95% CI)	Level 2, %	(95% CI)	Level 3, %	(95% CI)
LS	0.77	(0.74–0.80)	71.4	(67.5–75.1)	76.4	(70.1–82.7)	68.5	(63.0–74.0)	65.0	(58.6–71.5)
AB	0.71	(0.67–0.74)	66.2	(62.2–70.1)	73.7	(66.5–80.9)	70.8	(65.7–75.8)	61.7	(55.5–68.0)
AG	0.67	(0.63–0.71)	61.8	(57.7–65.8)	66.9	(60.7–73.1)	59.4	(53.5–65.3)	55.8	(48.5–63.1)
HB	0.72	(0.69–0.76)	68.3	(64.4–72.1)	69.7	(63.8–75.6)	68.5	(63.5–73.5)	65.8	(58.6–72.9)
RB	0.69	(0.65–0.72)	60.8	(56.6–64.8)	68.9	(62.8–74.9)	55.1	(48.4–61.7)	52.3	(44.9–59.7)
TU	0.65	(0.62–0.70)	58.8	(54.6–62.8)	63.0	(53.1–72.8)	50.0	(41.7–57.9)	55.4	(48.8–62.0)

Table 6. Overall interrater reliability [expressed as weighted kappa (κ_w)] and agreement [expressed as percentage of agreement (PA) and percentage of specific agreement] for locomotion score (LS), arched back (AB), asymmetric gait (AG), head bobbing (HB), reluctance to bear weight (RB), and tracking up (TU) scored with a 5-level scale by 10 raters in 2 sessions

Trait	Reliability		Agreement									
	κ_w	(95% CI)	PA, %	(95% CI)	Level 1, %	(95% CI)	Level 2, %	(95% CI)	Level 3, %	(95% CI)	Level 4, %	(95% CI)
LS	0.65	(0.64–0.66)	57.1	(55.7–58.4)	64.7	(62.2–67.2)	57.5	(55.4–59.5)	50.8	(48.5–53.2)	60.0	(57.3–62.5)
AB	0.59	(0.58–0.60)	48.7	(47.0–50.4)	32.7	(29.5–35.9)	58.1	(56.2–60.0)	49.2	(47.0–51.5)	50.9	(47.9–53.9)
AG	0.58	(0.57–0.59)	51.5	(49.8–53.1)	60.1	(57.9–62.3)	50.1	(48.0–52.1)	45.7	(43.2–48.2)	49.2	(46.2–52.3)
HB	0.61	(0.59–0.62)	56.6	(55.1–58.1)	57.0	(54.8–59.2)	57.3	(55.4–59.1)	51.0	(48.3–53.6)	61.9	(58.7–65.0)
RB	0.61	(0.60–0.62)	53.9	(52.4–55.6)	64.6	(62.5–66.7)	46.2	(43.9–48.5)	46.5	(44.0–49.1)	62.6	(60.1–65.0)
TU	0.53	(0.52–0.54)	45.7	(43.9–47.5)	47.1	(43.5–50.7)	37.7	(34.9–40.4)	39.6	(37.3–41.9)	46.9	(44.5–49.2)
											45.2	(39.6–50.7)
											39.3	(34.1–44.4)
											48.2	(42.7–53.6)
											58.6	(52.1–65.1)
											38.3	(32.7–44.0)
											62.5	(59.8–65.1)

