Performance of raters to assess locomotion in dairy cattle

Andrés Schlageter Tello

Performance of raters to assess locomotion in dairy cattle

Andrés Schlageter Tello

Thesis committee

Promotor

Prof. Dr P. W. G. Groot Koerkamp
Professor of Farm Technology
Wageningen University

Co-Promotors

Dr C. Lokhorst
Senior Researcher, Animal Welfare Group
Wageningen UR Livestock Research

Dr E. A. M. Bokkers
Assistant professor, Animal Production Systems Group
Wageningen University

Other members

Dr B. Engel, Wageningen University

Prof. Dr B. Kemp, Wageningen University

Prof. Dr T. J. G. M. Lam, Utrecht University

Prof. Dr C. Winckler, University of Natural Resources and Life Sciences (BOKU), Vienna

This research was conducted under the auspices of the Graduate School of Wageningen Institute of Animal Science (WIAS)

Performance of raters to assess locomotion in dairy cattle

Andrés Schlageter Tello

Thesis

submitted in fulfilment of the requirements for the degree of doctor

at Wageningen University

by the authority of the Rector Magnificus

Prof. Dr M. J. Kropff,

in the presence of the

Thesis Committee appointed by the Academic Board

to be defended in public

on Wednesday 27th May 2015

at 1.30 p.m. in the Aula.

Schlageter Tello, Andrés

Performance of raters to assess locomotion in dairy cattle,

162 pages

PhD thesis, Wageningen University, Wageningen, NL (2015)

With references, with summaries in English and Dutch

ISBN: 978-94-6257-275-1

Abstract

Locomotion scoring systems are procedures used to evaluate the quality of cows' locomotion. When scoring locomotion, raters focus their attention on gait and posture traits that are described in the protocol. Using these traits, raters assign a locomotion score to cows according to a pre-determined scale. Locomotion scoring systems are mostly used to classify cows as lame or non-lame. A preselected threshold within the scale determines whether a cow is classified as lame or non-lame. Since lameness is considered an important problem in modern dairy farming evaluation of locomotion scoring systems is utmost important. The objective of this thesis was to evaluate the performance of raters to assess locomotion in dairy cattle in terms of reliability (defined as the ability of a measuring device to differentiate among subjects) and agreement (defined as the degree to which scores or ratings are identical). This thesis also explores possibilities for the practical application of locomotion scoring systems. In a literature review comprising 244 peer-reviewed articles, twenty-five locomotion scoring systems were found. Most locomotion scoring systems varied in the scale used and traits observed. Some of the most used locomotion scoring systems were poorly evaluated and, when evaluated, raters showed an important variation in reliability and agreement estimates. The variation in reliability and agreement estimates was confirmed in different experiments aiming to estimate the performance of raters for scoring locomotion and traits under different practical conditions. For instance, experienced raters obtained better intrarater reliability and agreement when locomotion scoring was performed from video than by live observation. In another experiment, ten experienced raters scored 58 video records for locomotion and for five different gait and posture traits in two sessions. A similar number of cows was allocated in each level of the five-level scale for locomotion scoring. Raters showed a wide variation in intra- and interrater reliability and agreement estimates for scoring locomotion and traits, even under the same practical conditions. When agreement was calculated for specific levels when scoring locomotion and traits, the lowest agreement tended to be in level 3 of a five-level scale. When a multilevel scale was transformed into a two-level scale, agreement increased, however, this increment was likely due to chance. The variation in reliability and agreement is explained by different factors such as the lack of a standard procedure for assessing locomotion or the characteristics of the population sample that is assessed. The factor affecting reliability and agreement most, however, is the rater him/herself. Although the probability for obtaining acceptable reliability and agreement levels increases with training and experience, it is not possible to assure that raters score cows consistently in every scoring session. Given the large variation in reliability and agreement, it can be concluded that raters have a moderate performance to assess consistently locomotion in dairy cows. The variable performance of raters when assessing locomotion limits the practical utility of locomotion scoring systems as part of animal welfare assessment protocols or as golden standard for automatic locomotion scoring systems.

Table of Contents

Chapter 1

General introduction

A. Schlageter Tello

1.1. Dairy developments

Milk and milk products are a big business worldwide with a production of approximately 626 million tonnes of fresh cow's milk in 2012. In 2012, European farmers produced 40.8% of annual production volume while American and Asian farmers produced 28.5% and 21.8%, respectively (FAOSTAT, 2015).

Dairy farming in developed countries has been through a period of intensification, which is characterized by increased milk production per cow (Rauw et al., 1998; Lucy, 2001; Basset-Mens et al., 2009). In the Netherlands, for instance, in period 2002-2013 annual milk deliveries to dairy plants increased from 10.3 to 12.2 million tonnes while cow numbers remained constant (about 1.5 million heads) (EUROSTAT, 2015). In the United Kingdom, in the period 1995-2013, the average annual milk yield per cow increased from 5,512 to 7,300 litres (DEFRA, 2015). This increase in milk production was achieved based mainly on genetic selection which strongly focused on increasing milk yield (Rauw et al., 1998; Oltenacu and Broom, 2010), and on improvements in nutrition and reproductive dairy herd management (Van Saun and Sniffen, 1996; Roche, 2006). In order to achieve these high production levels, dairy cows are sometimes forced to stretch their metabolic and physiological limits (Rauw et al., 1998). In addition, modern dairy farms are often designed to keep cows indoors throughout the year. Modern barns are constructed with concrete floors and often have inappropriate bedding material in resting areas predisposing to health and other welfare problems (Kristula et al., 2008; Cramer et al., 2009; de Vries et al., 2015). Increases in production levels have been related to health (Rauw et al., 1998; Ingvartsen et al., 2003), reproduction (Lucy, 2001) and welfare problems (Oltenacu and Broom, 2010) in dairy cows. According to some authors, the main problems affecting production and causing most economic losses include infertility, reproduction related issues, mastitis and lameness (Enting et al., 1997; Kossaibati and Esslemont, 1997; Ahlman et al., 2011; Wu et al., 2012). The estimated cost of lameness on production performance was € 190 per case (Ettema and Ostergaard, 2006).

1.2. The impact of lameness in dairy farming

Lameness is highly prevalent in modern dairy farms with reported average prevalence of 37% in England and Wales (Barker et al., 2010); 33% in Austria and Germany (Dippel et al., 2009 a, b); and from 21% to 55% in the USA (Cook, 2003; Espejo et al., 2006; von Keyserlingk et al., 2012). Lameness has been associated to impaired production performance in several ways. Archer et al. (2010) showed that cows classified as severely lame reduced the 305-d milk production by 350 kg. Warnick et al. (2001) reported that after 2 weeks of being classified as lame, cows decreased milk yield by 1.5 kg/d. Lameness was associated with a higher somatic cell count (Archer et al., 2011), a decreased expression of oestrus behaviour (Walker et al., 2008), a prolonged lapse between calving

to first service and between first service and conception (Barkema et al., 1994), and a reduced embryo survival rate (Beltman et al., 2009). Lameness also increases the probability of culling cows from the production system according to Barkema et al. (1994). They found that approximately 26% of all culled cows were culled due to lameness. Awareness of the prevalence of lameness and its potential impact is of utmost importance to farmers and the dairy industry.

1.3. Lameness and locomotion scoring

Locomotion scoring systems are procedures used to evaluate the quality of the locomotion of animals. When scoring locomotion, raters focus their attention on gait and posture traits that are described in the protocol of the applied locomotion scoring system. Using these traits, raters assign a locomotion score to cows according to a pre-determined scale. In dairy farming, locomotion scoring systems are mainly used to classify cows as lame or non-lame (Whay, 2002; Flower and Weary, 2009). Locomotion scoring systems are also used in other livestock species, e.g. horses (Hewetson et al., 2006), pigs (Anil et al., 2007) and sheep (Kaler et al., 2009; Phythian et al., 2013). These scoring systems all aim to create comparable records for management and research of lameness (Whay, 2002; Flower and Weary, 2009).

Lameness management, in which locomotion scoring plays a crucial role, involves various steps (see Figure 1.1). During step 1, each cow is observed to evaluate gait and posture traits in order to assign a score for the quality of locomotion. This is usually done on a multilevel ordinal scale running from normal to severely impaired locomotion. In step 2, cows are classified as lame or non-lame when a predetermined threshold in the scale is exceeded, usually the middle level of the scale. Although in literature there are different definitions about when and how to classify a cow as lame (Alban et al., 1996; Murray et al., 1996; Flower and Weary, 2009; Bicalho and Oikonomou, 2013), in this thesis a cow was classified as lame when a cow is locomotion scored as level 3 or higher using a five-level scale with a range from 1 to 5, unless specified otherwise.

It is commonly assumed that cows classified as lame suffer pain due to either hoof or other limb lesions (Flower and Weary, 2009). Therefore, locomotion scoring systems are also used to detect hoof or other limb lesions (Step 3, Figure 1.1). In this regard, locomotion scoring systems have been included in programs aimed at improving hoof health (DairyCo., 2007; Alberta Dairy Hoof Health Project, 2014). In addition, since lameness is associated with hoof and other limb lesions that compromise welfare of cows (Whay et al., 2003; Bruijnis et al., 2012), locomotion scoring systems, have also been included in several animal welfare assessment protocols (University of Bristol, 2004; Bracke, 2009; Welfare Quality, 2009; Bayvel et al., 2012).

The final step within lameness management using locomotion scoring systems involves the choice between an appropriate treatment strategy or culling (Step 4, Figure 1.1).

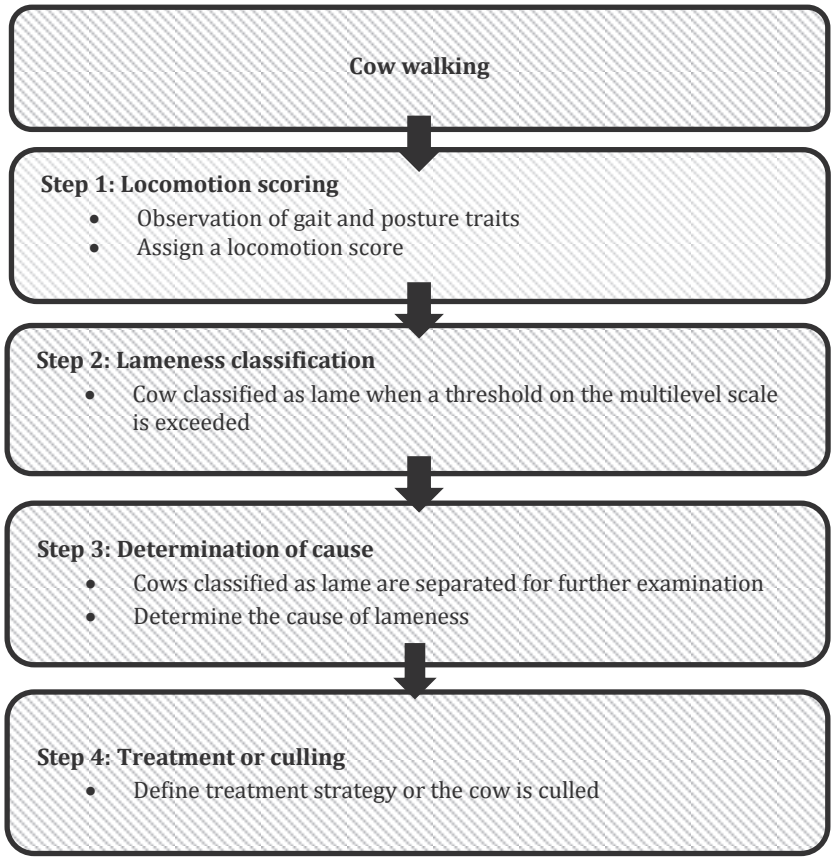


Figure 1.1 The steps in lameness management with the link between locomotion scoring, lameness classification, cause determination and treatment.

Lack of time is often mentioned by farmers as the main reason for not having good on-farm management and control of lameness (Leach et al., 2010). In order to improve the control and management of lameness, several attempts to develop automatic locomotion scoring systems have been made. Automatic locomotion scoring systems use sensors, instead of raters, to collect on-farm data. Data from these sensors are analysed using mathematical algorithms to assess the locomotion of cows and to classify them as either lame or not (Viazzi et al., 2013; Van Hertem et al., 2014). This thesis was conducted as part of a large

EU project (BioBusiness) aiming to develop an automatic locomotion scoring system for lameness management.

1.4. The BioBusiness Project

The BioBusiness project was part of the Marie Curie Initial Training Networks (http://www.bio-business.eu/index.php). The main objectives of the BioBusiness project were: a) to train biological and technological focused early stage researchers in emerging technologies and biological processes, product development, marketing and sales; and b) to integrate the knowledge of the aforementioned professionals in order to develop technological business solutions for the improvement of the welfare and performance of livestock.

The BioBusiness project sought solutions for three animal welfare problems. Part A investigated options to improve eggs incubation and broiler performance. Part B investigated automatic monitoring of pig aggression. And Part C which included the work presented in this thesis, investigated a business solution for lameness management on dairy farms using computer vision techniques.

Each partner within Part C of the BioBusiness project had a specific role towards the project goal. A team from the Agricultural Research Organization (Israel) was responsible for the construction of an experimental setup and for developing algorithms for locomotion scoring based on production and behaviour data. Another team from KU Leuven (Belgium) was responsible for the development of an algorithm to measure the back curvature of cows based on computer vision techniques. A Swedish team from DeLaval was responsible for project management and development of a commercial product. Our team from Wageningen UR (The Netherlands) supported the algorithms developed by the teams from Israel and Belgium by evaluating the reference or "golden standard" used for calibration and validation of the automatic locomotion scoring system.

The golden standard provides the definition of a case (e.g. lameness case) and the true reference to evaluate the performance of a new diagnostic tool (Coggon et al., 2005). In a perfect world, the golden standard is a theoretical method or procedure that is absolutely valid and consistent (Dohoo et al., 2003). However, in reality the golden standard is the best or closest method available to determine a case (Dohoo et al., 2003). By definition a mathematical algorithm (in this case to determine the locomotion score of a cow) cannot be more accurate or precise than its golden standard. Thus, the selection of the best available method to assess locomotion is a critical point in the development of automatic locomotion scoring systems. In this particular case, the logical selection for a golden standard is one of the commonly accepted locomotion scoring systems. Evaluation of existing locomotion scoring systems is of utmost important for supporting the development of an automatic locomotion scoring system.

1.5. Evaluation of locomotion scoring systems

According to Figure 1.1 locomotion scoring systems are procedures designed to help in lameness management and the identification of cows suffering hoof or other limb lesions by evaluating cows' locomotion. In this regard, the usefulness of locomotion scoring is dependent on a) the performance of raters for scoring locomotion consistently, and b) cows classified as lame are indeed affected by hoof or other limb lesions.

Although all the steps described in Figure 1.1 are important within the process of managing a lameness case, Step 1, locomotion scoring, is particularly important because it is the basic measurement on which lame or non-lame classification is based. Therefore, it is essential to gain more insight into the performance of raters to assess locomotion in dairy cattle. The performance of raters to assess locomotion using this scoring system can be evaluated by estimating their consistency in terms of reliability and agreement. In this thesis, the term consistency includes both concepts (reliability and agreement).

1.6. The concept of reliability and agreement in locomotion scoring

The performance of raters to assess locomotion consistently can be estimated by calculating reliability and agreement. In this thesis the definitions of reliability and agreement proposed by Kottner et al. (2011) were used. They defined agreement as "*the degree to which scores or ratings are identical*". Agreement is commonly estimated as percentage of agreement (PA), which is the number of agreements divided by the total number of observations expressed as percentage. Specific agreement is an agreement estimator used to determine the agreement in each level within the scale. Reliability is defined as "*the ability of a measuring device (e.g. locomotion scoring systems) to differentiate among subjects or objects*" (Kottner et al., 2011). Although other reliability estimators are available, reliability is commonly estimated using kappa (κ, for binary scales) or weighted kappa (κw, for ordinal multi-level scales). Both κ and κw were created to correct the PA by the expected agreement by chance (Cohen, 1960; 1968).

For a better understanding of the concept of reliability, knowledge on the effect of a heterogeneous or homogeneous population sample on reliability estimators is required. Heterogeneous population samples tend to contain individuals with the characteristic under study equally distributed across the various levels of the scale. On the other hand, homogeneous population samples tend to have most individuals distributed throughout a single level of the scale (de Vet et al., 2006; Kottner et al., 2011; Kottner and Streiner, 2011). Table 1.1 provides examples of locomotion scores from two raters in a population sample tending to be heterogeneous (examples A and B), and in a population sample tending to be homogeneous (examples C and D).

Table 1.1 Examples of locomotion scoring performed by two raters in population samples tending to be heterogeneous (Examples A and B) and homogeneous (Examples C and D) and different levels of disagreement by raters (Data are fictitious)

Example A. Similar number of agreements per level and disagreements are located close to agreement.

Levels	1	2	3	4	5
1	10	3	0	0	0
2	3	10	3	0	0
3	0	3	10	3	0
4	0	0	3	10	3
5	0	0	0	3	10

Example B. Similar number of agreements per level and disagreements are dispersed across levels.

Levels	1	2	3	4	5
1	10	2	1	1	0
2	2	10	2	1	0
3	1	2	10	2	1
4	1	1	2	10	2
5	0	0	1	2	10

Example C. Most agreements in levels 1, 2 and 3, and disagreements are located close to agreement (A situation that simulates locomotion scoring under practical conditions)

Levels	1	2	3	4	5
1	10	5	0	0	0
2	5	30	5	0	0
3	0	5	8	1	0
4	0	0	1	1	1
5	0	0	0	1	1

Example D. Most agreements in levels 1, 2 and 3, and disagreements dispersed across levels (A situation that simulates locomotion scoring under practical conditions)

Levels	1	2	3	4	5
1	10	2	2	2	0
2	2	30	3	2	0
3	2	3	8	1	0
4	2	2	1	1	0
5	0	0	0	0	1

Table 1.2 shows κw, PA and specific agreement for the examples shown in Table 1.1. Table 1.2 shows that κw values vary (0.42 - 0.79) for examples A, B, C and D while values for PA remain constant at 67.6%. The κw has higher values in heterogeneous (examples A and B, Table 1.2) than in homogeneous (examples C and D, Table 1.2) population samples with a constant PA. This is due to the probability of agreement by chance being higher in homogeneous than in heterogeneous population samples (Vach, 2005; de Vet et al., 2006).

The specific agreement is useful to estimate the performance of raters for identifying individual levels within the scale (Cicchetti and Feinstein, 1990; Kottner et al., 2011). Note that in a homogeneous population sample, specific agreement tends to be higher at the levels in which most individuals are distributed (levels 1 and 2, examples C and D, Table 1.2) and lower at those levels containing fewer individuals (Levels 3 and 4, examples C and D, Table 1.2). This is due to disagreements having a greater impact at levels with fewer individuals than at levels with more individuals. Although difficult, obtaining good specific

agreement in levels with few individuals is possible by minimizing the disagreements among raters.

As shown in examples C and D, low κw values are associated to a large difference between the higher and lower specific agreement, indicating that raters have difficulties to differentiate between levels of the scale. Thus, κw indicates the overall performance of raters for differentiating assessment between levels (e.g. Example D, levels 3 and 4) and specific agreement indicates which levels are difficult to differentiate.

Table 1.2. Reliability expressed as weighted kappa (κw), agreement expressed as percentage of agreement (PA) and specific agreement depending on the characteristics of the population sample and the level of disagreement between two raters for locomotion scoring assessed in a five-level (Lev) ordinal scale. (Data calculated from fictitious data in Table 1.1).

Example[a]	Reliability	Agreement					
	κw (-)	PA (%)	Lev 1 (%)	Lev 2 (%)	Lev 3 (%)	Lev 4 (%)	Lev 5 (%)
A	0.79	67.6	76.9	69.0	62.5	62.5	76.9
B	0.70	67.6	74.1	69.0	62.5	64.5	74.1
C	0.63	67.6	66.7	80.0	57.1	33.3	50.0
D	0.42	67.6	62.5	84.5	57.1	18.2	100.0

[a] Example A: Similar number of agreements per level and disagreements are located close to agreement; Example B: Similar number of agreements per level and disagreements are dispersed across levels; Example C: Most agreements in levels 1, 2 and 3, and disagreements are located close to agreement; Example D: Most agreements in levels 1, 2 and 3, and disagreements dispersed across levels

Reliability estimators are also affected by rater performance. This is reflected in lower κw values when disagreement between raters occurred at 2 or 3 levels (examples B and D, Table 1.2) than in cases when disagreements occurred at a single level (examples A and C, Table 1.2). This occurs because κw applies different weights for the level of disagreement between raters (e.g. disagreement at a single level is allocated more weight than disagreement at two, three or more levels). Thus, reliability indicators are affected by the characteristics of the population sample and rater performance.

Often a five-level scale is converted into a two-level scale to get a lame and non-lame classification. Table 1.3 shows κ coefficient, PA and specific agreement for lame and non-lame classifications from the cross-tables shown in Table 1.1. In Table 1.3, the κ coefficient is lower in examples with a low prevalence of lameness (Examples C and D, e.g. homogeneous population sample). The κ values tend to be low, even when PA is high (≥ 75%) as greater difference in the specific agreement for lame and non-lame cows. Thus low κ values indicates that raters are unable to differentiate properly between lame and non-lame cows in homogenous population samples.

Table 1.3. Reliability expressed as kappa (κ), agreement expressed as percentage of agreement (PA) and specific agreement depending on the characteristics of the population sample and the level of disagreement between raters in a two-level scale for lame or non-lame classification. (Data calculated from fictitious data in Table 1.1).

Examples[a]	Reliability	Agreement		
	κ (-)	PA (%)	Non-lame (%)	Lame (%)
A	0.83	91.8	89.6	93.3
B	0.71	86.5	82.7	88.8
C	0.65	86.5	90.9	73.6
D	0.40	75.6	83.2	57.1

[a] Example A: Similar number of agreements per level and disagreements are located close to agreement; Example B: Similar number of agreements per level and disagreements are dispersed across levels; Example C: Most agreements in levels 1, 2 and 3, and disagreements are located close to agreement; Example D: Most agreements in levels 1, 2 and 3, and disagreements dispersed across levels

The concepts of reliability and agreement as proposed by Kottner et al. (2011) have not been used to analyse the consistency of raters performing locomotion scoring. Performing locomotion scoring in a heterogeneous population sample provides a better evaluation of the performance of raters to score locomotion by minimizing the population sample effects on agreement by chance. However, under practical conditions, population samples tend to be homogeneous with most cows distributed across levels 1 and 2 of a five-level scale (Thomsen et al., 2008). Therefore, it is important to evaluate performance of raters to assess locomotion in both homogeneous and heterogeneous population samples.

1.7. General objective

Based on the importance associated to lameness in dairy farming and the positioning of this study within the BioBusiness project, the objective of this thesis is to evaluate the performance of raters to assess locomotion in dairy cattle in terms of reliability and agreement.

This thesis will also explore possibilities for the practical application of locomotion scoring systems related to lameness classification (Step 2, Figure 1.1) and detection of hoof and other limb lesions (Step 3, Figure 1.1). Finally, since the research was conducted within the framework of the BioBusiness project, this thesis will discuss the usefulness of automatic locomotion scoring systems for classifying cows as lame and the detection of hoof lesions and the possibilities for on-farm application.

1.8. Outline of the thesis

Many different types of locomotion scoring systems have been described in scientific literature, and no overview exists concerning the performance of raters to assess locomotion. In addition, no articles were found that focus on studying the performance of raters in terms of reliability and agreement as proposed by Kottner et al. (2011). Therefore, **Chapter 2**, contains a systematic literature review aimed at describing different locomotion scoring systems (manual and automatic), and an analysis of the performance of both systems to evaluate locomotion.

Since automatic locomotion scoring systems often make use of video imaging, it is important to determine whether or not video images of walking cows can be used to replace live observations for locomotion scoring. **Chapter 3** contains details of an investigation into reliability and agreement of live locomotion scoring in comparison to video.

It is common practice to merge adjacent levels within a locomotion score to improve consistency of rater evaluation. However, merging levels results in a loss of resolution and a reduction in information concerning locomotion scores (Engel et al., 2003). Until now, no studies have been done on the effect of merging levels. Therefore, in an attempt to fill this gap in information, **Chapter 4** contains an evaluation of ways of merging levels to optimize resolution, reliability and agreement of locomotion scoring in dairy cows.

Some automatic locomotion scoring systems attempt to mimic locomotion scoring performed by human raters by measuring traits using different types of sensors(Maertens et al., 2011; Viazzi et al., 2013; Van Hertem et al., 2014). Unlike raters, most automatic locomotion scoring systems focus on the measurement and analysis of only a single trait of gait or posture. Therefore, it might be beneficial to determine the relative importance of single traits within a locomotion scoring system. **Chapter 5** provides an investigation of the associations between scores assigned to locomotion and locomotion traits as made by experienced raters.

A discussion of the findings of this research is provided in **Chapter 6**. Furthermore, recommendations are provided for application of locomotion scoring systems and automatic locomotion scoring systems for lameness classification and lesion detection. In this chapter the conclusions of this thesis are also presents.

References

Ahlman, T., B. Berglund, L. Rydhmer, and E. Strandberg. 2011. Culling reasons in organic and conventional dairy herds and genotype by environment interaction for longevity. J. Dairy Sci. 94:1568-1575.

Alban, L., J. F. Agger, and L. G. Lawson. 1996. Lameness in tied Danish dairy cattle: The possible influence of housing systems, management, milk yield, and prior incidents of lameness. Prev. Vet. Med. 29:135-149.

Alberta Dairy Hoof Health Project. 2014. Dairy claw lesion identification. Access date: Jul 15, 2014.Web page:http://www.hoofhealth.ca/Section5/DD_Summer_2009_Lesion_ID.pdf.

Anil, L., S. S. Anil, and J. Deen. 2007. Analysis of the association of claw lesions with lameness in breeding sows. J. Dairy Sci. 90:4-4.

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.

Archer, S. C., M. J. Green, A. Madouasse, and J. N. Huxley. 2011. Association between somatic cell count and serial locomotion score assessments in UK dairy cows. J. Dairy Sci. 94:4383-4388.

Barkema, H. W., J. D. Westrik, K. A. S. Vankeulen, Y. H. Schukken, and A. Brand. 1994. The Effects of Lameness on Reproductive-Performance, Milk-Production and Culling in Dutch Dairy Farms. Prev. Vet. Med. 20:249-259.

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.

Basset-Mens, C., S. Ledgard, and M. Boyes. 2009. Eco-efficiency of intensification scenarios for milk production in New Zealand. Ecol. Econ. 68:1615-1625.

Bayvel, A. C. D., T. J. Diesch, and N. Cross. 2012. Animal welfare: a complex international public policy issue: economic, policy, societal, cultural and other drivers and constraints. A 20-year international perspective. Anim. Welfare 21:11-18.

Beltman, M. E., P. Lonergan, M. G. Diskin, J. F. Roche, and M. A. Crowe. 2009. Effect of progesterone supplementation in the first week post conception on embryo survival in beef heifers. Theriogenology 71:1173-1179.

Bicalho, R. C. and G. Oikonomou. 2013. Control and prevention of lameness associated with claw lesions in dairy cows. Livest. Sci. 156:96-105.

Bracke, M. B. M. 2009. Animal welfare in a global perspective - A survey of foreing agricultural services and case studies on poultry, aquaculture and wildlife. Wageningen UR Livestock Research, Lelystad.

Bruijnis, M. R. N., B. Beerda, H. Hogeveen, and E. N. Stassen. 2012. Assessing the welfare impact of foot disorders in dairy cattle by a modeling approach. Animal 6:962-970.

Cicchetti, D. V. and A. R. Feinstein. 1990. High agreement but low kappa: II. Resolving the paradoxes. J. Clin. Epidemiol. 43:551-558.

Coggon, D., C. Martyn, K. T. Palmer, and B. Evanoff. 2005. Assessing case definitions in the absence of a diagnostic gold standard. Int. J. Epidemiol. 34:949-952.

Cohen, J. 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20:37-46.

Cohen, J. 1968. Weighted Kappa - nominal scale agreement with provision for scaled disagreement or partial credit. Psychol. Bull. 70:213-220.

Cook, N. B. 2003. Prevalence of lameness among dairy cattle in Wisconsin as a function of housing type and stall surface. J. Am. Vet. Med. Assoc. 223:1324-1328.

Cramer, G., K. D. Lissemore, C. L. Guard, K. E. Leslie, and D. F. Kelton. 2009. Herd-level risk factors for seven different foot lesions in Ontario Holstein cattle housed in tie stalls or free stalls. J. Dairy Sci. 92:1404-1411.

DairyCo. 2007. DairyCo mobility score. Access date: May 10, 2011.Web page:http://www.dairyco.org.uk/.

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.

de Vries, M., E. A. M. Bokkers, C. G. van Reenen, B. Engel, G. van Schaik, T. Dijkstra, and I. J. M. de Boer. 2015. Housing and management factors associated with indicators of dairy cattle welfare. Prev. Vet. Med. 118:80-92.

DEFRA. 2015. UK supplies of milk products Access date: Jan. 27, 2015.Web page:https://www.gov.uk/government/statistics.

Dippel, S., M. Dolezal, C. Brenninkmeyer, J. Brinkmann, S. March, U. Knierim, and C. Winckler. 2009a. Risk factors for lameness in cubicle housed Austrian Simmental dairy cows. Prev. Vet. Med. 90:102-112.

Dippel, S., M. Dolezal, C. Brenninkmeyer, J. Brinkmann, S. March, U. Knierim, and C. Winckler. 2009b. Risk factors for lameness in freestall-housed dairy cows across two breeds, farming systems, and countries. J. Dairy Sci. 92:5476-5486.

Dohoo, I., W. Martin, and S. H. 2003. Screening and diagnostic tests. Pages 85-120. in Veterinary Epidemiologic Research. AVC Inc., Charlottetown. U. S. A.

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.

Enting, H., D. Kooij, A. A. Dijkhuizen, R. B. M. Huirne, and E. N. Noordhuizen-Stassen. 1997. Economic losses due to clinical lameness in dairy cattle. Livest. Prod. Sci. 49 259 - 267.

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

Ettema, J. F. and S. Ostergaard. 2006. Economic decision making on prevention and control of clinical lameness in Danish dairy herds. Livest. Sci. 102:92-106.

EUROSTAT. 2015. Agricultural production: Milk and milk products. Access date: Jan. 27, 2015.Web page:http://ec.europa.eu/eurostat.

FAOSTAT. 2015. Milk production quantities by country. Access date: Jan. 27, 2015.Web page:http://faostat3.fao.org/browse/Q/QL/E.

Flower, F. C. and D. M. Weary. 2009. Gait assessment in dairy cattle. Animal 3:87-95.

Hewetson, M., R. M. Christley, I. D. Hunt, and L. C. Voute. 2006. Investigations of the reliability of observational gait analysis for the assessment of lameness in horses. Vet. Rec. 158:852-858.

Ingvartsen, K. L., R. J. Dewhurst, and N. C. Friggens. 2003. On the relationship between lactational performance and health: is it yield or metabolic imbalance that cause production diseases in dairy cattle? A position paper. Livest. Prod. Sci. 83:277-308.

Kaler, J., G. J. Wassink, and L. E. Green. 2009. The inter- and intra-observer reliability of a locomotion scoring scale for sheep. Vet. J. 180:189-194.

Kossaibati, M. A. and R. J. Esslemont. 1997. The costs of production diseases in dairy herds in England. Vet. J. 154:41-51.

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.

Kottner, J. and D. L. Streiner. 2011. The difference between reliability and agreement. J. Clin. Epidemiol. 64:701-702.

Kristula, M. A., Z. Dou, J. D. Toth, B. I. Smith, N. Harvey, and M. Sabo. 2008. Evaluation of Free-Stall Mattress Bedding Treatments to Reduce Mastitis Bacterial Growth. J. Dairy Sci. 91:1885-1892.

Leach, K. A., H. R. Whay, C. M. Maggs, Z. E. Barker, E. S. Paul, A. K. Bell, and D. C. J. Main. 2010. Working towards a reduction in cattle lameness: 1. Understanding barriers to lameness control on dairy farms. Res. Vet. Sci. 89:311-317.

Lucy, M. C. 2001. Reproductive Loss in High-Producing Dairy Cattle: Where Will It End? J. Dairy Sci. 84:1277-1293.

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. Biosyst. Eng. 110:29-39.

Murray, R. D., D. Y. Downham, M. J. Clarkson, W. B. Faull, J. W. Hughes, F. J. Manson, J. B. Merritt, W. B. Russell, J. E. Sutherst, and W. R. Ward. 1996. Epidemiology of lameness in dairy cattle: Description and analysis of foot lesions. Vet. Rec. 138:586-591.

Oltenacu, P. A. and D. M. Broom. 2010. The impact of genetic selection for increased milk yield on the welfare of dairy cows. Anim. Welfare 19:39 - 49.

Phythian, C. J., P. C. Cripps, D. Grove-White, P. H. Jones, E. Michalopoulou, and J. S. Duncan. 2013. Observing lame sheep: evaluating test agreement between group-level and individual animal methods of assessment. Anim. Welfare 22:417-422.

Rauw, W. M., E. Kanis, E. N. Noordhuizen-Stassen, and F. J. Grommers. 1998. Undesirable side effects of selection for high production efficiency in farm animals: a review. Livest. Prod. Sci. 56:15-33.

Roche, J. F. 2006. The effect of nutritional management of the dairy cow on reproductive efficiency. Anim. Reprod. Sci. 96:282-296.

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.

University of Bristol. 2004. Bristol welfare assurance program: Cattle assessment, Version 2.0. University of Bristol, Bristol, UK.

Vach, W. 2005. The dependence of Cohen's kappa on the prevalence does not matter. J. Clin. Epidemiol. 58:655-661.

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. Biosyst. Eng. 119:108-116.

Van Saun, R. J. and C. J. Sniffen. 1996. Nutritional management of the pregnant dairy cow to optimize health, lactation and reproductive performance. Anim. Feed Sci. Tech. 59:13-26.

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.

von Keyserlingk, M. A. G., A. Barrientos, K. Ito, E. Galo, and D. M. Weary. 2012. Benchmarking cow comfort on North American freestall dairies: Lameness, leg injuries, lying time, facility design, and management for high-producing Holstein dairy cows. J. Dairy Sci. 95:7399-7408.

Walker, S. L., R. F. Smith, J. E. Routly, D. N. Jones, M. J. Morris, and H. Dobson. 2008. Lameness, Activity Time-Budgets, and Estrus Expression in Dairy Cattle. J. Dairy Sci. 91:4552-4559.

Warnick, L. D., D. Janssen, C. L. Guard, and Y. T. Grohn. 2001. The effect of lameness on milk production in dairy cows. J. Dairy Sci. 84:1988-1997.

Welfare Quality. 2009. Assessment Protocol for Cattle. in Welfare Quality Consortium. Lelystad, The Netherlands.

Whay, H. 2002. Locomotion scoring and lameness detection in dairy cattle. In Practice 24:444-449.

Whay, H. R., D. C. J. Main, L. E. Green, and A. J. F. 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.

Wu, J. J., D. C. Wathes, J. S. Brickell, L. G. Yang, Z. Cheng, H. Q. Zhao, Y. J. Xu, and S. J. Zhang. 2012. Reproductive performance and survival of Chinese Holstein dairy cows in central China. Anim. Prod. Sci. 52:11-19.

Chapter 2

Manual and automatic locomotion scoring systems in dairy cows: A review

Published in Preventive Veterinary Medicine 116 (2014) 12–25

A. Schlageter Tello, 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

Abstract

The objective of this review was to describe, compare and evaluate agreement, reliability, and validity of manual and automatic locomotion scoring systems (MLSSs and ALSSs, respectively) used in dairy cattle lameness research. There are many different types of MLSSs and ALSSs. Twenty-five MLSSs were found in 244 articles. MLSSs use different types of scale (ordinal or continuous) and different gait and posture traits need to be observed. The most used MLSS (used in 28% of the references) is based on asymmetric gait, reluctance to bear weight, and arched back, and is scored on a five-level scale. Fifteen ALSSs were found that could be categorized according to three approaches: (a) the kinetic approach measures forces involved in locomotion, (b) the kinematic approach measures time and distance of variables associated to limb movement and some specific posture variables, and (c) the indirect approach uses behavioural variables or production variables as indicators for impaired locomotion.

Agreement and reliability estimates were scarcely reported in articles related to MLSSs. When reported, inappropriate statistical methods such as PABAK and Pearson and Spear- man correlation coefficients were commonly used. Some of the most frequently used MLSSs were poorly evaluated for agreement and reliability. Agreement and reliability estimates for the original four-, five- or nine-level MLSS, expressed in percentage of agreement, kappa and weighted kappa, showed large ranges among and sometimes also within articles. After the transformation into a two-level scale, agreement and reliability estimates showed accept- able estimates (percentage of agreement ≥75%; kappa and weighted kappa ≥0.6), but still estimates showed a large variation between articles. Agreement and reliability estimates for ALSSs were not reported in any article.

Several ALSSs use MLSSs as a reference for model calibration and validation. However, varying agreement and reliability estimates of MLSSs make a clear definition of a lameness case difficult, and thus affect the validity of ALSSs. MLSSs and ALSSs showed limited validity for hoof lesion detection and pain assessment.

The utilization of MLSSs and ALSSs should aim to the prevention and efficient management of conditions that induce impaired locomotion. Long-term studies comparing MLSSs and ALSSs while applying various strategies to detect and control unfavourable conditions leading to impaired locomotion are required to determine the usefulness of MLSSs and ALSSs for securing optimal production and animal welfare in practice.

2.1. Introduction

Manual locomotion scoring systems (MLSSs) are used to indicate the quality of locomotion of cows. With MLSSs, human raters look at specific gait and posture traits to score the

locomotion of a cow on a scale indicating an increasing level of impaired locomotion. In literature, the term "lameness" is mainly associated to the presence of impaired locomotion in cows (Winckler and Willen, 2001; Flower and Weary, 2009). MLSSs have become a popular tool for lameness detection because they are inexpensive, non- invasive and easily applied under farm conditions (Whay, 2002).

Lameness is a major problem on dairy farms (Bruijnis et al., 2010). Lameness has been associated to a negative effect on milk yield (Warnick et al., 2001; Green et al., 2002; Archer et al., 2010), on reproductive performance (Garbarino et al., 2004; Walker et al., 2008, 2010), and it also increases the risk of culling (Barkema et al., 1994; Booth et al., 2004). Lameness is also considered an important welfare issue, because it is assumed to be a visual sign of underlying problems, e.g. hoof lesions or other painful limb injuries (Whay, 2002; Flower and Weary, 2009). Several programmes aiming to improve animal welfare (Welfare Quality, 2009) and hoof health (DairyCo., 2007) include a MLSSs. Thus, MLSSs are considered a tool for detection of hoof lesions and pain.

Regularly scoring locomotion should be a priority task for dairy farmers. However, as the number of cows per herd increases, farmers' time available to perform MLSS is likely to get less. This is one of the main reasons to develop automatic locomotion scoring systems (ALSSs). ALSSs collect on-farm data from cows using sensors. Data from these sensors is analysed using mathematical algorithms to assess the locomotion of cows.

Three approaches have been commonly used in ALSSs: kinetic, kinematic and indirect. The kinetic approach measures forces involved in locomotion (Flower and Weary, 2009). The kinematic approach measures time and distance of variables associated to limb movement and some specific posture variables (Flower and Weary, 2009). The indirect approach uses behavioural or production variables as indicators for impaired locomotion.