DISCUSSION

In the current study, all 5 individual traits had a significant effect on the probability of scoring a cow in 1 of the 5 levels of the locomotion score. In a logistic regression, the *F*-test value indicates the relative size of the fixed effect on explaining the dependent variable (McCullagh and Nelder, 1989). *F*-test value suggests that the traits that are most related to locomotion score are, in order, reluctance to bear weight, asymmetric gait, arched back, head bobbing, and tracking up. The relationship between traits and locomotion in literature has been established by others using different approaches not directly comparable with the methodology used in the current experiment. Based on the assessment of 2 raters, Borderas et al. (2008) and Chapinal et al. (2009) reported that reluctance to bear weight (range $r = 0.88$ – 0.90) and asymmetric gait (range $r = 0.84$ – 0.91), both scored on a continuous scale, were highly correlated with locomotion score, scored on an ordinal 9-level scale. For head bobbing, tracking up, and joint flexibility, these correlation coefficients ranged from 0.70 to 0.80, and for arched back from 0.41 to 0.68 (Borderas et al., 2008; Chapinal et al., 2009). Abduction or adduction, defined as a tendency to rotate the limb outward or inward, had a low correlation with the locomotion score ($r = 0.32$; Chapinal et al., 2009). Van Nuffel et al. (2009) used a regression model to analyze relations between 10 traits and a 3-level locomotion score assessed by 39 raters with different levels of experience. Asymmetric gait, reluctance to bear weight, arched back, and abduction or adduction were significant for predicting the locomotion score assigned to a cow. Based on the current study and other studies, we can summarize that all 5 traits included in the current study are associated with locomotion score, but especially reluctance to bear weight and asymmetric gait. The strength of the relation between traits such as arched back, head bobbing, and tracking up and locomotion score varied between studies. Differences in results reported by different studies for the traits arched back, head bobbing, and tracking up may be explained by the different approaches used to estimate the relation between traits and locomotion. Although we gave the raters the task to score each trait and locomotion separately and independently, this could not be controlled in the current experiment. However, the different patterns for specific agreement values for different traits suggest that raters did so.

In the current experiment, raters had different probabilities for scoring the locomotion of a cow in 1 of the 5 levels of the scale in different sessions, which is indicated by the significant interaction rater $\times$ session. This interaction indicates that raters have a large

variation in the score assigned to a cow using a 5-level scale. The variation of raters when performing locomotion scoring on a 5-level scale is also reflected in the moderate intra- and interrater agreement and specific agreement.

When the probability of classifying a cow as lame or severely lame was estimated, the interaction between rater $\times$ session was not significant. This finding suggests that when the 5-level scale is transformed into a binary classification, raters have similar probabilities to classify a cow as (severely) lame. The fact that the interaction between rater $\times$ session is not significant is in concordance with the increase in agreement and specific agreement when the 5-level scale is merged into a 2-level scale (Schlageter-Tello et al., 2014a).

The significant effect for the interaction between rater $\times$ tracking up indicates that different raters give different importance to tracking up on the probability of scoring a cow in 1 of the 5 levels of the locomotion score. The significant interaction could be explained by the fact that tracking up was poorly identified by raters as reflected in the moderate reliability and agreement for this trait. Similarly, for the lame or nonlame classification, the significant interaction between rater $\times$ tracking up indicates that some raters classified a cow as lame and some as not lame when the trait tracking up was scored as altered.

Odds ratio were included in the study to provide an alternative and probably easier explanation to the relation between traits and locomotion than *F*-test values. Odds ratios indicate that when a trait is classified as (severely) altered, the probability of classifying a cow as (severely) lame increases. The trait that is mostly related to lameness is reluctance to bear weight. Raters had about 11 times higher chance of classifying a cow as lame when reluctance to bear weight was classified as altered than when it was not altered. When (severely) altered, arched back, asymmetric gait, head bobbing, and tracking up also increased the probabilities of classifying a cow as (severely) lame. However, the confidence intervals for tracking up were wide and contained the value 1, which indicates that a cow has the same probability to be classified as lame when a trait is classified as altered or not altered (Cook, 2002). The confidence intervals indicated no differences in odds ratio when using the classification lame and altered traits (threshold at score ≥ 3) and the model using the classification severely lame and severely altered traits (threshold at score ≥ 4).