Development of new tools, such as ALSSs, requires a reference or golden standard for calibration and validation. ALSSs are validated mainly using MLSSs. However, there are several limitations of using MLSSs (Whay et al., 1997; Flower et al., 2006; Chapinal et al., 2009). MLSSs do not always have high interrater and intrarater agreement and reliability (O'Callaghan et al., 2003; Channon et al., 2009; Kottner et al., 2011). In addition, validity of MLSSs as a tool for detection of hoof lesions and pain is not always obvious (Winckler and Willen, 2001; Flower and Weary, 2006; Rushen et al., 2007). So far, no systematic overview to address these limitations has been published. The objectives of this review, therefore, were to describe, compare, and evaluate agreement, reliability and validity of current manual and automatic locomotion scoring systems.

2.2. Covered literature

This review covers peer-reviewed articles, published in English between 1945 and December 2013, reporting on the use of at least one MLSS or ALSS in the material and methods section. Literature search used three databases: ISI Web of Knowledge (Thomson Reuters, U.S.A.), MED- LINE/PubMed (U.S. National Library of Medicine, U.S.A.) and SciVerse Scopus (Elsevier C.V., The Netherlands), and was based on terms proposed by Hirst et al. (2002). Where relevant, publications cited in the included articles were also reviewed.

Information extracted from articles was stored in a database and MLSSs and ALSSs described were labelled according to their first published description (Tables 2.1 and 2.3). The database comprised 244 peer- reviewed articles published in 39 different journals, reporting research performed in 27 countries across five continents. Recent importance of lameness detection by using MLSSs and ALSSs was reflected by the fact that 70% of the articles were published after 2007. No articles published before 1988 were found.

Information extracted from MLSSs and ALSSs included gait and posture traits (Tables 2.1 and 2.2) or variables (Tables 2.3 and 2.4) that were either observed (for MLSSs) or measured (for ALSSs). These traits or variables were split into gait, which focused on alterations related to the limbs (e.g. asymmetric gait, reluctance to bear weight, and ground reaction force) and posture traits/variables, which focused on alterations related to parts of the body other than the limbs (e.g. back curvature, head bob) (Table 2.2). For ALSSs, also other behaviour and production variables were included (e.g. milk yield, lying time, walking speed, Table 2.4). Information extracted also included type of scale used (ordinal, continuous), range of the scale and number of levels for ordinal scales. If available, additional information was included, such as number of cows studied and duration of the experiment. For MLSSs, data on the background and level of training of the raters, the surface on which the cows walked, timing in relation to milking, whether locomotion scoring was performed live or from video, interrater and intrarater agreement and reliability, and statistical method used for agreement or reliability estimation were included.

For ALSSs, information on the approach (kinetic, kinematic or indirect), and the type of sensors used to measure gait and posture variables (e.g. force plates, accelerometers, camera) or behaviour (e.g. accelerometers) and production (e.g. milk metres) variables were included.

2.3. Manual locomotion scoring systems

Twenty-five MLSSs, which varied mainly in gait and posture traits observed and type of

scale used for locomotion scoring, were described in literature (Table 2.1). In 244 articles, 247 mentions were made to MLSSs and a five-level MLSS described by Sprecher et al. (1997) was mentioned most frequently (69 times of 247 references, about 28%). The nine-level MLSS described by Flower and Weary (2006) was mentioned 35 times (about 14%), and the MLSS developed by Manson and Leaver (1988) was mentioned 32 (13%) (Table 2.1). Other MLSSs were modifications or combinations of these three MLSSs (Garbarino et al., 2004; Amory et al., 2006; Thomsen et al., 2008).

Table 2.1. Manual locomotion scoring systems (MLSSs) described in 244 articles classified by type of scale (continuous or ordinal), minimum and maximum level of the scale (Min-Max), traits observed, and percentage (%) of papers using the MLSS

		Traits observed[a]			
MLSS	**Min – Max**	**Gait**	**Posture**	**Other**	**%[b]**
Continuous					**2.4**
Flower and Weary, 2006	0 - 100	AG–T-RBW-JF	AB-HB		1.6
Engel et al., 2003	0 - 1				0.4
Tuyttens et al., 2009[c]	0 - 10	AG-T-AA-RBW	AB-HB	Sp	0.4
Ordinal					**97.6**
13 Levels					0.4
Offinger et al., 2013	1-5 with +/-[d]	AG-T-RBW	AB-HB	Ris	0.4
9 Levels					29.6
Flower and Weary, 2006	1-5 with 0.5[e]	AG–T-RBW-JF	AB-HB		14.2
ICAR-Interbull	9 - 1	AG-S-AA			2.4
Manson and Leaver, 1988	1-5 with 0.5	AG-AA-RBW		Trn, Ris	13.0
6 Levels					4.0
Kestin et al., 1992	1 - 6				1.2
Garbarino et al, 2004	0 - 5	AG-S-RBW	AB		2.0
Fitzgerald et al., 2000	0 - 5				0.8
5 Levels					42.9
O'Callaghan et al., 2003	1 - 5	T-AA	AB-HB	Sp	3.6
Sprecher et al., 1997	1 - 5	AG-S-RBW	AB		27.9
Thomsen et al., 2008	1 - 5	AG-S-RBW	AB-HB		1.2
Thomsen, 2009	1 - 5		AB		0.4
Wells et al., 1993	0 - 4	AG		Ris	2.8
Winckler and Willen, 2001[f]	1 - 5	AG-S-RBW			6.9
4 Levels					8.5
Breuer et al., 2000	0 - 3	AG-RBW	HB		1.2
Cook, 2003	1 - 4	S-RBW	AB	Sp, Trn	2.0
DairyCO., 2007	0 - 3	AG-S-RBW	AB	Sp	4.9
Vokey et al., 2001	1 - 4	AG-S-AA	AB-HB-HH	Sp	0.4
3 Levels					6.9
Amory et al., 2006	1 - 3		AB		1.2
Welfare Quality®, 2009	0 - 2	AG-S-RBW			2.4
Sogstad et al., 2005	0 - 2	AG - RBW			1.6
Van Nuffel et al., 2009	1 - 3	AG-T-AA-RBW	AB-HB	Sp	1.6
2 Levels					5.3
Groehn et al., 1992 (Lame/Non-lame)	0 - 1				5.3

[a]: AG=Asymmetric Gait; T=Tracking up; AA=Abduction; JF=Joint flexibility; S=Step/stride length; RBW=Reluctance Bear Weight; AB= Arched Back; HB=Head Bob; HH=Hip Hick; Sp=Walking Speed; Trn=difficult turning; Ris=Difficult rising.
[b]: Percentage of utilization based on 247 manual locomotion scoring systems found in 244 articles.
[c]: VAS divided in three sections with different colours.
[d]: Range from 1 to 5, in each level is possible to assign a + or – (e.g: 3+ or 3-).
[e]: Range from 1 to 5, scale graded in half points.
[f]: Including papers using scores developed by Bicalho et al., (2007), with same characteristics as Winckler and Willen (2001).

The fact that 25 different MLSSs were described indicates that there is no consensus on a single MLSS. Several attempts, nevertheless, were made to design a standardized MLSS. The EU project Welfare Quality (2009) included a three-level MLSS in its assessment protocol for dairy cattle based on the MLSS described by Winckler and Willen (2001). The UK dairy industry introduced a four-level MLSS, commonly known as DairyCo. (2007). The International Committee for Animal Recording (ICAR) recommends a nine-level MLSS that includes not only observation of gait traits but also conformation traits, such as foot angle and conformation of rear legs from rear view. ICAR MLSS have been used mainly in studies related with genetics parameters (van der Waaij et al., 2005; Onyiro et al., 2008; Laursen et al., 2009).

Table 2.2. Abbreviation (Abb) and definition of gait and posture traits used in manual locomotion scoring systems.

Traits	Abb	Definition[a]
Gait		
Abduction or adduction	AA	A tendency to rotate the limb outwards and hock inwards (Abduction) or tendency to rotate the limb inwards (Adduction).
Asymmetric gait	AG	Asymmetry of distance/time in the imprints between two consecutive strides.
Joint flexibility	JF	Obvious joint stiffness characterized by lack of joint flexion.
Reluctance bear weight	RBW	Cow avoids bearing weight in the affected limb(s).
Short step	S	Diminished distance/time between two consecutive imprints of left and right hoof.
Tracking–up	T	Distance between the position of the front foot and hind foot on the same body side on the floor in the subsequent step
Posture		
Arched back	AB	Convex back line formed by the spine between the withers and tailbone.
Hip hick	HH	From behind, inclination of the imaginary horizontal line that joins the two pin bones
Head bob	HB	Exaggerated movement of the head when affected limb is lifted from the ground
Other		
Difficult turning	Trn	Difficulty in changing direction while walking
Difficult rising	Ris	Increase in time taken to stand up
Speed	Sp	Reduction in speed of displacement, compared with humans

[a] Based on definitions proposed by Whay (2002); Telezhenko and Bergsten (2005) and Maertens et al. (2011).

There is no standard protocol on how to perform MLSSs. Several studies scored cows walking on concrete surfaces, probably because this is common farm practice (90 of 244 articles). Independent of the surface selected, most studies agreed that manual locomotion scoring should be performed with cows walking on a flat, firm, and non-slippery surface. In 39 articles, scoring was performed after milking, probably because it

was compatible with normal farm work routines.

2.4. Automatic locomotion scoring systems

Fifteen different ALSSs were described (Table 2.3). In 244 articles, 30 mentions were made to ALSSs. The kinetic approach was mentioned ten times (33% of 30 mentions). Locomotion scoring was done either by measuring forces exerted on the floor surface by the hoofs while cows walked on two parallel force plates (Rajkondawar et al., 2002, 2006), or by measuring changes in weight distribution while cows stood on a platform containing four independent weight recording units (Pastell et al., 2008, 2010). Practical limitations associated with the kinetic approach are related to the positioning of the cow's hoof on the weighing units during measurement (Pastell and Kujala, 2007), or to the walking speed of cows that may affect the accuracy of the system (Scott, 1988).

The kinematic approach was mentioned 11 times (37% of 30 mentions) (Table 2.3). ALSSs use different techniques to obtain kinematic variables of locomotion (e.g. step length, step height, or back curvature). One technique in the kinematic approach uses markers (e.g. yellow circles) attached to hooves, limb joints, withers, or back-line contour. Video recordings of cows walking with markers are later analyzed with software for kinematic variables (Flower et al., 2005; Aoki et al., 2006; Blackie et al., 2013). Another technique uses image pre-processing, in which video recordings are transformed into sequences of binary images to facilitate the detection of anatomical parts of cows (Song et al., 2008; Van Hertem et al., 2013; Viazzi et al., 2013). A third technique involves pressure sensitive walkways (Van Nuffel et al., 2009; Maertens et al., 2011), which contain an array of pressure sensors. These sensors record the footprint of walking cows, which can be analyzed as kinematic variables of locomotion (Maertens et al., 2011). Finally, accelerometers attached to limbs allow measurements of acceleration of legs while cows walk (Pastell et al., 2009). Although pressure sensitive walkways and accelerometers are able to measure forces associated with locomotion (kinetic), force itself has not been shown a useful indicator for locomotion scoring.

ALSSs using the indirect approach, based on behavioural and production variables, were mentioned in nine articles (30% of 30 mentions) (Table 2.3). ALSSs based on behaviour use two-dimensional or three-dimensional accelerometers attached to the limbs or neck of cows to detect alterations in behaviour, such as duration of lying or standing bouts, and total time lying or standing per day (Ito et al., 2010; Alsaaod et al., 2012). Production data may be obtained by combining several sensors, such as milk metres or weight scales (for feed or live weight) (de Mol et al., 2013; Kamphuis et al., 2013). Behaviour and production are affected not only by lameness, but also by other common diseases, such as mastitis (DeVries et al., 2011) and ketosis (Goldhawk et al., 2009) as well as management and feeding.

Table 2.3. Automatic locomotion scoring systems (ALSSs) described in 244 articles classified by approach (kinetic, kinematic, and indirect), sensor used, gait, posture, behaviour and production variables measured, and percentage (%) of papers using ALSS.

ALSS	Sensor(s)[a]	Variables[b]			%[c]
		Gait	Posture	Other	
Kinetic approach					33.3
Rajkondawar et al., 2002	2PFP	GRF, ST			16.7
Pastell and Kujala, 2007	4WP	LWR, KN, StN, SDRL			16.7
Kinematic approach					36.7
Song et al., 2008	Cm	Tm			6.8
Pastell et al., 2009	Acc	Var Acc			3.3
Poursaberi et al., 2010	Cm		ABm		3.3
Blackie et al., 2011b	Cm	StrL		Spm	10.0
Maertens et al., 2011	PSW	ASpL, AST, ASpW, ASpT			6.7
Pluk et al., 2012	Cm/PSW	T&R angle			3.3
Viazzi et al., 2013	Cm		BMP		3.3
Indirect approach					26.7
Borderas et al., 2008	MR/Sl			MF, FV, MY, FT	6.7
Ito et al., 2010	Acc			LT, LBn, LBt	10.0
Blackie et al., 2011a	Acc/MM			LT, Sta, MY	3.3
Kamphuis et al., 2013	Acc/Sl/MM			Act, MO, MY, MD, LW	3.3
de Mol et al., 2013	Acc/MR/Sl			LT, LBn, Sta, MY, CLO	3.3
Chapinal et al., 2010b	4WP/Acc/Cm	SDRL		LBt, Spm	3.3

[a]: Type of sensor used for lameness detection for ALSS: 4WP=4 independent weighting platforms; 2PFP=2 parallel force plates; Cm=video camera; PSW=pressure sensitive walkway; MR=milking robot; Acc=accelerometer; Sl=Scale (for feed or live weight); MM=milk meters.
[b]: Act=Activity; ABm=Arched back measurement; ASpL=Asymmetry of step length; ASpT=Asymmetry of step time; ASpW=Asymmetry of step width; AST=Asymmetry of stance time; BMP=Body movement pattern; CLO=Concentrate left over; FV=Feed bunk visits; FT=Feeding time; GRF=Ground reaction force; KN=Kicks number; LWR=Leg weigh ratio; LBn=Lying bouts number; LBt=Lying bouts time; LT=Lying time; MF=Milking frequency; MD= Milking duration; MO=Order in which a cow enter to milking; MY=Milk yield; Spm=Speed measurement; ST=Stance time; SDRL=Standard deviation of weight rear legs; Sta=Standing time; StN=Steps number; StrL=Stride length; Tm=Tracking-up measurement; T&R angle=Touch and release angle; Var acc=Variance of acceleration
[c]: Percentage of utilization, calculated considering only references to automatic locomotion scoring systems found in complete review. 30 automatic locomotion scoring systems used in 244 articles
[d]: Mix approach is a combination of different approaches such as kinetic, kinematic and indirect approach

2.5. Traits and variables considered in locomotion scoring systems

2.5.1. Traits observed in manual locomotion scoring systems

MLSSs are based on the observation and judgement of several gait and posture traits. The review disclosed twelve gait, posture, or other traits used in 25 MLSSs (Tables 2.1 and 2.2). Gait traits focused on detecting alterations related to the limbs included: asymmetric gait (uneven gait) used in 17 MLSSs; reluctance to bear weight (also tenderness or affected

leg), used in 15; short steps, used in nine; abduction/adduction, used in six; tracking up (step overlap), used in five; and joint flexibility, used in two.

Posture traits are alterations in locomotion related to parts of the body other than the limbs, including arched back (also back or spine curvature), used in 14 MLSSs, and head bobbing (also head carriage) used in eight. Other locomotion traits focused on attributes that could not be classified in previous categories such as walking speed, used in six MLSSs; and difficulty in turning and difficulty in rising used in three and two MLSSs, respectively.

Cows presenting impaired locomotion do not always express all gait and posture traits described in MLSSs. Bach et al. (2007) and Thomsen et al. (2008) reported that not all cows presented an arched back when cows presented impaired locomotion, 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 must combine different gait or posture traits and decide which of them is more important to assign a locomotion score.

The importance that raters assign to individual gait and posture traits has been studied by estimating correlation coefficients between scores of specific gait and posture traits and the locomotion score. Borderas et al. (2008) and Chapinal et al. (2009) reported that asymmetric gait (range $r = 0.84 – 0.91$) and reluctance to bear weight (range $r = 0.88 – 0.90$) showed high correlations with locomotion score. Correlation coefficients ranging from 0.70 to 0.80 were estimated between scores for head bobbing, tracking up, joint flexibility, and locomotion score. Low to medium correlation coefficients ($r = 0.41 – 0.68$) were estimated between arched back and locomotion score and between abduction/adduction and locomotion score ($r = 0.32$). Van Nuffel et al. (2009) used a different approach based on the frequency of detection of ten gait, posture, and other traits assessed by 39 raters with different levels of experience. Asymmetric gait, reluctance to bear weight, arched back, and abduction/adduction had a significant effect ($p < 0.05$) when predicting locomotion score by a regression model.

In general, the reviewed articles show that some of the most used traits in MLSSs, such as asymmetric gait and reluctance to bear weight, are also the most associated with the final locomotion score assigned to a cow. Contradictory results for the importance assigned to individual trait, and especially arched back, indicates that raters give different importance to different traits based on their personal criteria.

2.5.2. Variables measured in automatic locomotion scoring systems

The kinetic ALSS first described by Rajkondawar et al. (2002), uses different ground reaction forces and stance time of individual limbs (Tables 2.3 and 2.4). The kinetic ALSS using four independent weighing units as sensors described by Pastell and Kujala (2007), measures the weight distribution among limbs when the cow is standing. Measured variables are weight ratio, standard deviation of weight in front and hind limbs, number of kicks, and number of steps (Tables 2.3 and 2.4) (Pastell et al., 2010; Chapinal and Tucker, 2012).

In the kinematic approach, the measured gait variables were asymmetry of step length, asymmetry of step time, asymmetry of step width, stance time, stride length and tracking up (Tables 2.3 and 2.4). Posture variables were related mostly to measurements of back-arching and included the radius of an imaginary circle fitted to the back-line of a cow (Poursaberi et al., 2010) or body movement pattern (Viazzi et al., 2013) (Tables 2.3 and 2.4). Acceleration variables included the variance of the forward, lateral-horizontal and vertical acceleration relative to the cow's leg while walking (Pastell et al., 2009).

Most used behaviour and production variables were, milk yield, used in four ALSSs; lying time, used in three; and number of lying bouts and standing time, used in two. All other behaviour and production variables were used in only one ALSS each (Tables 2.3 and 2.4).

In general, ALSSs using the kinetic and kinematic approach are based on locomotion analysis in a similar way as MLSSs. Kinetic variables may be related to traits such as reluctance to bear weight whereas, kinematic variables such as step length and body movement pattern may be considered equivalent to traits such as asymmetric gait and arched back, respectively. Since kinetic and kinematic ALSSs try to mimic MLSSs, the selection of the variables to be measured should be based on the importance assigned to individual traits used in MLSSs (discussed in Section 5.1). Thus, probably ALSSs based on the measurement of kinetic and kinematic of gait variables, should be more related to locomotion score than ALSSs based on posture kinematic variables.

Table 2.4. Abbreviation (Abb) and definition of gait, posture, behaviour and production variables used in automatic locomotion scoring systems.

Variables	Abb	Definitions[a]
Gait		
Asymmetry of step length	ASpL	Mean difference in step length between left and right hoof imprints
Asymmetry of step time	ASpT	Mean difference in step time between left and right hoof imprints
Asymmetry of step width	ASpW	Mean difference in step width between left and right hoof imprints
Asymmetry of stance time	AST	Mean difference in time that a hoof is on the ground between left and right hoof imprints
Ground reaction force	GRF	Force transmitted by the hoof to the ground while walking
Number of kicks	KN	Lifting of the leg when the weight decreased to less than 5 kg (in kinetic approach)
Leg weigh ratio	LWR	Ratio between lighter and heavier leg
Stance time	ST	Time during which a hoof is in contact with the floor
Standard deviation weight rear legs	SDRL	Standard deviation of weight of rear legs
Number of steps	StN	Lifting of the leg when the weight decreased to between 5 and 20 kg (in kinetic approach)
Stride length	StrL	Distance between two consecutive imprints of the same hoof
Tracking-up measurement	Tm	Distance between the position of the front foot and hind foot on the same body side on the floor in the subsequent step
Touch and release angle	T&R angle	Angle of the metacarpus and metatarsus bones with respect to a vertical line during stance phase of a hoof
Variance of acceleration	Var acc	Variance of forward, lateral-horizontal and vertical acceleration relative to cow's leg while walking
Posture		
Arched back measurement	ABm	Measurement of back curvature expressed as inverse of the radius of an imaginary circle fitted in back-line of the cow
Body movement pattern	BMP	Coefficient obtained by weighting different angles and distances in the cow's posture
Behaviour-Production		
Activity	Act	Activity indicator depends on the anatomical location of the accelerometer, e.g. on the neck or on the leg
Concentrate left over	CLO	Concentrate left in concentrate dispenser
Feed bunk visits	FV	Number of visits to the feed bunk
Feeding time	FT	Time spend in the feed bunk
Number of lying bouts	LBn	Number of lying bouts in a day
Duration of lying bout	LBt	Mean time of lying bouts
Lying time	LT	Mean time that a cow spend lying in a day
Milking frequency	MF	Number of visits to the milking robot in a voluntary milking system
Milking duration	MD	Time needed for milking
Milking Order	MO	Entering order for milking
Milk yield	MY	Daily milk production
Speed measurement	Spm	Lapse of time to cover a known distance
Standing time	Sta	Time spend in standing posture (still or moving)

[a] Based on definitions proposed by Ito et al. (2010), Kamphuis et al. (2013), Maertens et al. (2011), Pastell and Kujala (2007), Pastell et al. (2009), Rajkondawar et al. (2002) and Viazzi et al. (2013).

2.6. Types of scale

The 25 described MLSSs used two types of scale: continuous (six of 247 references, 2%), and ordinal (241 of 247 references, 98%). Commonly a low score indicated normal locomotion and a high score indicated extremely impaired locomotion, with the exception of the nine-level ICAR MLSS, which uses a "reverse scale" (Table 2.1).

A continuous scale is constructed by drawing a straight line, normally 100 mm, of which the endpoints are defined as the minimum and maximum values of the trait recorded (e.g. from normal locomotion to extremely impaired locomotion) (Paul-Dauphin et al., 1999). In a continuous scale, the rater marks on the line in the location believed to correspond best to the observed trait. The value assigned to the trait is equivalent to the distance (in mm) between an endpoint (normally the minimum value) and the mark by the rater.

Ordinal scales as used in different MLSSs have two (13 references, 5%), three (17 references, 7%), four (21 references, 9%), five (106 references, 43%), six (10 references, 4%) or nine (73 references, 30%) levels (Table 2.1). Each level of the scale includes a description of traits to be assessed and raters use this description as a guideline to assign a locomotion score to a cow. Preference for ordinal over continuous scales in MLSSs may be explained by the notion that ordinal scales are more easily taught and easier to use on farm (Engel et al., 2003; Tuyttens et al., 2009). In addition, the description of traits at each level of an ordinal scale may help to define a standardized method for locomotion scoring.

A cow was classified as lame when a defined threshold on the scale was exceeded. In most MLSSs, the threshold to classify a cow as lame was when the locomotion score exceeded the middle level of the scale (e.g. locomotion score ≥3 in five-level scales) (Winckler and Willen, 2001; Channon et al., 2009; Chapinal et al., 2009; Hoffman et al., 2013). An alternative approach to classify a cow as lame was when two of the five gait and posture traits scored ≥3 on a five-level scale (O'Callaghan et al., 2003). Van Nuffel et al. (2009) classified cows as mildly lame when a rater detected one of the ten gait and posture traits, and as lame when two or more traits were detected.

Several ALSSs use binary (e.g. lame/not lame) (Rajkondawar et al., 2006; Ito et al., 2010; Pastell et al., 2010) or three-level ordinal scales (e.g. not lame, mildly lame and severely lame) (Pluk et al., 2010; Maertens et al., 2011).

2.7. Agreement and reliability

Agreement and reliability are important indicators of consistency and reproducibility of a test (Martin and Bateson, 1993; Kottner et al., 2011). Agreement indicates the capability of raters using MLSSs to assign identical locomotion scores to a cow (Kottner et al., 2011).

Agreement is a characteristic of the quality of the test (de Vet et al., 2006). Reliability is the capability of raters using MLSSs to differentiate among levels (e.g. lame and not lame) (Kottner et al., 2011). Unlike agreement, reliability is not only an indicator of the quality of the test, but it is also highly dependent on the homogeneity of the population sample (de Vet et al., 2006) (e.g. populations with low lameness prevalence can be considered homogenous).

2.7.1. Statistics used for agreement and reliability

The only agreement statistic used in studies using MLSSs was percentage or proportion of agreement (PA). The most commonly used reliability statistics in MLSSs were kappa (κ) and weighted kappa (κw). Prevalence-adjusted bias-adjusted kappa (PABAK), Pearson (r), and Spearman (rs) correlation coefficients have also been used as expression of agreement or reliability in MLSSs (Table 2.5).

The PA is calculated by dividing the number of agreements by the total number of agreements and disagreements (Martin and Bateson, 1993). PA is commonly used because it is easy to calculate. However, reporting PA of a homogeneous population sample (e.g. low lameness prevalence) may be misleading, because PA will be representative only for the majority portion of the population sample (e.g. non-lame) (Kaufman and Rosenthal, 2009). Acceptance threshold indicating good PA estimates is commonly indicated around 75% (Burn and Weir, 2011).

The κ coefficient (Cohen, 1960) corrects PA for the possibility of agreements obtained by chance in categorical scales. Since MLSSs are ordinal scales, κ coefficient should not be used in multi-level MLSSs but only for the binary scale for lame or non-lame classification. The κ coefficient has been criticized for being affected by the prevalence of the measured characteristic (in this case, low lameness prevalence would result in a relatively low κ) (Sim and Wright, 2005; Burn and Weir, 2011). Many authors, however, indicate that the effect of prevalence is useful for a correct interpretation of κ coefficient as reliability indicator. A low κ coefficient indicates that raters presented high agreement in only one of the two levels, indicating the incapability of raters to differentiate among levels when a characteristic has low prevalence (Cicchetti and Feinstein, 1990; Vach, 2005; Kottner et al., 2011). The acceptance threshold for κ coefficient is usually set around 0.6 (Landis and Koch, 1977).

The κw coefficient (Cohen, 1968) is considered a suitable statistic of reliability estimation for multiple level ordinal scales because it introduces different weightings according to the magnitude of disagreement of the raters. Thus, a high κw coefficient indicates that rater disagreements are mainly due to one level difference, whereas differences for two or three levels are less common. A common critic to κw coefficient is that there is not a standard method to decide upon weights (Graham and Jackson, 1993). The acceptance level for κw coefficient is usually set around 0.6.

Another used statistic is PABAK which corrects for the effects of prevalence of the studied characteristic and rater bias in the κ coefficient (Byrt et al., 1993). Since PABAK is corrected for the effect of prevalence it cannot be considered a reliability statistic but it is an agreement statistic with a difficult interpretation. Therefore it should not be used according to Hoehler (2000).

A correlation coefficient describes the linear relationship or interdependence between two measures (Kirk, 2007). The principal criticism for Pearson and Spearman correlation coefficients is that they only indicate linear relationships. Therefore, Pearson and Spearman correlation coefficients should not be used as agreement or reliability estimates according to Gallagher et al. (2003) and Kottner et al. (2011).

2.7.2. Agreement and reliability in manual locomotion scoring systems

Although agreement and reliability of subjective tests are considered important, only 31 articles reported agreement or reliability. In none of the articles, there was a distinction made between the concepts of agreement and reliability. In many cases, the concept of agreement was used, however reliability statistics were reported (Bicalho et al., 2007; Thomsen et al., 2008; Danscher et al., 2009). From 31 articles, eight were not included in Table 2.5 because authors did not report whether or not the agreement or reliability estimates corresponded to the original scale or to transformation into a binary scale (lame or non-lame) (Espejo et al., 2006; Katsoulos and Christodoulopoulos, 2009; Eicher et al., 2013). Most articles had a different aim than evaluating agreement and reliability, thus agreement or reliability were reported only briefly and in many cases important facts for data interpretation were missing. For instance, the number of cows scored or the lameness prevalence was not reported, or it was not indicated if raters were allowed to comment on locomotion scores assigned. Most articles focused on reporting interrater comparisons (Table 2.5).

Agreement or reliability were reported for nine MLSSs (Table 2.5). Several studies aimed to evaluate the MLSS (continuous and ordinal scale) of Flower and Weary (2006). However, most studies used the inappropriate Pearson correlation coefficient (Table 2.5). The Pearson correlation coefficient was probably selected as indicator to make an easier comparison among continuous scales (Flower and Weary, 2006; Borderas et al., 2008). A better statistic to estimate reliability for continuous scales is the intra-class correlation coefficient (Kottner et al., 2011). No agreement estimation has been reported for the Flower and Weary (2006) MLSS (Table 2.5). No study was found that aimed to estimate agreement and reliability of the Sprecher et al. (1997) and DairyCo. (2007) MLSSs. Most articles reporting agreement or reliability using these MLSS did it briefly in the material and methods or results sections. Some of the best evaluated MLSSs for agreement and reliability are Manson and Leaver (1988) and Winckler and Willen (2001) MLSSs.

Table 2.5. Interrater and intrarater agreement and reliability for original locomotion score and a two-level scale (lame/non-lame) for different manual locomotion scoring systems (MLSSs) found in the literature. Agreement or reliability were expressed as percentage of agreement (P_A), kappa coefficient (κ), weighted kappa coefficient (κ_w), prevalence-adjusted bias-adjusted kappa (PABAK), Pearson correlation coefficient (r) and Spearman rank correlation coefficient (r_s).

MLSS	Statistic	Original score		Two levels		Citation
		Interrater	Intrarater	Interrater	Intrarater	
DairyCo., 2007	P_A	61.3 - 83.3		83.9 - 96.8		Barker et al., 2010
	P_A				90.6 - 100	Main et al., 2010
	P_A	67.2		90.5		Rutherford et al., 2009
	κ			0.67 - 0.93		Barker et al., 2010
	κ				0.81 - 1.00	Main et al., 2010
	κ_w	0.42 - 0.73				Rutherford et al., 2009
	PABAK			0.67 - 0.94		Rutherford et al., 2009
Flower and Weary, 2006 (Dis)[a]	κW		0.67			Yamamoto et al., 2013
	PABAK			0.83 - 0.93		Chapinal et al., 2013
	r	0.71 - 0.76				Bernardi et al., 2009
	r	0.88				Chapinal et al., 2010a
	r	0.83	0.87 - 0.92			Flower and Weary, 2006[c]
	r		0.88 - 0.99			Flower et al., 2008[c]
	r_s		0.76			Yamamoto et al., 2013
Flower and Weary, 2006 (Con)[d]	r	0.78				Chapinal et al., 2010a
	r	0.85	0.87 - 0.90			Flower and Weary, 2006[c]
Manson and Leaver, 1988	P_A	17.0 - 42.0	30.0	88.3	86.7	Channon et al., 2009
	P_A	25.0 - 47.0				Engel et al., 2003
	κ	0.05 - 0.27		0.79		Channon et al., 2009
	κ_w	0.80 - 0.85				Channon et al., 2009
O'Callaghan et al., 2003	P_A	37.0	56.0			O'Callaghan et al., 2003

Table 2.5, continuation

Sogstad et al., 2005	PABAK			0.49 - 0.80		Otten et al., 2013
Sprecher et al., 1997	P_A	83.0		96.0		Hoffman et al., 2013
	κ			0.72 - 0.80		Hoffman et al., 2013
	κ_w	0.30 - 0.40				Danscher et al., 2009
	κ_w	0.57 - 0.68				Hoffman et al., 2013
	PABAK			0.36 - 0.80	0.60 - 0.85	Thomsen and Baadsgaard, 2006
	PABAK			0.79		Hoffman et al., 2013
Thomsen et al., 2008	κ	0.01 - 0.54	0.30 - 0.68			Thomsen et al., 2008
	κ_w	0.24 - 0.68	0.38 - 0.78			Thomsen et al., 2008
Winkler and Willen, 2001	P_A	46.0 - 95.0				March et al., 2007
	P_A			91.0		Leach et al., 2009[d]
	P_A	63.0 - 74.0				Winkler and Willen, 2001
	κ			0.81		Leach et al., 2009[d]
	κ_w	0.46 - 0.48				Bicalho et al., 2007
	κ_w	0.41- 0.86				March et al., 2007
	PABAK	0.25 - 0.68		0.59 - 0.70		Brenninkmeyer et al., 2007
	PABAK	0.32 - 0.94		0.52 - 0.95		March et al., 2007
	r_s	0.55 - 0.89				March et al., 2007

[a] Manual locomotion scoring system in discrete scale

[b] Manual locomotion scoring system in continuous scale

[c] Repeatability originally expressed as coefficient of determination (R^2) and transformed to Pearson correlation coefficient by calculating the square root of the original value reported in the paper

[d] Modification of locomotion score proposed by Winckler and Willen, (2001) for cows in tie stall barns

However, the number of raters evaluating both MLSSs was relatively low ranging from two (March et al., 2007; Leach et al., 2009) to nine (Engel et al., 2003) raters. In addition, some articles, based their conclusions on inappropriate statistical indicators such as PABAK (Brenninkmeyer et al., 2007) or reported κ coefficient in multiple level MLSSs (Channon et al., 2009).

Lowest interrater agreement estimates were reported for the MLSS proposed by Manson and Leaver (1988) with PA ranging from 17 to 47% for the original nine-level scale. Agreement estimates for Manson and Leaver's (1988) MLSS are relatively low, which can be explained by the fact that a higher number of levels results in lower PA estimates. Each of the four- and five-level MLSSs showed a large range for interrater agreement estimates with PA ranging from 37% (O'Callaghan et al., 2003) to 95% (March et al., 2007) (Table 2.5). For two-level scales (lame/non-lame), PA estimates were ≥80%, exceeding the acceptance threshold of PA for interrater and intrarater agreement. The range across articles of PA estimates for two-level scales was large (from 83 to 97%; Table 2.5).

In general, agreement estimates showed relatively large ranges across articles or even within articles which may partly be due to the lack of a standard to perform MLSSs, as mentioned in Section 3, but probably raters had the largest effect (Engel et al., 2003; Channon et al., 2009). Training of raters is mentioned as the main factor affecting performance of raters (Kazdin, 1977). A rater is considered sufficiently proficient when the agreement estimates are above the acceptance threshold of the used statistical method (Martin and Bateson, 1993). There is no standard available, however, for training raters to perform locomotion scoring (March et al., 2007). Engel et al. (2003) reported that different raters performed differently, with some raters obtaining better agreement estimates while other performed worse after a short training. Improved agreement estimates of raters were also obtained as more cows were assessed (March et al., 2007). March et al. (2007) considered 300 cows as a sufficient number to score to reach the acceptance threshold for agreement and reliability using a five-level MLSS. Even after obtaining the acceptance threshold, raters should receive periodic training to avoid any "drift" which refers to the tendency of raters to change over time how they apply the definitions of a measurement (Kazdin, 1977).

As agreement, reliability estimates presented a large variation. Interrater reliability estimates for the original scale showed a range for κw from 0.24 to 0.86. Intrarater reliability expressed as κw ranged from 0.38 to 0.78 (Table 2.5). For two-level scales (lame or non-lame), interrater reliability estimates presented ranged for κ coefficient from 0.67 to 0.93 (Table 2.5). Intrarater reliability for two-level scales expressed as κ ranged from 0.81 to 1. Variability in reliability estimates may be explained, in part, by the level of training of raters. Thomsen et al. (2008) reported limited improvement in reliability estimates after training of experienced raters. However, prevalence of the studied

characteristic has an important effect on reliability estimates (especially κ coefficient). Thus, comparison of reliability estimates in different articles must be done taking into account the prevalence of the studied characteristic (de Vet et al., 2006).

As explained in Section 2.5, some traits have more importance than others for assigning a locomotion score to a cow; however, it is also important that individual traits present high agreement and reliability. In two articles, Pearson correlation coefficients were reported for gait and posture traits (Flower and Weary, 2006; Borderas et al., 2008). Estimates for r > 0.7 were for tracking up, head bob, arched back and reluctance to bear weight, whereas for asymmetric gait and joint flexion presented r was <0.7 (Flower and Weary, 2006). Slightly different results were reported by Borderas et al. (2008) where tracking up and joint flexion resulted in r < 0.7. Both articles, reported scores from only two raters and using Pearson correlation coefficient as agreement or reliability estimate. Further research is required in this topic using more raters and the correct statistics. Utilization of individual traits with high weights and high agreement and reliability is important to obtain consistent MLSSs.

2.8. Validation of locomotion scoring systems

The term validity refers to the meaning and usefulness of the conclusions that can be drawn from a test (Wainer and Braun, 1988). Validity is not a property of the test itself, but rather of the meaning of the test (Messick, 1995). In this regard, it is possible to draw different conclusions from the same test (e.g. performance of a test detecting lameness or hoof lesions). Validation, i.e. the process to assess validity, can be performed using several approaches and statistical analyses (Wainer and Braun, 1988; Franzen, 2000).

2.8.1. Validation of ALSSs for lameness detection

Validation of ALSSs is mainly performed using MLSSs as golden standard for lameness and calculating sensitivity (Se) and specificity (Sp). Furthermore, a ROC curve can be constructed as an additional measure of validity by calculating the area under the curve (AUC) (Hanley and McNeil, 1982).

The Sp, Se and AUC of several ALSSs for lameness detection are shown in Table 2.6. Most ALSSs had acceptable Sp (≥80%). ALSSs, however, had a large range for Se, from 39 to 90%. These results indicate that ALSSs are better at detecting non-lame cows than at detecting lame cows.

Although some ALSSs had high Se, Sp, and AUC estimates, these results must be interpreted with caution. In many cases, validation was performed on experimental farms under controlled conditions, with a small number of lame cows (Chapinal et al., 2010b;

Pastell et al., 2010; Poursaberi et al., 2010; Maertens et al., 2011). Therefore, Se, Sp and AUC may be overestimated.

The major concern for validation of ALSSs is the utilization of MLSSs as golden standard. The agreement and reliability of the rater(s) performing locomotion scoring has an important effect in the definition of a lameness case and thus on the validity of ALSSs.

2.8.2. Validation of MLSSs and ALSSs for hoof lesion detection

MLSSs and ALSSs can be used for prevention and management of hoof lesions. Using a nine-level MLSSs and a threshold of 3.5 to detect sole ulcers, Se was 54%, and Sp was 70% (Chapinal et al., 2009). Acceptable AUC estimates ranging from 0.75 to 0.84 were reported for kinetic ALSSs described by Rajkondawar et al. (2002) for hoof lesion detection (Rajkondawar et al., 2006). For the ALSS described by Pastell and Kujala (2007), AUC was 0.71 using sole haemorrhage as a reference, and 0.87 using sole ulcer as a reference (Pastell et al., 2010).

A comparison between a five-level MLSS and the ALSSs described by Rajkondawar et al. (2002) for their capability of detecting painful lesions (defined as limb retraction when digital pressure was applied on the lesion) was performed under practical farm conditions (Bicalho et al., 2007). Using a threshold of 3 to classify a cow as lame, MLSS had a higher Se and slightly lower Sp than ALSS (Se = 67% for MLSS and 33% for ALSS; Sp = 84% MLSS and 90% for ALSS). The MLSS also presented better AUC than ALSS (0.77 vs. 0.62) for painful lesion detection (Bicalho et al., 2007).