Using a similar approach as in the current study for determining the relationship between traits (binary classification) and lameness (locomotion score ≥ 3 , using a 5-level scale), Hoffman et al. (2014) reported that arched back, cow-hocked (defined as "hocks were

rotated medially such that the point of the hock was directed medially and the lateral claw could be seen directed laterally when viewed from behind"), and favoring limb (defined as "if any part of a hind or fore hoof was elevated from the ground at the time of observation") were significantly associated with lameness. The odd ratio for lame or nonlame classification for having an arched back was 2.5, which is close to the values of 4.8 and 3.2 found in the current study. The odds ratio for favoring limbs and cow-hocked were 3 and 1.6, respectively. Differences between the study of Hoffman et al. (2014) and the current study, especially for the trait favoring limbs (comparable to reluctance to bear weight), could be explained by differences in the definition of the trait, method, and perspective of observation (live caudal view vs. video lateral view in the current experiment) or raters (3 raters observed different cows vs. 10 raters observing the same cows in the current article).

Acceptance thresholds for overall intrarater reliability were exceeded for locomotion and all traits, whereas for overall interrater reliability the acceptance thresholds were only exceeded by locomotion, head bobbing, and reluctance to bear weight. Exceeding the acceptance threshold for κ_w indicates that disagreements for intra- and interrater comparisons were mainly by one level difference and that disagreements with 2 or 3 levels were less frequent and that, in general, raters have an acceptable level for differentiating among levels within the scale. The specific agreement may be helpful for a better interpretation of the κ_w coefficient by showing the levels in which raters have more difficulties identifying. In this regard, the lowest values for specific intrarater and interrater agreement were often found in level 3 and 5 of the scale. Low values for specific intrarater and interrater agreement in level 3 indicate that a slight alteration of locomotion and traits are difficult to detect consistently by experienced raters. The low values for specific agreement in level 5 indicate that raters were not able to identify consistently cows with extremely impaired locomotion or traits. This finding, however, should be interpreted carefully because in the current study most raters scored few cows in level 5. Previously, it has been reported that when performing locomotion scoring, most disagreement occurs between level 1 and 2 on a 5-level scale. This is explained because usually about 80% of cows are classified in level 1 and 2 on a 5-level scale (Winckler and Willen, 2001). In the current experiment, a similar number of cows per level were included (Schlageter-Tello et al., 2014a). This allows a better evaluation of reliability and agreement of raters because it minimizes the effect of the population distribution on the expected agreement by chance (Hoehler, 2000; de Vet et al., 2006). Acceptance

thresholds for overall intrarater and interrater agreement were not exceeded by locomotion or by traits. This indicates that even for rather experienced raters as present in this study, obtaining acceptable agreement values is difficult when using a 5-level scale, which is in agreement with previous studies (Winckler and Wilken, 2001; Rutherford et al., 2009; Schlageter-Tello et al., 2014a). Only few and even not directly comparable results have been reported in the literature for reliability and agreement for traits. Flower and Weary (2006) and Borderas et al. (2008) reported acceptable intra- and interrater reliability values (expressed as Pearson correlation coefficient, $r \geq 0.7$) for tracking up, arched back, head bobbing, and reluctance to bear weight as scored in a 0 to 100 continuous scale. Flower and Weary (2006) reported $r = 0.48$ for asymmetric gait for interrater comparison. Both the literature and the current study suggest large variation for intra- and interrater reliability and agreement for the different traits.

One of the best ways to improve reliability and agreement of raters is training (Kaufman and Rosenthal, 2009). The objective of training is to establish similar opinions among raters about which score to assign to a cow, to improve reliability and agreement of raters at acceptable levels, and to maintain this over time (Martin and Bateson, 2007). Training alone, however, is not enough. Schlageter-Tello et al. (2015) showed that experienced and inexperienced raters did not increase reliability and agreement after a short training and scoring of about 600 cows. Engel et al. (2003) and Thomsen et al. (2008) also reported different responses of raters to training, with some raters improving while others did not or got even worse reliability and agreement values (Engel et al., 2003; Thomsen et al., 2008). March et al. (2007) reported that during the training period (about 140 cows scored), raters obtained moderate reliability ($\kappa_w = 0.52$) and agreement (PA = 52%) values. Acceptable levels of reliability ($\kappa_w = 0.69$) and agreement (PA = 73%) were only achieved after assessment of 650 cows (March et al., 2007). This result indicates that besides training, experience of raters is a fundamental aspect in attaining an acceptable reliability and agreement. Even after training and with experience, maintenance of acceptable levels of reliability and agreement over time is difficult because of drift. Drift refers to the tendency of raters to change over time how they apply the definitions of a measurement (Kazdin, 1977), and therefore, regular update training sessions are required to avoid this drift. The requirement for regular training affects the feasibility of an effective training program, because it might demand time and effort from the raters to accomplish. Generally, under practical conditions, raters scoring locomotion are ex-