Limited capability of MLSSs and ALSSs to detect hoof lesions might be because locomotion seems to be affected only by certain types of hoof lesions, mainly sole ulcers (Whay et al., 1997; Flower and Weary, 2006; Chapinal et al., 2009), severe cases of digital dermatitis (Frankena et al., 2009), and double sole and inter-digital purulent inflammation (Tadich et al., 2010). Other common hoof lesions, such as white line disease and sole haemorrhage had no effect on locomotion (Flower and Weary, 2006; Chapinal et al., 2009). The limited available literature on this topic only associates impaired locomotion and hoof lesions and do not consider other possible causes such as acute laminitis (Nordlund et al., 2004; Thoefner et al., 2004), hock lesions (Rutherford et al., 2008) or other traumatic limb injuries.

Although validity of MLSSs and ALSSs seems limited, it should be noted that results of most studies cited in this section were single measurements. In this regard, there is a need for long-term studies aiming to evaluate the practical utility of MLSSs and ALSSs for preventing and managing different types of hoof lesions. These studies should also aim to compare MLSSs and ALSSs with different methods for detection and control of hoof lesions.

Table 2.6. Sensitivity (Se), Specificity (Sp) and area under the curve (AUC) of automatic locomotion scoring systems (ALSSs) for lameness detection using manual locomotion scoring systems (MLSS) as reference.

ALSSs	Reference		Measure			Citation
	MLSS[a]	Lame[b]	Se	Sp	AUC	
Kinetic approach						
Rajkondawar et al., 2002	Spr	LS ≥ 3			0.63 - 0.73	Rajkondawar et al., 2006 [c]
	Spr	LS ≥ 3	51.9	88.4		Liu et al., 2011
Pastell and Kujala, 2007	F&W	LS ≥ 3			0.71	Pastell et al., 2010
	F&W	LS ≥ 3.5			0.88	Pastell et al., 2010
	F&W	LS ≥ 3			0.69 - 0.71	Chapinal et al., 2010b [d]
	F&W	LS ≥ 3			0.67	Chapinal and Tucker, 2012
Kinematic approach						
Viazzi et al., 2013	F&W	LS ≥ 3	76	91		Viazzi et al., 2013
Maertens et al., 2011	VN	LS 3 lev	76 - 90			Maertens et al., 2011 [e]
Indirect approach						
Ito et al., 2010	F&W	LS ≥ 3			0.64 - 0.65	Chapinal et al., 2010b [f]
	F&W	LS ≥ 4	39.1 - 56.5	72.8 - 96.4		Ito et al., 2010 [f]
	F&W	LS ≥ 3	72.0	81.0		Alsaaod et al., 2012
Kamphuis et al., 2013	Spr	LS ≥ 3	40.1 - 56.8	80 - 90	0.75	Kamphuis et al., 2012[g]
de Mol et al., 2013	W&W	LS ≥ 3	85.5	89.9		de Mol et al., 2013
Chapinal et al., 2010b	F&W	LS > 3			0.83	Chapinal et al., 2010b

[a] F&W=Flower and Weary, (2006); Spr=Sprecher et al., (1997); VN=Van Nuffel et al., (2009); W&W=Winckler and Willen, (2001).
[b] LS ≥ n: Threshold level at which a cow is considered lame; LS 3 lev: Locomotion is classified as, not lame, mildly lame or lame.
[c] Range indicate AUC values calculated using data from 1, 2 or 3 days of observation.
[d] Range of values indicate AUC for lameness detection using individual kinetic variables, leg weight ratio of rear limbs and standard deviation of the weight of front and hind limbs.
[e] Range of values are true positive detection rate obtained for each of the three levels used for locomotion scoring.
[f] Range of values indicate Se, Sp and AUC for lameness detection using individual behavior variables, daily lying time and lying bout duration.
[g] Range of values indicate sensitivity values when specificity if set at 80% and 90%.

2.8.3. Validation of locomotion scoring systems by pain assessment

It is assumed that cows change their way of walking to relieve pain (Flower and Weary, 2009). Thus, impaired locomotion is considered as the indicator of an underlying problem that induces pain (Flower and Weary, 2009).

An approach to assess pain was to apply noxious stimuli (e.g. thermal stimulus) to induce a response from the animal (e.g. limb retraction) (Gagliese and Melzack, 2000). The relationship was studied between locomotion score (performed with MLSSs) assigned to cows and the amount of pressure required to produce limb retraction when the pressure was applied to the dorsal aspect of the metatarsus (Whay et al., 1997) or to hooves (Dyer et al., 2007). Cows with higher locomotion scores required, on average, less pressure to initiate the response of limb retraction than cows with lower locomotion scores (Whay et al., 1997; Dyer et al., 2007). Cows with higher locomotion scores, therefore, would be more

likely to experience pain than those with lower locomotion scores (Whay et al., 1997; Dyer et al., 2007). Dyer et al. (2007), however, reported that 37% out of 262 cows did not have a locomotion score higher than 2 on a five-level MLSS.

A second approach to assess pain assumed that the use of analgesics or anaesthetics would improve the locomotion score of cows. Small, but significant improvements in locomotion score, expressed as a decrease of 0.3 (Rushen et al., 2007) and 0.25 (Flower et al., 2008) locomotion score points, were found in lame cows after injection of lidocaine (Rushen et al., 2007) and ketoprofen (Flower et al., 2008). On the other hand, a combination of hoof trimming and analgesia (flunixin meglumine) did not have an effect on the locomotion score (Chapinal et al., 2010c). Analgesics and anaesthetics also have been used to validate the ALSS described by Pastell and Kujala (2007). Rushen et al. (2007) reported that cows bear more weight on lame limbs after an injection with an anaesthetic. In addition, the injection of ketoprofen decreased the standard deviation of weight applied to rear legs in lame cows in lame cows by 18% and in non-lame cows by 12% (Chapinal et al., 2010b) (Table 2.4). Finally, a combination of hoof trimming and analgesia did not affect any measure of weight distribution in lame and not lame cows (Chapinal et al., 2010c) (Table 2.3).

Both manual and automatic locomotion scoring systems presented significant changes after the application of analgesics or anaesthetics. In case of MLSSs changes in locomotion scores should be interpreted with caution, because the statistical analysis was done using methodology more suitable for continuous data instead of ordinal data. Thus, the fact that locomotion score decreased with less than 0.5 score point is meaningless because raters tend to disagree at least one point of score (Winckler and Willen, 2001; O'Callaghan et al., 2003). Result obtained by ALSSs indicates that weight distribution over limbs might be a promising approach for assessment of pain-in-limbs in cows. However, agreement and reliability of the ALSSs need to be evaluated to determine the usefulness of the system for pain assessment. The limited validity of MLSSs and ALSSs must also be interpreted taking into account that the methodologies for pain assessment in animals are limited.

Better validation of MLSSs and ALSSs for pain assessment may be performed if more reliable methods for pain assessment are developed.

2.9. General discussion and conclusions

In conclusion, there are many different types of manual (MLSSs) and automatic (ALSSs) locomotion scoring systems. The most used gait and posture traits in MLSSs were asymmetric gait, reluctance to bear weight, short strides, arched back, and head bobbing. A five-level, ordinal scale was used most often. Lameness classification of cows depends on the established threshold of the scale, which was commonly decided to be the middle level.

We found 15 ALSSs that could be assigned to three different approaches: the kinetic, the kinematic and the indirect, each using sensors such as force plates, weighing units, cameras, pressure sensitive walkways and accelerometers. Kinetic and kinematic ALSSs try to mimic MLSS by measuring gait and posture variables and classifying cows in a scale with three- or two-levels. Indirect approaches use different sensors and variables available in common farming routine (e.g. milk metres, accelerometers, scales). ALSSs using the indirect approach, however, are unspecific since different illnesses may affect the same variables.

Agreement and reliability are important indicators of consistency and reproducibility of MLSSs and ALSSs. Agreement and reliability are different concepts that are often used interchangeably. Confusion in concepts of agreement and reliability leads to an incorrect interpretation of appropriate statistics and to the utilization of inappropriate statistics, such as PABAK, Pearson and Spearman correlation coefficients.

Agreement and reliability in locomotion scoring systems is an underestimated topic in scientific literature. Some of the most used MLSSs have not been properly evaluated for agreement and reliability mainly because of the use of incorrect statistics or a relatively low number of raters. Agreement presented large variability for the original four-, five- or nine-level MLSS. Some of the main factors affecting agreement are probably level of training of raters and the number of levels of the scale used. Like agreement, reliability also presented large variation. An extra factor affecting reliability is homogeneity (e.g. low lameness prevalence) of the sample population. No data for agreement and reliability of ALSSs was found.

Lameness detection is the main purpose of using MLSSs and ALSSs. Several ALSSs use MLSSs as reference for model calibration and validation. However, variable agreement and reliability of MLSSs make a clear definition of a lameness case difficult, which affects the validity of ALSSs.

MLSSs and ALSSs presented limited capability of detecting cows with hoof lesions. Other possible reasons for impaired locomotion (e.g. hock lesions or other limb injuries) have not been considered. Associating MLSSs and ALSSs to indicators of pain (noxious stimuli in limbs and use of analgesics or anaesthetics) showed contradicting and limited results. However, limited current methods for pain assessment in animals make it difficult to establish a better association between impaired locomotion and pain.

Limited validity of MLSSs and ALSSs for hoof lesions and pain assessment may be explained by various factors affecting locomotion, such as material of the walking surface (Telezhenko and Bergsten, 2005; Flower et al., 2007; Haufe et al., 2009); anatomical conformation of cows (Boettcher et al., 1998); parity (Chapinal et al., 2009); breed (Baird

et al., 2009); hoof trimming (Chapinal et al., 2010a); and degree of udder distension (Flower et al., 2006).

The utilization of MLSSs and ALSSs should aim to the prevention, detection and efficient management of conditions that induce impaired locomotion. Long-term studies comparing MLSSs and ALSSs with various strategies aiming to detect and control unfavourable conditions leading to impaired locomotion are required to determine the usefulness of MLSSs and ALSSs for securing optimal production and animal welfare in practice.

Acknowledgments

This study is part of the Marie Curie Initial Training Network BioBusiness project (FP7-PEOPLE-ITN-2008). The authors are very grateful to Jos Metz and the reviewers for their valuable comments.

References

Alsaaod, M., C. Romer, J. Kleinmanns, K. Hendriksen, S. Rose-Meierhofer, L. Plumer, and W. Buscher. 2012. Electronic detection of lameness in dairy cows through measuring pedometric activity and lying behavior. Appl. Anim. Behav. Sci. 142:134-141.

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.

Aoki, Y., M. Kamo, H. Kawamoto, J. G. Zhang, and A. Yamada. 2006. Changes in walking parameters of milking cows after hoof trimming. Anim. Sci. J. 77:103-109.

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.

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.

Baird, L. G., N. E. O'Connell, M. A. McCoy, T. W. J. Keady, and D. J. Kilpatrick. 2009. Effects of breed and production system on lameness parameters in dairy cattle. J. Dairy Sci. 92:2174-2182.

Barkema, H. W., J. D. Westrik, K. A. S. Vankeulen, Y. H. Schukken, and A. Brand. 1994. The effects of lameness on reproductive-performance, milk-production and culling in Dutch dairy farms. Prev. Vet. Med. 20:249-259.

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.

Bernardi, F., J. Fregonesi, C. Winckler, D.M. Veira, M.A.G. von Keyserlingk, and D.M. Weary. 2009. The stall-design paradox: Neck rails increase lameness but improve udder and stall hygiene. J. Dairy Sci. 92: 3074-3080.

Bicalho, R. C., S. H. Cheong, G. Cramer, and C. L. Guard. 2007. Association between a visual and an automated locomotion score in lactating holstein cows. J. Dairy Sci. 90:3294-3300.

Blackie, N., J. Amory, E. Bleach, and J Scaife. 2011a. The effect of lameness on lying behaviour of zero grazed Holstein dairy cattle. Appl. Anim. Behav. Sci. 134: 85-91.

Blackie, N., E. Bleach, J. Amory, and J., Scaife. 2011b. Impact of lameness on gait characteristics and lying behaviour of zero grazed dairy cattle in early lactation. Appl. Anim. Behav. Sci. 129: 67-73.

Blackie, N., E. C. L. Bleach, J. R. Amory, and J. R. Scaife. 2013. Associations between locomotion score and kinematic measures in dairy cows with varying hoof lesion types. J. Dairy Sci. 96:3564-3572.

Boettcher, P. J., J. C. M. Dekkers, L. D. Warnick, and S. J. Wells. 1998. Genetic analysis of clinical lameness in dairy cattle. J. Dairy Sci. 81:1148-1156.

Booth, C. J., L. D. Warnick, Y. T. Grohn, D. O. Maizon, C. L. Guard, and D. Janssen. 2004. Effect of lameness on culling in dairy cows. J. Dairy Sci. 87:4115-4122.

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.

Brenninkmeyer, C., S. Dippel, S. March, J. Brinkmann, C. Winckler, and U. Knierim. 2007. Reliability of a subjective lameness scoring system for dairy cows. Anim. Welfare 16:127-129.

Breuer, K., P.H. Hemsworth, J.L. Barnett, L.R. Matthews, andG.J. Coleman. 2000. Behavioural response to humans and the productivity of commercial dairy cows. Appl. Anim. Behav. Sci. 66: 273-288.

Bruijnis, M. R. N., H. Hogeveen, and E. N. Stassen. 2010. Assessing economic consequences of foot disorders in dairy cattle using a dynamic stochastic simulation model. J. Dairy Sci. 93:2419-2432.

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.

Byrt, T., J. Bishop, and J. B. Carlin. 1993. Bias, prevalence and kappa. J. Clin. Epidemiol. 46:423-429.

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. and C. B. Tucker. 2012. Validation of an automated method to count steps while cows stand on a weighing platform and its application as a measure to detect lameness. J. Dairy Sci. 95:6523-6528.

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.

Chapinal, N., A. M. de Passille, and J. Rushen. 2010a. Correlated changes in behavioral indicators of lameness in dairy cows following hoof trimming. J. Dairy Sci. 93:5758-5763.

Chapinal, N., A. M. de Passille, J. Rushen, and S. Wagner. 2010b. Automated methods for detecting lameness and measuring analgesia in dairy cattle. J. Dairy Sci. 93:2007-2013.

Chapinal, N., A. M. de Passille, J. Rushen, and S. A. Wagner. 2010c. Effect of analgesia during hoof trimming on gait, weight distribution, and activity of dairy cattle. J. Dairy Sci. 93:3039-3046.

Chapinal, N., M.A.G. von Keyserlingk, R.L.A. Cerri, K. Ito, S.J. LeBlanc, and D.M. Weary, 2013. Short communication: Herd-level reproductive performance and its relationship with lameness and leg injuries in freestall dairy herds in the northeastern United States. J. Dairy Sci. 96: 7066-7072.

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. 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20:37-46.

Cohen, J. 1968. Weighted Kappa - nominal scale agreement with provision for scaled disagreement or partial credit. Psychol. Bull. 70:213-220.

Cook, N.B., 2003. Prevalence of lameness among dairy cattle in Wisconsin as a function of housing type and stall surface. J. Am. Vet. Med. Assoc. 223: 1324-1328.

DairyCo. 2007. DairyCo mobility score. Access date: May 10, 2011.Web page:http://www.dairyco.org.uk/.

Danscher, A. M., J. M. D. Enemark, E. Telezhenko, N. Capion, C. T. Ekstrom, and M. B. Thoefner. 2009. Oligofructose overload induces lameness in cattle. J. Dairy Sci. 92:607-616.

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.

DeVries, T. J., J. A. Deming, J. Rodenburg, G. Seguin, K. E. Leslie, and H. W. Barkema. 2011. Association of standing and lying behavior patterns and incidence of intramammary infection in dairy cows milked with an automatic milking system. J. Dairy Sci. 94:3845-3855.

Dyer, R. M., N. K. Neerchal, U. Tasch, Y. Wu, P. Dyer, and P. G. Rajkondawar. 2007. Objective determination of claw pain and its relationship to limb locomotion score in dairy cattle. J. Dairy Sci. 90:4592-4602.

Eicher, S. D., D. C. Lay, J. D. Arthington, and M. M. Schutz. 2013. Effects of rubber flooring during the first 2 lactations on production, locomotion, hoof health, immune functions, and stress. J. Dairy Sci. 96:3639-3651.

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.

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

Fitzgerald, T., B.W. Norton, R. Elliott, , H. Podlich and O.L. Svendsen. 2000. The influence of long-term supplementation with biotin on the prevention of lameness in pasture fed dairy cows. J. Dairy Sci. 83: 338-344.

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.

Flower, F. C., D. J. Sanderson, and D. M. Weary. 2005. Hoof pathologies influence kinematic measures of dairy cow gait. J. Dairy Sci. 88:3166-3173.

Flower, F. C., D. J. Sanderson, and D. M. Weary. 2006. Effects of milking on dairy cow gait. J. Dairy Sci. 89:2084-2089.

Flower, F. C., A. M. de Passille, D. M. Weary, D. J. Sanderson, and J. Rushen. 2007. Softer, higher-friction flooring improves gait of cows with and without sole ulcers. J. Dairy Sci. 90:1235-1242.

Flower, F. C., M. Sedlbauer, E. Carter, M. A. G. von Keyserlingk, D. J. Sanderson, and D. M. Weary. 2008. Analgesics improve the gait of lame dairy cattle. J. Dairy Sci. 91:3010-3014.

Frankena, K., J. Somers, W. G. P. Schouten, J. V. van Stek, J. H. M. Metz, E. N. Stassen, and E. A. M. Graat. 2009. The effect of digital lesions and floor type on locomotion score in Dutch dairy cows. Prev. Vet. Med. 88:150-157.

Franzen, M. D. 2000. Reliability and validity in neuropsychilogical assessment. Kluwer Academic/Plenum Publisher, New York.

Gagliese, L. and R. Melzack. 2000. Age differences in nociception and pain behaviours in the rat. Neurosci. Biobehav. Rev. 24:843-854.

Gallagher, A. G., E. M. Ritter, and R. M. Satava. 2003. Fundamental principles of validation, and reliability: rigorous science for the assessment of surgical education and training. Surg. Endosc. 17:1525-1529.

Garbarino, E. J., J. A. Hernandez, J. K. Shearer, C. A. Risco, and W. W. Thatcher. 2004. Effect of lameness on ovarian activity in postpartum Holstein cows. J. Dairy Sci. 87:4123-4131.

Goldhawk, C., N. Chapinal, D. M. Veira, D. M. Weary, and M. A. G. von Keyserlingk. 2009. Prepartum feeding behavior is an early indicator of subclinical ketosis. J. Dairy Sci. 92:4971-4977.

Graham, P. and R. Jackson. 1993. The analysis of ordinal agreement data - beyond weighted kappa. J. Clin. Epidemiol. 46:1055-1062.

Green, L. E., V. J. Hedges, Y. H. Schukken, R. W. Blowey, and A. J. Packington. 2002. The impact of clinical lameness on the milk yield of dairy cows. J. Dairy Sci. 85:2250-2256.

Groehn, J.A., J.B. Kaneene, and D. Foster. 1992. Risk-factors associated with lameness in lactating dairy-cattle in Michigan. Prev. Vet. Med. 14: 77-85.

Hanley, J. A. and B. J. McNeil. 1982. The meaning and use of the area under the curve a receiver operating characteristic (ROC) curve. Radiology 143:29 - 36.

Haufe, H. C., L. Gygax, B. Steiner, K. Friedli, M. Stauffacher, and B. Wechsler. 2009. Influence of floor type in the walking area of cubicle housing systems on the behaviour of dairy cows. Appl. Anim. Behav. Sci. 116:21-27.

Hirst, W. M., A. M. Le Fevre, D. N. Logue, J. E. Offer, S. J. Chaplin, R. D. Murray, W. R. Ward, and N. P. French. 2002. A systematic compilation and classification of the literature on lameness in cattle. Vet. J. 164:7-19.

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. R. Wenz, and J. Vanegas. 2013. Comparison of modeled sampling strategies for estimation of dairy herd lameness prevalence and cow-level variables associated with lameness. J. Dairy Sci. 96:5746-5755.

Ito, K., M. A. G. von Keyserlingk, S. J. LeBlanc, and D. M. Weary. 2010. Lying behavior as an indicator of lameness in dairy cows. J. Dairy Sci. 93:3553-3560.

Kamphuis, C., E. Frank, J. K. Burke, G. A. Verkerk, and J. G. Jago. 2013. Applying additive logistic regression to data derived from sensors monitoring behavioral and physiological characteristics of dairy cows to detect lameness. J. Dairy Sci. 96:7043-7053.

Katsoulos, P. D. and G. Christodoulopoulos. 2009. Prevalence of lameness and of associated claw disorders in Greek dairy cattle industry. Livest. Sci. 122:354-358.

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.

Kestin, S.C., T.G. Knowles, A.E. Tinch, N.G. and Gregory. 1992. Prevalence of leg weakness in broiler-chickens and its relationship with genotype. Vet. Rec. 131: 190-194.

Kirk, R. E. 2007. Correlation. Pages 123 - 151 in Statistics an introduction. Thomson Wadsworth, Belmont, U.S.A.

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.

Laursen, M. V., D. Boelling, and T. Mark. 2009. Genetic parameters for claw and leg health, foot and leg conformation, and locomotion in Danish Holsteins. J. Dairy Sci. 92:1770-1777.

Leach, K. A., S. Dippel, J. Huber, S. March, C. Winckler, and H. R. Whay. 2009. Assessing lameness in cows kept in tie-stalls. J. Dairy Sci. 92:1567-1574.

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. Biosyst. Eng. 110:29-39.

Main, D.C.J., Z.E. Barker, K.A. Leach, N.J. Bell, H.R. Whay, and W.J. Browne. 2010. Sampling strategies for monitoring lameness in dairy cattle. J. Dairy Sci. 93: 1970-1978.

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. Welfare 16:131-133.

Martin, P. and P. Bateson. 1993. Measuring behaviour: an introductory guide. 2nd edition ed. Cambridge University Press, Cambridge.

Messick, S. 1995. Validity of psycological assessment. Am. Psycol 50:741-749.

Nordlund, K. V., N. B. Cook, and G. R. Oetzel. 2004. Investigation Strategies for Laminitis Problem Herds. J. Dairy Sci. 87:E27-E35.

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. Welfare 12:605-610.

Offinger, J., S. Herdtweck, A. Rizk, A. Starke, M. Heppelmann, H. Meyer, S. Janssen, M. Beyerbach, and J. Rehage, 2013. Postoperative analgesic efficacy of meloxicam in lame dairy cows undergoing resection of the distal interphalangeal joint. J. Dairy Sci. 96: 866-876.

Onyiro, O. M., L. J. Andrews, and S. Brotherstone. 2008. Genetic parameters for digital dermatitis and correlations with locomotion, production, fertility traits, and longevity in Holstein-Friesian dairy cows. J. Dairy Sci. 91:4037-4046.

Otten, N.D., N. Toft, H. Houe, P.T. Thomsen, J.T. Sorensen. 2013. Adjusting for multiple clinical observers in an unbalanced study design using latent class models of true within-herd lameness prevalence in Danish dairy herds. Prev. Vet. Med. 112: 348-354.

Pastell, M. E. and M. Kujala. 2007. A probabilistic neural network model for lameness detection. J. Dairy Sci. 90:2283-2292.

Pastell, M., M. Hautala, V. Poikalainen, J. Praks, I. Veermae, M. Kujala, and J. Ahokas. 2008. Automatic observation of cow leg health using load sensors. Comput. Electron. Agr. 62:48-53.

Pastell, M., J. Tiusanen, M. Hakojarvi, and L. Hanninen. 2009. A wireless accelerometer system with wavelet analysis for assessing lameness in cattle. Biosyst. Eng. 104:545-551.

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.

Paul-Dauphin, A., F. Guillemin, J. M. Virion, and S. Briancon. 1999. Bias and precision in visual analogue scales: A randomized controlled trial. Am. J. Epidemiol. 150:1117-1127.

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.

Pluk, A., C. Bahr, A. Poursaberi, W. Maertens, A. van Nuffel, D. Berckmans. 2012. Automatic measurement of touch and release angles of the fetlock joint for lameness detection in dairy cattle using vision techniques. J. Dairy Sci. 95: 1738-1748.

Poursaberi, A., C. Bahr, A. Pluk, A. Van Nuffel, and D. Berckmans. 2010. Real-time automatic lameness detection based on back posture extraction in dairy cattle: Shape analysis of cow with image processing techniques. Comput. Electron. Agr. 74:110-119.

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. 2008. Hock injury prevalence and associated risk factors on organic and nonorganic dairy farms in the United Kingdom. J. Dairy Sci. 91:2265-2274.

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.

Scott, G. B. 1988. Lameness and pregnancy in Friesian dairy-cows. Br. Vet. J. 144:273-281.

Sim, J. and C. C. Wright. 2005. The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. Phys. Ther. 85:257-268.

Sogstad, A.M., T. Fjeldaas, and O. Osteras. 2005. Lameness and claw lesions of the Norwegian red dairy cattle housed in free stalls in relation to environment, parity and stage of lactation. Acta Vet. Scand. 46, 203-217.

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. Agr. 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.

Tadich, N., E. Flor, and L. Green. 2010. Associations between hoof lesions and locomotion score in 1098 unsound dairy cows. Vet. J. 184:60-65.

Telezhenko, E. and C. Bergsten. 2005. Influence of floor type on the locomotion of dairy cows. Appl. Anim. Behav. Sci. 93:183-197.

Thoefner, M. B., C. C. Pollitt, A. W. van Eps, G. J. Milinovich, D. J. Trott, O. Wattle, and P. H. Andersen. 2004. Acute Bovine Laminitis: A New Induction Model Using Alimentary Oligofructose Overload. J. Dairy Sci. 87:2932-2940.

Thomsen, P.T., and N.P. Baadsgaard, 2006. Intra- and inter-observer agreement of a protocol for clinical examination of dairy cows. Prev. Vet. Med. 75: 133-139.

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.

Thomsen, P.T., 2009. Rapid screening method for lameness in dairy cows. Vet. Rec. 164: 689-690.

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. Welfare 18:399-405.

Vach, W. 2005. The dependence of Cohen's kappa on the prevalence does not matter. J. Clin. Epidemiol. 58:655-661.

Van der Waaij, E. H., M. Holzhauer, E. Ellen, C. Kamphuis, and G. de Jong. 2005. Genetic parameters for claw disorders in dutch dairy cattle and correlations with conformation traits. J. Dairy Sci. 88:3672-3678.

Van Hertem, T., V. Alchanatis, A. Antler, E. Maltz, I. Halachmi, A. Schlageter-Tello, C. Lokhorst, S. Viazzi, C. E. B. Romanini, A. Pluk, C. Bahr, and D. Berckmans. 2013. Comparison of segmentation algorithms for cow contour extraction from natural barn background in side view images. Comput. Electron. Agr. 91:65-74.

Van Nuffel, A., M. Sprenger, F. A. M. Tuyttens, and W. Maertens. 2009. Cow gait scores and kinematic gait data: can people see gait irregularities? Anim. Welfare 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.

Vokey, F.J., C.L. Guard, H.N. Erb, and D.M., Galton. 2001. Effects of alley and stall surfaces on indices of claw and leg health in dairy cattle housed in a free-stall barn. J. Dairy Sci. 84: 2686-2699.

Wainer, H. and H. I. Braun. 1988. Test Validity. First ed. Lawrence Erlbaum Associates, Inc, New Jersey.

Walker, S. L., R. F. Smith, D. N. Jones, J. E. Routly, M. J. Morris, and H. Dobson. 2010. The Effect of a Chronic Stressor, Lameness, on Detailed Sexual Behaviour and Hormonal Profiles in Milk and Plasma of Dairy Cattle. Reprod. Domest. Anim. 45:109-117.

Walker, S. L., R. F. Smith, J. E. Routly, D. N. Jones, M. J. Morris, and H. Dobson. 2008. Lameness, Activity Time-Budgets, and Estrus Expression in Dairy Cattle. J. Dairy Sci. 91:4552-4559.

Warnick, L. D., D. Janssen, C. L. Guard, and Y. T. Grohn. 2001. The effect of lameness on milk production in dairy cows. J. Dairy Sci. 84:1988-1997.

Welfare Quality. 2009. Assessment Protocol for Cattle. in Welfare Quality Consortium. Lelystad, The Netherlands.

Wells, S.J., A.M. Trent, W.E. Marsh, P.G. Mcgovern, and R.A. Robinson. 1993. Individual cow risk-factors for clinical lameness in lactating dairy-cows. Prev. Vet. Med. 17: 95-109.

Whay, H. 2002. Locomotion scoring and lameness detection in dairy cattle. In Practice 24:444-449.

Whay, H. R., A. E. Waterman, and A. J. F. Webster. 1997. Associations between locomotion, claw lesions and nociceptive threshold in dairy heifers during the peri-partum period. Vet. J. 154:155-161.

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 Agr. Scand. a-An. 30:103-107.

Yamamoto, S., K. Ito, K. Suzuki, Y. Matsushima, I. Watanabe, Y. Watanabe, K. Abiko, T. Kamada, and K. Sato, 2013. Kinematic gait analysis and lactation performance in dairy cows fed a diet supplemented with zinc, manganese, copper and cobalt. Anim. Sci. J. 85:330-335

Chapter 3

Comparison of locomotion scoring for dairy cows by experienced and inexperienced raters using live or video observation methods

Published in Animal Welfare (2015) 24: 69-79

A. Schlageter Tello, 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

Abstract

Lameness is considered a major problem in dairy production. Lameness is commonly detected with locomotion scores assigned to cows under farm conditions, but raters are often trained and assessed for reliability and agreement by using video recordings. The aim of this research was to evaluate intrarater and interrater reliability and agreement of experienced and inexperienced raters for locomotion scoring performed live and from video, and to calculate the influence of raters and the method of observation (live or video) on the probability of classifying a cow as lame. Using a five-level locomotion score, cows were scored twice live and twice from video by three experienced and two inexperienced raters for three weeks. Every week different cows were scored. Intrarater and interrater reliability (expressed as weighted kappa, κw) and agreement (expressed as percentage of agreement, PA) for live/live, live/video and video/video comparisons were determined. A logistic regression was performed to estimate the influence of the rater and method of observation on the probability of classifying a cow as lame in live and video observation. Experienced raters had higher values for intrarater reliability and agreement for video/video than for live/live and live/video comparison. Inexperienced raters, however, did not differ for intrarater and interrater reliability and agreement for live/live, live/video and video/video comparisons. The logistic regression indicated that raters were responsible for the main effect and the method of observation (live or from video) had a minor effect on the probability for classifying a cow as lame (locomotion score ≥ 3). In conclusion, under the present experimental conditions experienced raters performed better than unexperienced raters when locomotion scoring was done from video. Since raters are the most important factors influencing the probability of classifying a cow as lame, video observation seems to be an acceptable method for locomotion scoring and lameness assessment in dairy cows.

3.1. Introduction

Lameness is considered a major problem in dairy production (Bruijnis et al., 2010). Mean prevalence of lameness in dairy herds during the last decade was 33% in Austria and Germany (Dippel et al., 2009b; a); 37% in England and Wales (Barker et al., 2010); and ranged from 21% to 55% in the USA (Cook, 2003; Espejo et al., 2006; von Keyserlingk et al., 2012). Lameness is associated with reduced milk yield (Warnick et al., 2001; Green et al., 2002; Archer et al., 2010), impaired reproductive performance (Garbarino et al., 2004; Walker et al., 2008; Walker et al., 2010), increased risk of culling (Barkema et al., 1994; Booth et al., 2004), and impaired animal welfare (Nordlund et al., 2004; Rushen et al., 2007). These effects result generally in increased production costs (Bruijnis et al., 2010; Cha et al., 2010).

Lameness is commonly detected with locomotion scoring methods. Locomotion scoring can be done quickly on-site, requires no technical equipment, and can be applied easily to a large number of animals (Whay, 2002; Flower and Weary, 2009; Ito et al., 2010). On the other hand, locomotion scoring is sensitive to variation between and within raters (Engel et al., 2003; O'Callaghan et al., 2003; Thomsen et al., 2008; Channon et al., 2009). The quality of subjective measurements is commonly expressed by calculating intra and interrater reliability and agreement (Martin and Bateson, 1993; Kottner et al., 2011). Reliability is defined as the capability of raters using locomotion scores to differentiate among individuals (Kottner et al., 2011) e.g. capability to differentiate between cows scored in level 1 and level 2. Agreement indicates the capability of raters to assign identical locomotion scores to an individual (Kottner et al., 2011).

Locomotion scoring is performed in different environmental conditions and by raters with different background and experience levels. In literature, locomotion scoring is generally conducted under farm conditions by live observations with cows walking across a flat and even surface. Reliability and agreement, however, are often estimated on observations from video recordings on a sample of cows (Flower and Weary, 2006; Borderas et al., 2008; Channon et al., 2009; Hoffman et al., 2013). Compared to live locomotion scoring, locomotion scoring from video enables registration of details that occur too fast or that are too complex to detect during live scoring and allows multiple scoring of the same cow (Martin and Bateson 1993). On the other hand, video recordings provide a limited context for observation of cows and the quality of recordings may have an important effect on the decision of the raters (Bench et al., 1974; Rogowitz et al., 2001). In this regard, locomotion scores obtained from live observations may differ from locomotion scores obtained from video observations (Martin and Bateson, 1993). Therefore, it is important to evaluate the reliability and agreement when locomotion scoring is done live and from video. In addition, it is relevant to know if locomotion scoring from video, as an alternative for live scoring, determines the same cows as lame and which factors influence this most. Therefore the aim of this study was to evaluate intrarater and interrater reliability and agreement of experienced and inexperienced raters for locomotion scoring performed live and from video, and to calculate the influence of raters and the method of observation (live or video) on the probability of classifying a cow as lame.

3.2. Materials and Methods

3.2.1. Animals and housing

This study was carried out on a commercial dairy farm located in Yifat, Israel. The dairy herd comprised 951 lactating Holstein cows distributed over 11 production groups. Each group was housed in a separate roofed cowshed without cubicles with dry manure bedding. The cows were milked three times a day (03:00 h, 11:00 h and 19:00 h) in a 2 x

32 parallel milking parlor. Annual milk production was on average 11,500 kg/cow. A total mixed ration supplied by a local feed company was provided twice daily. Drinking water was available ad libitum.

3.2.2. Locomotion scoring method, raters and training

The locomotion scoring method used in the experiment was based on the score proposed by Flower and Weary (2006). It consisted of a five-level scale based on judging gait asymmetry, reluctance to bear weight, arched back and head bobbing. The locomotion scoring method used described cows in level 1 as having smooth and fluid gait; level 2 imperfect locomotion but with ability to move freely; level 3 compromised capability to move freely; level 4 obviously diminished capability to move freely and level 5 severely restricted capability to move and must be vigorously encouraged to move.

Locomotion scoring was performed by five raters with different backgrounds and experience levels. The raters were part of a multidisciplinary project team and had to work together on the development of an automatic locomotion scoring system (Viazzi et al., 2013). Experienced Rater 1 (Rater-Exp 1) was a veterinarian with three different trainings in locomotion scoring. Prior to the experiment, Rater-Exp 1 conducted locomotion scoring live and from video on approximately 200 cows weekly for six months. Experienced rater 2 (Rater-Exp 2) and 3 (Rater-Exp 3) had agricultural backgrounds and joined one training in locomotion scoring prior the present experiment. In the last six months prior to the experiment, Rater-Exp 2 and 3 scored approximately 100 cows every two weeks by live observation. Inexperienced Rater 4 (Rater-Inexp 4) and 5 (Rater-Inexp 5) had no agricultural background and no previous experience in locomotion scoring in cows.

One week prior to the beginning of data gathering, Rater-Exp 2 and 3, and Rater-Inexp 4 and 5 were trained by Rater-Exp 1. The objective of the training was to introduce raters to the locomotion scoring method used in the experiment and to the practical experimental conditions. Training was divided into three sessions. During the first session, five videos per level of the locomotion scoring method used in this experiment were shown and the gait and posture traits were discussed among raters. In session 2 the live locomotion scoring was performed and in session 3 the video scoring session. At the beginning of session 2 and 3 approximately 20 cows were observed in order to discuss locomotion and individual gait and posture traits of cows. Thereafter, raters scored 140 cows live and 50 cows from video. Interrater reliability of the training sessions is shown in Table 3.1. The training was the only period in which raters were allowed to discuss locomotion scoring. The cows observed during the training period were not included in the experiment.

3.2.3. Locomotion scoring live and from video

Live locomotion scoring was performed while cows walked through an alley (1.5 m wide, 7 m long) with a flat concrete floor. This alley was situated at the exit of the milking area. Depending on the walking speed of the cow, raters had between 7 to 45 s to identify the cow, to score locomotion, and to write the results on a predefined form. Rater-Exp 1 and 2 and Rater-Inexp 4 and 5 were positioned 6.5 m perpendicular to the progression line of the alley. Rater-Exp 3 was positioned in the vicinity of the entrance to the alley to control cow access (Figure 3.1).

At the same time, a camera (Canon EOS 60D, Canon Inc, Tokyo, Japan) equipped with a lens Canon EF-S 17-85 mm IS USM, (Canon Inc, Tokyo, Japan) recorded continuously the cows walking through the alley. The camera was positioned in close proximity to the raters, 6.5 m perpendicular to the progression line of the alley and 1.35 m above ground level, in order to obtain flank views of a similar perspective as raters (Figure 3.1 and 3.2). Video recordings had a resolution of 1920 x 1080 pixels at a frame rate of 25 frames per second in .mov file format. To obtain individual video recordings of each cow, the videos were edited with Quick Time 7 Pro (Apple Inc, CA, U.S.A). All video recordings were stored on an external hard drive (WD elements, CA, U.S.A).

The edited video recordings of individual cows were used to perform locomotion scoring from video. The videos were projected onto a 20 inch screen (Fujicom FJ-2040-LED, Fujicom HK Ltd, Kowloon, Hong Kong) with a resolution of 1600 x 900 pixels. During locomotion scoring from video the five raters were located approximately 1.5 m away from the screen on which the videos were shown.

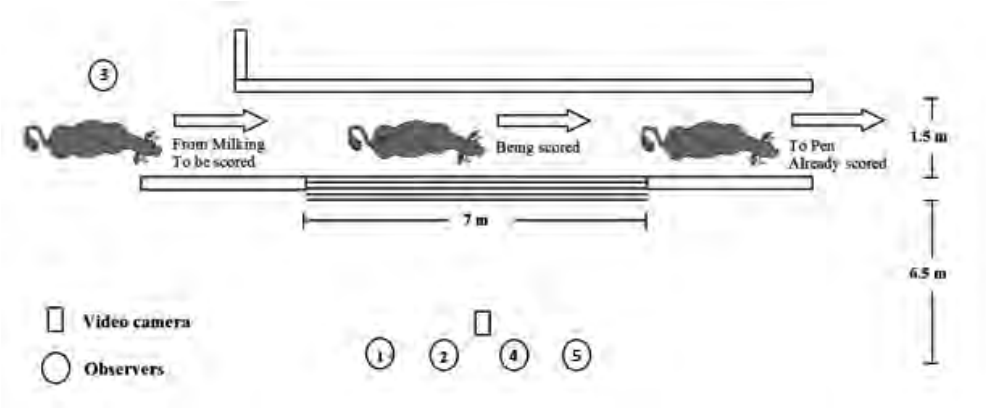


Figure 3.1. Schematic view of situation at the barn for live scoring and video recordings

Figure 3.2. Picture of a video recording shown during the video scoring

3.2.4. Data gathering schedule

Sample size (N = 492) was calculated considering a lameness prevalence (lameness defined as locomotion score ≥ 3) of 20% (which was measured for the whole herd by Rater-Exp 1 and 3 before the experiment) and a confidence interval of 10%. In order to increase the probability of cows with different locomotion scores eight production groups with mainly multiparous cows were selected for the experiment.