perienced but do not get any refresher training. Studies that use experienced raters without further training may provide reliability and agreement values similar to those obtained by farmers, veterinarians, or technicians applying locomotion scoring during their daily work.

As stated in the introduction, to know the relation between traits and locomotion would be useful to establish guidelines about which traits are priority to assess by raters. Based on the relation between locomotion and traits, we recommend that raters should give more importance to the traits reluctance to bear weight, asymmetric gait, and arched back when scoring locomotion. Head bobbing can be included in locomotion scoring using multiple traits because it was significantly related to locomotion scoring and raters were able to score head bobbing consistently. However, it has been reported that head bobbing is not frequently shown in cows classified as lame (Chapinal et al., 2009). Although the interaction between rater $\times$ tracking up was significant, it is important to remark that raters were not trained to detect this trait. Thus, judging tracking up while scoring locomotion still may be useful to detect some cases that otherwise would be missed. However, it should not be a priority trait to detect when assessing locomotion.

Usually, raters look at a combination of traits when assessing locomotion, without specifically scoring each one independently. It has been proposed, however, that scoring a single trait instead of locomotion would be easier (Chapinal et al., 2009; Thomsen, 2009). For this to be true, the reliability and agreement must be higher for scoring the traits separately than for scoring locomotion all at once. On average none of the traits had a better reliability and agreement than locomotion. This finding suggests that locomotion scoring based on a single trait, for example, arched back (Thomsen, 2009), a priori is not simpler than locomotion scoring based on a combination of traits. An explanation might be that raters were experienced in locomotion scoring and not in scoring single traits. Training on single traits, therefore, might help in improving the reliability and agreement of scoring.

Most automatic locomotion scoring systems mimic locomotion scoring performed by humans. Because of technical limitations, however, many automatic locomotion scoring systems assess locomotion based on the measurement of only one trait (Schlageter-Tello et al., 2014b). Hence, it has been stated that knowing the trait that is most related to locomotion scoring would be helpful to create better automatic locomotion scoring systems for lameness detection (Van Nuffel et al., 2009; Schlageter-Tello et al., 2014b). According to the results of the current study, automatic locomotion scor-

ing systems measuring forces exerted by hoofs on the floor (Rajkondawar et al., 2006) or weight distribution of limbs (Neveux et al., 2006; Pastell et al., 2010), which are related to reluctance to bear weight, should perform better than automatic locomotion scoring systems measuring gait asymmetries (Maertens et al., 2011), arched back (Viazzi et al., 2013; Van Hertem et al., 2014), or tracking up (Song et al., 2008; Pluk et al., 2010). In this regard, Liu et al. (2011) reported a sensitivity of 52% and a specificity of 88% for detecting lame or nonlame cows with automatic locomotion scoring system based on measuring forces exerted by hoofs on force plates (comparable with reluctance to bear weight). Whereas in an automatic locomotion scoring system based on measuring back curvature with computer vision techniques, Viazzi et al. (2013) reported a sensitivity of 76% and a specificity of 91% for detecting lame or nonlame cows. Better performance for detecting lame and nonlame cows (sensitivity and specificity >85%) were reported by de Mol et al. (2013) with an automatic locomotion scoring system measuring and analyzing different behaviors in cows without measuring any locomotion trait. According to the sensitivities and specificities values for the classification of lame and nonlame, the performance of automatic locomotion scoring systems had an important variability and seems to be independent of the measured trait for classifying cows as lame. However, this statement is only based on performance of 3 automatic locomotion scoring systems that were validated with limited conditions (Schlageter-Tello et al., 2014b). Further research is required to determine whether automatic locomotion scoring systems based on the measurement of variables than can be comparable to reluctance to bear weight perform better than systems measuring variables compared with the other traits used in the current experiment.

CONCLUSIONS

When scored by experienced raters without further training together, the 5 locomotion traits (arched back, asymmetric gait, head bobbing, reluctance to bear weight, and tracking up) were significantly related to the locomotion score of cows, of which reluctance to bear weight, asymmetric gait, and arched back had the strongest relation with the locomotion score. Raters performing locomotion scoring and developers of automatic locomotion scoring systems should give priority to reluctance to bear weight, asymmetric gait, and arched back. The high variation for reliability and agreement values showed that experienced raters had a limited capability for scoring locomotion traits consistently. Slight alterations in locomotion traits were hard to detect consistently by the experienced raters. None

of the traits that were scored separately had better reliability and agreement values than locomotion scoring.

ACKNOWLEDGMENTS

This study is part of the Marie Curie Initial Training Network BioBusiness project (FP7-PEOPLE-ITN-2008). We thank the 10 raters for their time and effort to observe and assess the videos for this experiment. Thanks are extended to Bas Engel (Wageningen UR, Wageningen, the Netherlands) for his valuable advice on the statistical analysis. This study was funded by the Marie Curie Initial Training Network BioBusiness Project (KU Leuven, Leuven, Belgium, FP7-PEOPLE-ITN-2008).

REFERENCES

- Bach, A., M. Dinares, M. Devant, and X. Carre. 2007. Associations between lameness and production, feeding and milking attendance of Holstein cows milked with an automatic milking system. *J. Dairy Res.* 74:40–46.
- Borderas, T. F., A. Fournier, J. Rushen, and A. M. B. De Passille. 2008. Effect of lameness on dairy cows' visits to automatic milking systems. *Can. J. Anim. Sci.* 88:1–8.
- Burn, C. C., and A. A. S. Weir. 2011. Using prevalence indices to aid interpretation and comparison of agreement ratings between two or more observers. *Vet. J.* 188:166–170.
- Channon, A. J., A. M. Walker, T. Pfau, I. M. Sheldon, and A. M. Wilson. 2009. Variability of Manson and Leaver locomotion scores assigned to dairy cows by different observers. *Vet. Rec.* 164:388–392.
- Chapinal, N., A. M. de Passille, D. M. Weary, M. A. G. von Keyserlingk, and J. Rushen. 2009. Using gait score, walking speed, and lying behavior to detect hoof lesions in dairy cows. *J. Dairy Sci.* 92:4365–4374.
- Cicchetti, D. V., and T. Allison. 1971. A new procedure for assessing reliability of scoring EEG sleep recordings. *Am. J. EEG Technol.* 11:101–109.
- Cicchetti, D. V., and A. R. Feinstein. 1990. High agreement but low kappa: II. Resolving the paradoxes. *J. Clin. Epidemiol.* 43:551–558.
- Cohen, J. 1968. Weighted kappa—Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.* 70:213–220.
- Cook, T. D. 2002. Advanced statistics: Up with odds ratios! A case for odds ratios when outcomes are common. *Acad. Emerg. Med.* 9:1430–1434.
- de Mol, R. M., G. André, E. J. B. Bleumer, J. T. N. van der Werf, Y. de Haas, and C. G. van Reenen. 2013. Applicability of day-to-day variation in behavior for the automated detection of lameness in dairy cows. *J. Dairy Sci.* 96:3703–3712.
- de Vet, H. C. W., C. B. Terwee, D. L. Knol, and L. M. Bouter. 2006. When to use agreement versus reliability measures. *J. Clin. Epidemiol.* 59:1033–1039.
- Engel, B., G. Bruin, G. Andre, and W. Buist. 2003. Assessment of observer performance in a subjective scoring system: Visual classification of the gait of cows. *J. Agric. Sci.* 140:317–333.
- Flower, F. C., and D. M. Weary. 2006. Effect of hoof pathologies on subjective assessments of dairy cow gait. *J. Dairy Sci.* 89:139–146.
- Flower, F. C., and D. M. Weary. 2009. Gait assessment in dairy cattle. *Animal* 3:87–95.
- Graham, P., and B. Bull. 1998. Approximate standard errors and confidence intervals for indices of positive and negative agreement. *J. Clin. Epidemiol.* 51:763–771.
- Hoehler, F. K. 2000. Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *J. Clin. Epidemiol.* 53:499–503.