To be able to observe the cows of eight production groups, locomotion scoring was performed in three consecutive weeks (week 1, 2 and 3). In each week two live scorings (live scoring 1 and 2) and two video scorings (video scoring 1 and 2) were performed. Within the same week the same groups were scored in the two live and two video scorings. Different production groups were scored between weeks. In week 1, live scoring 1 and 2 were performed on one day at 6:00 and 14:00 including production groups that consisted of multiparous cows, cows to be culled and the hospital group. In week 2, due to the lack of light at dawn when the cows left the milking parlour, live scoring 1 and 2 were performed on two consecutive days at 13:00 including production groups that consisted of primiparous and multiparous cows with low milk yield. In week 3, production groups that consisted of primiparous and slow milking cows were scored on two consecutive days according to a schedule similar to week 2. The order in which different groups were scored live was done in a way not to interfere with the normal farm routine.

In week 1, 2 and 3 scoring from video was performed two days after the live scoring and followed the schedule for live scoring for the different production groups. Two video scorings were performed in one day: video scoring 1 was performed at 9:00 and video scoring 2 at 14:00. The same videos, recorded during live scoring 2, were shown in video scoring 1 and 2, but in different random order in each scoring session to reduce the risk of cow recognition by the raters. After every 100 videos there was a 10 minute break. In an attempt to simulate live scoring, each video was shown only once and raters had to record the cow identification number and the locomotion score. Raters were not allowed to comment on locomotion scoring during and after live and video scorings.

3.2.5. Statistical analysis

Due to the relatively short observation time per cow and the large number of cows, the raters did not score exactly the same number of cows. Particularly in the live sessions, cows were sometimes missed for scoring. The exact number of cows differed per analysis and is stated where relevant. Since Rater-Exp 3 was located in a different position, his live locomotion scores were not included in the statistical analysis.

Average distribution for the five-level locomotion score was calculated for 208 cows that were scored by all raters in both live and video scorings. Difference in distributions for the same rater for live scorings and video scorings was estimated with Bowker's symmetry test for the five-level scale. Level of significance was stated at $P < 0.05$.

The intrarater reliability and agreement were calculated by comparing the scores assigned by the same rater to the same cow. The interrater reliability and agreement were calculated by comparing scores assigned by each rater in relation to Rater-Exp 1. Intrarater and interrater reliability and agreement were calculated for live/live, live/video and video/video comparisons considering individual raters in each of the three weeks and as overall considering all locomotion scores assigned in the experiment per rater.

Intrarater and interrater reliability were expressed as weighted kappa coefficient (κw) which is a suitable reliability indicator for ordinal scales with multiple levels (Cohen, 1968). The κw was calculated using linear weighting as proposed by Cicchetti and Allison (1971). Intrarater and interrater agreement were expressed as percentage of agreement (PA) for a five-level scale. The PA was calculated by dividing the number of agreements by the total number of agreements and disagreements (Martin and Bateson, 1993). The 95% confidence interval (CI) for κw was calculated as proposed by Fleiss *et al* (1969), whereas Clopper-Pearson CI was calculated for PA (Brown et al., 2001). The acceptance threshold was set at $\kappa w \geq 0.4$ (March et al., 2007; Burn and Weir, 2011). In addition a $\kappa w \geq 0.6$ can be classified as substantial and $\kappa w \geq 0.8$ as excellent (Landis and Koch, 1977). Acceptance threshold for PA was ≥ 75% (Burn and Weir, 2011). Intrarater and interrater percentage of disagreement was calculated dividing the disagreements obtained by raters among specific

levels within the five-level scale divided by the total number of cows locomotion scored during the three weeks of experiment. All above mentioned analyses were performed using PROC FREQ within the statistical software package SAS 9.2 (SAS Institute Inc., Cary, NC).

A generalized linear mixed model was used to calculate the relative size of the fixed effects on the probability of classifying a cow as lame by performing locomotion scoring live and from video. This model was performed on a logistic scale. The model comprised the fixed effects of rater (Rater-Exp 1 and 2, Rater-Inexp 4 and 5), method (live scoring 2 and video scoring 2) and interactions between raters and method. Cows were included as random effect. In a logistic regression, the Wald statistics divided by the degrees of freedom (Wald/df) indicate the relative size of the fixed effect (McCullagh and Nelder, 1989). Logistic regression was performed using GenStat Version 14.2.0.6297 (VSN International Ltd, Hemel Hempstead, UK)

3.3. Results

3.3.1. Training

Interrater reliability and agreement values obtained by experienced and inexperienced raters in comparison to Rater-Exp 1 during the training session are shown in Table 3.1. For live/live comparison only comparison between Rater-Exp 1 and 3, exceeded the acceptance threshold, κw = 0.48 (Table 3.1), whereas for video/video comparison experienced and inexperienced exceeded the acceptance threshold for κw when compared with Rater-Exp 1 (Range κw = 0.48 – 0.53) (Table 3.1). Interrater agreement did not exceed the threshold in any of the comparisons among raters (Table 3.1).

Table 3.1. Interrater reliability (expressed as weighted kappa, κw) and agreement (expressed as percentage of agreement, PA) of the training sessions for live/live (L/L) and video/video (V/V) comparisons for two experienced (Rater-Exp) and two inexperienced (Rater-Inexp) with the trainer (Rater-Exp 1). CI indicates 95% confidence interval.

	Rater	N[a]	κw (CI)	PA (CI)
L/L	Rater–Exp 2	103	0.39 (0.23 – 0.55)	53.4 (43.3 – 63.3)
	Rater–Exp 3	79	0.48 (0.32 – 0.64)	58.2 (46.6 – 69.2)
	Rater–Inexp 4	101	0.35 (0.17 – 0.52)	49.5 (39.4 – 59.6)
	Rater–Inexp 5	77	0.14 (0.00 - 0.29)	50.6 (39.7 – 62.2)
V/V	Rater–Exp 2	38	0.52 (0.35 – 0.70)	52.6 (35.8 – 67.5)
	Rater–Exp 3	36	0.48 (0.27 – 0.68)	52.8 (35.5 – 69.6)
	Rater–Inexp 4	39	0.53 (0.35 – 0.72)	56.4 (39.6 – 72.2)
	Rater–Inexp 5	39	0.48 (0.31 – 0.66)	53.8 (37.2 – 69.9)

[a] Number of comparisons

3.3.2. Distribution of locomotion scores

The distribution of the locomotion scores of the 208 cows scored by all five raters in all live and video scorings are shown in Table 3.2. The distribution for live and video scoring was only different for Rater-Exp 2 for the five-level and non-lame/lame classification ($P < 0.05$). For video observation, experienced raters reported lameness prevalence of about 25% whereas for inexperienced raters lameness prevalence was about 15% (Table 3.2).

Table 3.2. Distribution of scores for live and video locomotion scoring on a five–level scale scored by experienced (Exp–rater) and inexperienced raters (Inexp–rater) across all sessions (208 cows) scored by all raters in all sessions.

	Five Levels					Two levels	
	Level 1, %	Level 2, %	Level 3, %	Level 4, %	Level 5, %	Non–Lame %	Lame %
Exp–rater 1							
Live	24.5	46.2	21.8	6.3	1.2	70.7	29.3
Video	30.0	40.9	22.3	6.3	0.5	70.9	29.1
Exp–rater 2							
Live	41.3	42.3	13.0	2.9	0.5	83.6	16.4
Video	25.0	49.1	19.2	5.5	1.2	74.1	25.9
Exp–rater 3							
Live	–	–	–	–	–	–	–
Video	36.8	39.2	15.6	6.0	2.4	76.0	24.0
Inexp–rater 4							
Live	40.1	43.3	12.7	2.9	1.0	83.4	16.6
Video	33.2	50.5	10.8	4.3	1.2	83.7	16.3
Inexp–rater 5							
Live	50.0	34.2	10.3	4.8	0.7	84.2	15.8
Video	50.7	34.1	10.8	3.4	1.0	84.8	15.2

3.3.3. Intrarater reliability, agreement and disagreement

Overall intrarater reliability, agreement and disagreements for live/live, live/video, and video/video comparisons for different raters using the five-level scale are shown in Table 3.3. The CIs indicate that intrarater reliability and agreement for Rater-Exp 1 and 2 was lower for live/live than for video/video (Table 3.3). Overall intrarater reliability and agreement for inexperienced raters for live/live showed no difference with video/video comparison (Table 3.3). Overall intrarater reliability and agreement for live/video comparison was similar to values obtained in live/live comparison for experienced and inexperienced raters (Table 3.3).

Percentage of disagreement for intrarater comparison showed that most disagreements are due to one level difference. Percentage of disagreement was high for level 1 and 2 and for level 2 and 3 (Table 3.3).

Intrarater reliability and agreement for live/live, live/video, and video/video comparisons in three different weeks are shown in Table 3.4. The CIs suggest that experienced raters had lower intrarater reliability and agreement values for live/live than for video/video comparison in three weeks of experiment (Table 3.4). For inexperienced raters CIs suggest that there was no difference for intrarater reliability and agreement in live/live and video/video comparison in the three weeks of the experiment (Table 3.4). During the three weeks of observation live/video comparison showed values similar to those obtained in live/live comparison for experienced and inexperienced raters (Table 3.4)

3.3.4. Interrater reliability, agreement and disagreement

Interrater reliability and agreement for live/live, live/video, and video/video comparisons for experienced and inexperienced raters compared with Rater-Exp 1 for the five-level scale are shown in Table 3.5. The CIs indicate that interrater reliability and agreement for experienced raters was lower for live/live than for video/video comparison (Table 3.5). When compared with inexperienced raters interrater reliability and agreement showed no differences for live/live and video/video (Table 3.5).

Percentage of disagreement for interrater comparison showed that most of disagreements are due to one level difference. Percentage of disagreement was high for levels 1 and 2 and for levels 2 and 3 (Table 3.5).

The CIs for interrater reliability and agreement of experienced raters compared to Rater-Exp 1 indicated that live/live comparison had lower values than video/video comparison in week 1 and 2 (Table 3.6). When compared with inexperienced raters interrater reliability and agreement showed no differences for live/live, live/video and video/video comparisons along the three weeks of experiment (Table 3.6).

Table 3.3. Overall intrarater reliability (expressed as weighted kappa, κw), agreement (expressed as percentage of agreement, PA) and disagreements (expressed as percentage of disagreements) among specific levels for a five–level locomotion score performed by three experienced (Rater-Exp) and two inexperienced (Rater-Inexp) raters for live/live (L/L), live/video (L/V) and video/video (V/V). CI indicates 95% confidence interval.

		Reliability and agreement			Disagreements									
	Rater	**N[a]**	**κw (CI)**	**PA (CI)**	**1-2**	**1-3**	**1-4**	**1-5**	**2-3**	**2-4**	**2-5**	**3-4**	**3-5**	**4-5**
L/L	Rater–Exp 1	409	0.39 (0.32 – 0.46)	48.9 (43.9 - 53.9)	19.6	4.2	0.2	0.0	19.3	1.7	0.0	4.2	0.5	1.5
	Rater–Exp 2	452	0.45 (0.39 – 0.52)	56.7 (52.0 - 61.4)	27.2	2.4	0.2	0.0	10.2	0.7	0.0	2.0	0.0	0.7
	Rater–Exp 3	–	–	–	–	–	–	–	–	–	–	–	–	–
	Rater–Inexp 4	427	0.42 (0.35 – 0.50)	56.3 (51.5 - 61.1)	26.2	1.9	0.5	0.0	11.0	1.9	0.0	2.1	0.0	0.2
	Rater–Inexp 5	504	0.39 (0.31 – 0.46)	56.2 (51.7 - 60.5)	26.2	5.8	1.2	0.0	7.5	1.2	0.0	1.2	0.0	0.8
L/V	Rater–Exp 1	466	0.46 (0.40 – 0.53)	54.1 (49.4 – 58.7)	18.2	4.5	0.4	0.0	17.4	1.1	0.0	3.9	0.0	0.4
	Rater–Exp 2	511	0.34 (0.28 – 0.41)	48.5 (44.1 – 52.9)	29.9	4.7	0.4	0.0	12.3	1.6	0.0	1.8	0.2	0.6
	Rater–Exp 3	-	-	-	-	-	-	-	-	-	-			
	Rater–Inexp 4	515	0.31 (0.24 – 0.38)	47.9 (43.5 - 52.9)	29.3	2.5	1.0	0.0	14.2	1.4	0.0	3.3	0.2	0.2
	Rater–Inexp 5	539	0.32 (0.24 – 0.39)	52.1 (47.8 – 53.4)	29.9	6.7	0.7	0.0	6.9	0.7	0.0	2.2	0.0	0.7
V/V	Rater–Exp 1	544	0.60 (0.55 – 0.66)	64.8 (60.6 - 68.8)	17.6	1.3	0.2	0.0	12.1	0.6	0.0	2.9	0.2	0.2
	Rater–Exp 2	564	0.59 (0.54 – 0.64)	66.1 (62.1 - 70.0)	18.1	0.9	0.2	0.0	9.6	0.7	0.0	3.7	0.0	0.5
	Rater–Exp 3	563	0.60 (0.55 – 0.65)	62.2 (58.1 - 66.2)	18.5	1.1	0.0	0.0	13.9	0.2	0.0	3.7	0.0	0.4
	Rater–Inexp 4	554	0.41 (0.35 – 0.47)	56.6 (52.4 - 60.8)	24.9	2.0	0.5	0.0	11.6	0.7	0.0	2.7	0.0	0.9
	Rater–Inexp 5	573	0.34 (0.27 – 0.42)	52.2 (48.0 - 56.3)	32.5	4.0	0.7	0.0	8.4	0.9	0.0	0.9	0.0	0.7

a Number of comparisons

Table 3.4. Intrarater reliability (expressed as weighted kappa, κw) and agreement (expressed as percentage of agreement, PA) for a five–level locomotion score performed by three experienced (Rater-Exp) and two inexperienced (Rater-Inexp) raters for live/live (L/L), live/video (L/V) and video/video (V/V) comparisons in three consecutive weeks (Weeks 1, 2 and 3). CI indicates 95% confidence interval.

		Week 1			Week 2			Week 3		
		N[a]	κw (CI)	PA (CI)	N[a]	κw (CI)	PA (CI)	N[a]	κw (CI)	PA (CI)
L/L	Rater–Exp 1	181	0.46 (0.35 – 0.55)	50.2 (42.8 – 57.8)	102	0.25 (0.11 – 0.37)	47.1 (37.1 – 57.2)	126	0.38 (0.26 – 0.51)	48.4 (39.4 – 57.5)
	Rater–Exp 2	243	0.49 (0.41 – 0.58)	58.4 (52.2 – 64.7)	115	0.25 (0.10 – 0.40)	51.3 (41.8 – 60.7)	94	0.51 (0.36 – 0.66)	58.5 (47.8 – 68.6)
	Rater–Exp 3	–	–	–	–	–	–		–	–
	Rater–Inexp 4	171	0.48 (0.37 – 0.60)	57.3 (49.5 – 64.8)	113	0.39 (0.26 – 0.52)	59.3 (49.7 – 68.4)	143	0.33 (0.19 – 0.48)	52.8 (44.3 – 61.2)
	Rater–Inexp 5	212	0.29 (0.18 – 0.40)	43.8 (37.1 – 50.8)	143	0.44 (0.29 – 0.59)	73.4 (65.4 – 80.4)	149	0.37 (0.23 – 0.51)	57.1 (48.7 – 65.1)
L/V	Rater–Exp 1	213	0.49 (0.49 – 0.57)	51.6 (44.7 – 58.5)	127	0.28 (0.15 – 0.41)	48.0 (39.1 – 57.1)	126	0.55 (0.42 – 0.67)	64.3 (55.3 – 72.6)
	Rater–Exp 2	254	0.30 (0.21 – 0.38)	41.3 (35.2 – 47.7)	142	0.31 (0.19 – 0.41)	52.1 (43.6 – 60.6)	115	0.49 (0.36 – 0.62)	60.0 (50.5 – 69.0)
	Rater–Exp 3	–	–	–	–	–	–	–	–	–
	Rater–Inexp 4	243	0.38 (0.29 – 0.47)	49.0 (42.5 – 55.4)	137	0.09 (0.0 – 0.15)	42.3 (33.9 – 51.1)	135	0.36 (0.22 – 0.50)	51.9 (43.1 – 60.5)
	Rater–Inexp 5	248	0.32 (0.22 – 0.41)	45.2 (38.9 – 51.6)	155	0.26 (0.14 – 0.37)	60.0 (51.8 – 67.8)	136	0.31 (0.16 – 0.47)	55.9 (47.1 – 64.4)
V/V	Rater–Exp 1	244	0.66 (0.59 – 0.73)	66.0 (59.7 – 71.9)	157	0.53 (0.42 – 0.63)	62.5 (55.2 – 69.4)	143	0.52 (0.40 – 0.65)	65.0 (56.6 – 72.8)
	Rater–Exp 2	259	0.60 (0.53 – 0.68)	65.3 (59.1 – 71.0)	159	0.64 (0.54 – 0.74)	73.6 (66.1 – 80.3)	146	0.50 (0.38 – 0.61)	60.3 (51.9 – 68.3)
	Rater–Exp 3	252	0.60 (0.53 – 0.68)	58.7 (52.4 – 64.9)	164	0.53 (0.42 – 0.64)	65.1 (57.4 – 72.5)	147	0.64 (0.55 – 0.73)	65.3 (56.4 – 73.1)
	Rater–Inexp 4	257	0.46 (0.37 – 0.56)	58.0 (51.7 – 64.1)	155	0.27 (0.14 – 0.40)	54.8 (46.7 – 62.8)	142	0.41 (0.27 – 0.54)	56.3 (47.8 – 64.6)
	Rater–Inexp 5	263	0.41 (0.31 – 0.50)	49.8 (43.6 – 55.7)	164	0.24 (0.12 – 0.36)	52.4 (44.5 – 60.3)	146	0.28 (0.12 – 0.44)	55.5 (47.4 – 63.7)

[a] Number of comparisons.

3.3.5. Effect of raters and method of observation on lameness classification

The size of Wald/df, obtained in the logistic regression suggested that the rater (Wald/df = 59.9; $P < 0.05$) was the most important factor affecting the classification of lame cows. To a lesser extent the interaction between observer and method (Wald/df = 12.9;$P < 0.05$) and the method (Wald/df = 4.8; $P < 0.05$) were also affecting the classification for lame cows.

3.4. Discussion

In the present study, raters showed differences in the distribution of locomotion scores using a five-level scale. In addition, differences in the distribution of locomotion scores between experienced and inexperienced raters might be a result of inexperienced raters tending to classify less cows as lame (locomotion score ≥ 3) when compared to experienced raters.

Although experienced raters reached substantial κw values in weeks 1 and 2 for intrarater reliability in video/video comparison, they were not able to obtain the same substantial κw values in live/live and live/video comparison. Cows may have displayed variations in locomotion in the live scoring 1 and 2 due to factors related to the cows (e.g. hoof disorders not present in live scoring 1 but present in live scoring 2), or factors related to the environment (e.g floor conditions). The concentration and performance of the raters also might have been different, for example, due to other groups of cows going to the milking parlour, background noise, or people passing by. All these factors are commonly present in practical farm conditions for live scoring. Other factors that may explain a higher intrarater and interrater reliability and agreement when locomotion scoring was performed from video are: the possibility for the raters to focus on a single cow, elimination of variation associated to observing a cow at different moments. Given the large number of cows included in the experiment, the effect of memorizing cows seems of minor importance; only a few cows with exceptional characteristics (e.g. completely white cows or severely lame cows) were remembered sometimes. The moderate values for κw and PA give an additional, unforeseen indication for this.