- Hoffman, A. C., D. A. Moore, J. Vanegas, and J. R. Wenz. 2014. Association of abnormal hind-limb postures and back arch with gait abnormality in dairy cattle. *J. Dairy Sci.* 97:2178–2185.
- Kaufman, A. B., and R. Rosenthal. 2009. Can you believe my eyes? The importance of interobserver reliability statistics in observations of animal behaviour. *Anim. Behav.* 78:1487–1491.
- Kazdin, A. E. 1977. Artifact, bias, and complexity of assessment: The ABCs of reliability. *J. Appl. Behav. Anal.* 10:141–150.
- Kottner, J., L. Audigé, S. Brorson, A. Donner, B. J. Gajewski, A. Hróbjartsson, C. Roberts, M. Shoukri, and D. L. Streiner. 2011. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *J. Clin. Epidemiol.* 64:96–106.
- Landis, J. R., and G. G. Koch. 1977. Measurement of observer agreement for categorical data. *Biometrics* 33:159–174.
- Liu, J. B., R. M. Dyer, N. K. Neerchal, U. Tasch, and P. G. Rajkondawar. 2011. Diversity in the magnitude of hind limb unloading occurs with similar forms of lameness in dairy cows. *J. Dairy Res.* 78:168–177.
- Maertens, W., J. Vangeyte, J. Baert, A. Jantuan, K. C. Mertens, S. De Campeneere, A. Pluk, G. Opsomer, S. Van Weyenberg, and A. Van Nuffel. 2011. Development of a real time cow gait tracking and analysing tool to assess lameness using a pressure sensitive walkway: The GAITWISE system. *Biosystems Eng.* 110:29–39.
- Manson, F. J., and J. D. Leaver. 1988. The influence of concentrate amount on locomotion and clinical lameness in dairy-cattle. *Anim. Prod.* 47:185–190.
- March, S., J. Brinkmann, and C. Winkler. 2007. Effect of training on the inter-observer reliability of lameness scoring in dairy cattle. *Anim. Welf.* 16:131–133.
- Martin, P., and P. Bateson. 2007. *Measuring Behaviour: An Introductory Guide*. 3rd ed. Cambridge Univ. Press, Cambridge.
- McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*. 2nd ed. Chapman Hall, London, UK.
- Neveux, S., D. M. Weary, J. Rushen, M. A. G. von Keyserlingk, and A. M. de Passille. 2006. Hoof discomfort changes how dairy cattle distribute their body weight. *J. Dairy Sci.* 89:2503–2509.
- Pastell, M., L. Hanninen, A. M. de Passille, and J. Rushen. 2010. Measures of weight distribution of dairy cows to detect lameness and the presence of hoof lesions. *J. Dairy Sci.* 93:954–960.
- Pluk, A., C. Bahr, T. Leroy, A. Poursaberi, X. Song, E. Vranken, W. Maertens, A. Van Nuffel, and D. Berckmans. 2010. Evaluation of step overlap as an automatic measure in dairy cow locomotion. *Trans. ASABE* 53:1305–1312.
- Rajkondawar, P. G., A. M. Lefcourt, N. K. Neerchal, R. M. Dyer, M. A. Varner, B. Erez, and U. Tasch. 2002. The development of an objective lameness scoring system for dairy herds: Pilot study. *Trans. ASAE* 45:1123–1125.
- Rajkondawar, P. G., M. Liu, R. M. Dyer, N. K. Neerchal, U. Tasch, A. M. Lefcourt, B. Erez, and M. A. Varner. 2006. Comparison of models to identify lame cows based on gait and lesion scores, and limb movement variables. *J. Dairy Sci.* 89:4267–4275.
- Rushen, J., E. Pombourcq, and A. M. de Passille. 2007. Validation of two measures of lameness in dairy cows. *Appl. Anim. Behav. Sci.* 106:173–177.
- Rutherford, K. M. D., F. M. Langford, M. C. Jack, L. Sherwood, A. B. Lawrence, and M. J. Haskell. 2009. Lameness prevalence and risk factors in organic and non-organic dairy herds in the United Kingdom. *Vet. J.* 180:95–105.
- 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 intrarater 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.
- 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. 2015. Comparison of locomotion scoring for dairy cows by experienced and inexperienced raters using live or video observation methods. *Anim. Welf.* 24:69–79.
- Scott, G. B. 1988. Lameness and pregnancy in Friesian dairy-cows. *Br. Vet. J.* 144:273–281.
- Song, X. Y., T. Leroy, E. Vranken, W. Maertens, B. Sonck, and D. Berckmans. 2008. Automatic detection of lameness in dairy cattle—Vision-based trackway analysis in cow's locomotion. *Comput. Electron. Agric.* 64:39–44.
- 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.
- Thomsen, P. T. 2009. Rapid screening method for lameness in dairy cows. *Vet. Rec.* 164:689–690.
- Thomsen, P. T., L. Munksgaard, and F. A. Togersen. 2008. Evaluation of a lameness scoring system for dairy cows. *J. Dairy Sci.* 91:119–126.
- Van Hertem, T., E. Maltz, A. Antler, C. E. B. Romanini, S. Viazzi, C. Bahr, A. Schlageter-Tello, C. Lokhorst, D. Berckmans, and I. Halachmi. 2013. Lameness detection based on multivariate continuous sensing of milk yield, rumination, and neck activity. *J. Dairy Sci.* 96:4286–4298.
- Van Hertem, T., S. Viazzi, M. Steensels, E. Maltz, A. Antler, V. Alchanatis, A. A. Schlageter-Tello, K. Lokhorst, E. C. B. Romanini, C. Bahr, D. Berckmans, and I. Halachmi. 2014. Automatic lameness detection based on consecutive 3D-video recordings. *Biosystems Eng.* 119:108–116.
- Van Nuffel, A., M. Sprenger, F. A. M. Tuytens, and W. Maertens. 2009. Cow gait scores and kinematic gait data: can people see gait irregularities? *Anim. Welf.* 18:433–439.
- Viazzi, S., C. Bahr, A. Schlageter-Tello, T. Van Hertem, C. E. B. Romanini, A. Pluk, I. Halachmi, C. Lokhorst, and D. Berckmans. 2013. Analysis of individual classification of lameness using automatic measurement of back posture in dairy cattle. *J. Dairy Sci.* 96:257–266.
- Warrens, M. J. 2013. Conditional inequalities between Cohen's kappa and weighted kappas. *Stat. Methodol.* 10:14–22.
- Welfare Quality. 2009. *Assessment Protocol for Cattle*. Welfare Quality Consortium. Lelystad, the Netherlands.
- Whay, H. 2002. Locomotion scoring and lameness detection in dairy cattle. In *Pract.* 24:444–449.
- Winckler, C., and S. Willen. 2001. The reliability and repeatability of a lameness scoring system for use as an indicator of welfare in dairy cattle. *Acta Agric. Scand. A Anim. Sci.* 30:103–107.