In the literature, few articles reported a comparison between live and video locomotion scoring. In agreement with our study, Bernardi *et al* (2009) found no differences in interrater reliability when two raters were compared for live/live and live/video locomotion scoring (Bernardi et al., 2009). Another study showed no differences in interrater agreement calculated for live/live or video/video comparison (Channon et al., 2009).

Table 3.5. Overall interrater reliability (expressed as weighted kappa, κw), agreement (expressed as percentage of agreement, PA) and disagreements (expressed as percentage of disagreements) among specific levels for a five–level locomotion score for two experienced (Rater-Exp) and two inexperienced (Rater-Inexp) in comparison with experienced rater 1 for live/live (L/L), live/video (L/V) and video/video (V/V) comparisons. CI, indicates 95% confidence interval.

		Reliability and agreement			Disagreements									
	Rater	**N[a]**	**κw (CI)**	**PA (CI)**	**1-2**	**1-3**	**1-4**	**1-5**	**2-3**	**2-4**	**2-5**	**3-4**	**3-5**	**4-5**
L/L	Rater–Exp 2	952	0.35 (0.39 – 0.40)	45.3 (42.1 - 48.5)	29.4	5.5	0.3	0.0	14.3	1.4	0.0	2.5	0.3	1.1
	Rater–Exp 3	–	–	–	–	–	–	–	–	–	–	–	–	–
	Rater–Inexp 4	924	0.33 (0.28 – 0.38)	45.8 (42.4 – 48.9)	26.4	5.3	0.5	0.0	16.2	2.1	0.1	2.8	0.1	0.6
	Rater–Inexp 5	970	0.29 (0.25 – 0.34)	41.0 (38.0 – 44.2)	31.2	8.8	0.7	0.0	12.6	1.5	0.0	2.6	0.3	1.2
L/V	Rater–Exp 2	974	0.37 (0.33 – 0.42)	50.2 (47.1 – 53.4)	22.8	5.3	0.5	0.0	15.5	1.7	0.1	3.1	0.2	0.5
	Rater–Exp 3	–	–	–	–	–	–	–	–	–	–	–	–	–
	Rater–Inexp 4	977	0.32 (0.27 – 0.37)	46.9 (43.7 – 50.1)	25.6	5.7	0.6	0.0	15.9	2.1	0.2	2.5	0.2	0.3
	Rater–Inexp 5	1004	0.26 (0.22 – 0.31)	41.7 (38.7 – 44.9)	30.0	10.4	0.7	0.0	11.7	1.9	0.0	2.8	0.3	0.6
V/V	Rater–Exp 2	1097	0.47 (0.43 – 0.52)	56.2 (53.2 - 59.1)	18.7	1.9	0.0	0.1	17.5	1.0	0.0	3.7	0.3	0.6
	Rater–Exp 3	1102	0.47 (0.43 – 0.52)	54.2 (51.2 – 57.1)	20.8	3.0	0.0	0.1	15.9	1.5	0.0	3.4	0.5	0.7
	Rater–Inexp 4	1076	0.35 (0.31 – 0.40)	50.3 (47.2 – 53.3)	24.0	3.9	0.8	0.0	15.8	1.1	0.1	3.1	0.2	0.7
	Rater–Inexp 5	1099	0.33 (0.29 – 0.37)	46.0 (43.1 – 49.0)	28.8	6.4	0.6	0.0	12.9	1.2	0.1	3.1	0.1	0.7

[a] Number of comparisons

Table 3.6. Interrater reliability (expressed as weighted kappa, κw) and agreement (expressed as percentage of agreement, PA) for a five–level locomotion score for two experienced (Rater-Exp) and two inexperienced (Rater-Inexp) in comparison with experienced rater 1 for live/live (L/L), live/video (L/V) and video/video (V/V) comparisons in three consecutive weeks (Weeks 1, 2 and 3). CI, indicates 95% confidence interval.

		Week 1			Week 2			Week 3		
		N[a]	κw (CI)	PA (CI)	N[a]	κw (CI)	PA (CI)	N[a]	κw (CI)	PA (CI)
L/L	Rater–Exp 2	457	0.38 (0.32 – 0.44)	42.7 (38.1 – 47.2)	247	0.24 (0.16 – 0.33)	45.3 (39.8 – 51.8)	248	0.39 (0.29 – 0.49)	50.0 (43.6 – 56.4)
	Rater–Exp 3	–	–	–	–	–	–	–	–	–
	Rater– Inexp 4	402	0.33 (0.26 – 0.41)	40.0 (35.2 – 45.0)	238	0.17 (0.07 – 0.26)	45.0 (38.5 – 51.5)	284	0.44 (0.35 – 0.52)	54.2 (48.2 – 60.1)
	Rater– Inexp 5	426	0.36 (0.30 – 42.8)	42.3 (35.7 – 47.1)	261	0.13 (0.07 – 0.19)	36.8 (30.9 – 42.9)	283	0.27 (0.17 – 0.36)	43.1 (37.3 – 49.1)
L/V	Rater–Exp 2	465	0.36 (0.30 – 0.42)	44.5 (39.9 – 49.2)	269	0.28 (0.19 – 0.38)	49.8 (43.7 – 55.9)	240	0.49 (0.40 – 0.59)	61.7 (55.2 – 67.8)
	Rater–Exp 3	–	–	–	–	–	–	–	–	–
	Rater– Inexp 4	455	0.34 (0.27 – 0.41)	43.7 (39.1 – 48.4)	262	0.20 (0.11 – 0.29)	45.5 (39.5 – 51.7)	260	0.37 (0.26 – 0.47)	53.8 (47.6 – 59.8)
	Rater– Inexp 5	458	0.32 (0.26 – 0.38)	41.5 (36.9 – 46.1)	285	0.15 (0.08 – 0.22)	40.0 (34.3 – 45.9)	261	0.23 (0.13 – 0.34)	44.1 (37.9 – 50.3)
V/V	Rater–Exp 2	495	0.52 (0.47 – 0.58)	55.2 (50.6 – 59.6)	315	0.41 (0.33 – 0.49)	55.9 (50.2 – 61.3)	287	0.42 (0.33 – 0.51)	58.2 (52.2 – 64.0)
	Rater–Exp 3	492	0.50 (0.45 – 0.56)	52.4 (47.9 – 56.9)	332	0.38 (0.30 – 0.46)	52.4 (46.9 – 57.9)	289	0.47 (0.39 – 0.55)	57.4 (51.5 – 63.2)
	Rater– Inexp 4	491	0.42 (0.36 – 0.48)	49.7 (45.2 – 54.2)	307	0.23 (0.15 – 0.32)	50.8 (45.1 – 56.5)	278	0.31 (0.21 – 0.42)	50.7 (44.7 – 56.7)
	Rater– Inexp 5	496	0.38 (0.32 – 0.44)	45.6 (41.1 – 50.1)	318	0.23 (0.16 – 0.30)	45.0 (39.4 – 50.6)	285	0.28 (0.17 – 0.38)	48.1 (42.1 – 54.0)

[a] Number of comparisons

Intrarater reliability values (expressed as κw) for live/live comparison in the present study for experienced raters were similar to the results obtained by Thomsen *et al* (2008) who reported κw values ranging from 0.38 to 0.64, reaching in most cases moderate agreement. Intrarater agreement values (expressed in PA) were similar to results of O'Callaghan *et al* (2003) who reported a PA of 56%. Both articles (O'Callaghan et al., 2003; Thomsen et al., 2008) used a similar live/live comparison and a five-level scale for locomotion scoring as in the current study. In contrast to our experiment, locomotion scoring performed by Thomsen *et al* (2008), was done under experimental conditions and raters were allowed to score cows from different positions. For video/video comparison intrarater reliability and agreement for experienced raters in the current study were lower than values reported by Schlageter-Tello *et al* (2014a) with κw ranging from 0.63 to 0.83 and PA ranging from 60.3% to 82.8%. Values reported by Schlageter-Tello *et al* (2014a) were obtained scoring a relatively small number of cows (N = 58), each video was showed two times and all raters were experienced. Other articles reporting intrarater reliability or agreement for live/video or video/video comparisons are not directly comparable with the results obtained in the present study. Channon *et al* (2009) reported an intrarater agreement of 30% for a similar live/video comparison using a nine-level scale for locomotion scoring. High intrarater reliability for video/video comparison were reported, however, those are expressed as coefficient of determination (R^2 ranging from 0.75 to 0.98) (Flower and Weary, 2006; Flower et al., 2008) or Pearson correlation coefficient (r = 0.92) (Borderas et al., 2008). The acceptance threshold for reliability expressed as r is ≥ 0.7 (Martin and Bateson, 1993).

Interrater reliability and agreement were below the threshold of moderate reliability (κw < 0.4) for all pairwise comparisons with Rater-Exp 1 for live scorings and below the threshold for substantial reliability for video scorings. In the literature, reported interrater reliability and agreement showed high variation among or even within articles (Schlageter-Tello et al., 2014b). Values for interrater reliability obtained in the present experiment were lower than those reported by Thomsen *et al* (2008) (κw values ranging from 0.24 to 0.68) and March *et al* (2007) (κw ranging from 0.41 to 0.86) with a similar live/live comparison and five-level scale. Interrater agreement for live/live comparison in the present study were similar (PA = 36%) (O'Callaghan et al., 2003) or lower (PA = 63% to 74%, Winckler and Willen 2001; and PA = 45% to 96%, March *et al* 2007) than other values reported in the literature using a similar five-level scale. Interrater reliability with κw ranging from 0.30 to 0.40 was reported for live/video comparison with a similar five-level scale (Danscher et al., 2009). For video/video comparison and similar five-level scale, interrater reliability ranged from κw = 0.57 to 0.68, whereas interrater agreement was 83% (Hoffman et al., 2013). Recently, Schlageter-Tello *et al* (2014a) reported a large variation for interrater reliability and agreement obtained with a similar video/video comparison obtained by experienced raters without further training; κw values in that study ranged from 0.28 to 0.82 and PA from 22.6% to 77.2%.

Fair to moderate reliability and agreement values obtained in the present study suggest that although experienced and inexperienced raters received training, the training performed was not sufficient to improve reliability and agreement. Inexperienced raters showed no improvement for the intrarater and interrater reliability and agreement during the three weeks. Experienced raters showed an increment in the intrarater and interrater reliability and agreement for the live/video comparison in week 3 suggesting that training may decrease differences in reliability and agreement between live and video scoring. However, measurements along more weeks and with more experienced raters are required to confirm this finding. In this regard, both experienced and inexperienced raters would have needed more training to for higher reliability and agreement values. Different studies indicate that training is one of the main factors to improve reliability and agreement of raters (Winckler and Willen, 2001; March et al., 2007; Thomsen et al., 2008). Though, there are studies confirming the limited and variable improvement in reliability and agreement after training (Engel et al., 2003; Thomsen et al., 2008). Variable results for improvement of reliability and agreement indicates that there is not a standard training for locomotion scoring as existing for other scoring systems such as body condition score (Vasseur et al., 2013) and injury score (Gibbons et al., 2012). Possible solutions for the improvement of reliability and agreement values in raters may be the inclusion of a mid-experiment control for reliability and agreement and include extra training sessions if required or to allow raters to comment on the scores assigned to cows among sessions. In addition, the utilization of a simpler locomotion score (with less levels and traits to be observed) would be useful in the improvement of reliability and agreement of raters.

In accordance with our study, Winckler and Willen (2001) reported that the highest number of disagreements in a similar five-level scale was between level 1 and 2. In both studies, however, about 80% of cows were scored in level 1 and 2. In a recent study in which raters had to classify cows from video that were selected to have a similar number of videos for each level the lowest agreement was for level 2 and 3, suggesting that it is more difficult for raters to differentiate between these two levels (Schlageter-Tello et al., 2014a).

The acceptance threshold of ($\kappa w \geq 0.4$) in the current study was selected because it was used in most studies using κw and κ (Brenninkmeyer et al., 2007; March et al., 2007; Burn and Weir, 2011) at the time our experiment was performed. This acceptance threshold may be considered low when compared with the acceptance threshold used in other studies estimating reliability and agreement of other observations. An acceptance threshold $\kappa w \geq 0.6$ was used for injuries scores in cows (Gibbons et al., 2012), and an acceptance threshold of $\kappa w \geq 0.8$ was proposed for body condition scoring in cows (Vasseur et al., 2013). However, it is stated that application of such thresholds may lead to questionable interpretations of κw values (Warrens, 2013). An example of this, is the fact that when calculated with the quadratic weighting, κw tend to have higher values than

when calculated with linear weighting (Warrens, 2013). In addition, reliability estimators are affected by the homogeneity of the population sample (e.g. only non-lame cows, (de Vet et al., 2006). The acceptance threshold for PA ≥ 75% was never exceeded in the current experiment for the five-level scale, which is in line with previous studies that showed that it is hard to exceed this threshold (Winckler and Willen, 2001; Schlageter-Tello et al., 2014a).

The logistic regression was included to detect possible influence of video scoring on the classification of cows as lame. Although logistic regression showed a significant effect for the factor method (live or video scoring) the small size of the Wald/df indicated that this effect was of less importance than the effect of the raters to influence the probability of classifying of cows as lame. The utilization of highly trained raters may contribute to decrease the effect of raters in the current study. However, an important variation in reliability and agreement values even for lame/non-lame classification has been reported in the literature (Schlageter-Tello et al., 2014b). In this regard, it would be unlikely that the effect of method (live or video) will be more important than the rater effect.

Important facts that may limit the conclusions obtained in the present study are the relative low values for κw and PA for intrarater and interrater reliability and agreement which suggest a high variation in the locomotion scores assigned to the cows by the raters, and the low number of experienced and inexperienced raters included in the experiment. Repeating the experiment with a larger number of raters with a similar training level (all raters experienced or all raters inexperienced) would provide stronger conclusions than in the current experiment.

3.5. Conclusions

Under the present experimental conditions, experienced raters showed lower intrarater and interrater reliability and agreement in live scoring than in video scoring. Inexperienced raters did not show differences in reliability and agreement when scoring live or from video. The live/video comparison showed reliability and agreement values similar to those obtained from live scoring for experienced and inexperienced raters. Since raters are the most important factors influencing the probability of classifying a cow as lame, video observation seems to be an acceptable method for locomotion scoring and lameness assessment in dairy cows.

Animal welfare implications

Lameness is considered an important welfare issue and it is commonly assessed with locomotion scoring methods. Video locomotion scoring showed no differences in relation to live scoring for classifying cows as lame. That means that video recording might be used for lameness detection. This gives further opportunities to develop technological tools for

lameness detection that make use of video recordings and automatic computer vision analysis (Viazzi et al., 2013) or simpler systems based on automatic selection of video records that may be shown to farmers/veterinarian for further analysis (Bruyere et al., 2012).

No standardised description of training protocols for locomotion scoring in dairy cows was found. It would be beneficial to develop training protocols that can help to improve reliability and agreement in both live and video locomotion scoring.

Acknowledgments
This study is part of the Marie Curie Initial Training Network BioBusiness project (FP7-PEOPLE-ITN-2008). Many thanks to Machteld Steensels and Daniel Rozen for their help with the data gathering; to the ARO technician Aharon Antler for building and maintaining the cow guiding construction; to Bas Engel and Willem Buist for their collaboration in the statistical analysis and to Jos Metz for his valuable advice.

References

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.

Barkema, H. W., J. D. Westrik, K. A. S. Vankeulen, Y. H. Schukken, and A. Brand. 1994. The Effects of Lameness on Reproductive-Performance, Milk-Production and Culling in Dutch Dairy Farms. Prev. Vet. Med. 20:249-259.

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.

Bench, J., E. Hoffman, and I. Wilson. 1974. A comparison of live and videorecord viewing of infant behavior under sound stimulation. I. Neonates. Dev. Psychobiol. 7:455-464.

Bernardi, F., J. Fregonesi, C. Winckler, D. M. Veira, M. A. G. von Keyserlingk, and D. M. Weary. 2009. The stall-design paradox: Neck rails increase lameness but improve udder and stall hygiene. J. Dairy Sci. 92:3074-3080.

Booth, C. J., L. D. Warnick, Y. T. Grohn, D. O. Maizon, C. L. Guard, and D. Janssen. 2004. Effect of lameness on culling in dairy cows. J. Dairy Sci. 87:4115-4122.

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.

Brenninkmeyer, C., S. Dippel, S. March, J. Brinkmann, C. Winckler, and U. Knierim. 2007. Reliability of a subjective lameness scoring system for dairy cows. Anim. Welfare 16:127-129.

Brown, L. D., T. T. Cai, A. DasGupta, A. Agresti, B. A. Coull, G. Casella, C. Corcoran, C. Mehta, M. Ghosh, and T. J. Santner. 2001. Interval estimation for a binomial proportion. Stat. Sci. 16:101-133.

Bruijnis, M. R. N., H. Hogeveen, and E. N. Stassen. 2010. Assessing economic consequences of foot disorders in dairy cattle using a dynamic stochastic simulation model. J. Dairy Sci. 93:2419-2432.

Bruyere, P., T. Hetreau, C. Ponsart, J. Gatien, S. Buff, C. Disenhaus, O. Giroud, and P. Guerin. 2012. Can video cameras replace visual estrus detection in dairy cows? Theriogenology 77:525-530.

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.

Cha, E., J. A. Hertl, D. Bar, and Y. T. Grohn. 2010. The cost of different types of lameness in dairy cows calculated by dynamic programming. Prev. Vet. Med. 97:1-8.

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.

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.

Cohen, J. 1968. Weighted Kappa - nominal scale agreement with provision for scaled disagreement or partial credit. Psychol. Bull. 70:213-220.

Cook, N. B. 2003. Prevalence of lameness among dairy cattle in Wisconsin as a function of housing type and stall surface. J. Am. Vet. Med. Assoc. 223:1324-1328.

Danscher, A. M., J. M. D. Enemark, E. Telezhenko, N. Capion, C. T. Ekstrom, and M. B. Thoefner. 2009. Oligofructose overload induces lameness in cattle. J. Dairy Sci. 92:607-616.

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.

Dippel, S., M. Dolezal, C. Brenninkmeyer, J. Brinkmann, S. March, U. Knierim, and C. Winckler. 2009a. Risk factors for lameness in cubicle housed Austrian Simmental dairy cows. Prev. Vet. Med. 90:102-112.

Dippel, S., M. Dolezal, C. Brenninkmeyer, J. Brinkmann, S. March, U. Knierim, and C. Winckler. 2009b. Risk factors for lameness in freestall-housed dairy cows across two breeds, farming systems, and countries. J. Dairy Sci. 92:5476-5486.

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.

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

Fleiss, J. L., J. Cohen, and B. S. Everitt. 1969. Large-sample standard errors of kappa and weighted kappa. Psychol. Bull. 72:323-327.

Flower, F. C., M. Sedlbauer, E. Carter, M. A. G. von Keyserlingk, D. J. Sanderson, and D. M. Weary. 2008. Analgesics improve the gait of lame dairy cattle. J. Dairy Sci. 91:3010-3014.

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.

Garbarino, E. J., J. A. Hernandez, J. K. Shearer, C. A. Risco, and W. W. Thatcher. 2004. Effect of lameness on ovarian activity in postpartum Holstein cows. J. Dairy Sci. 87:4123-4131.

Gibbons, J., E. Vasseur, J. Rushen, and A. M. de Passille. 2012. A training programme to ensure high repeatability of injury scoring of dairy cows. Anim. Welfare 21:379-388.

Green, L. E., V. J. Hedges, Y. H. Schukken, R. W. Blowey, and A. J. Packington. 2002. The impact of clinical lameness on the milk yield of dairy cows. J. Dairy Sci. 85:2250-2256.

Hoffman, A. C., D. A. Moore, J. R. Wenz, and J. Vanegas. 2013. Comparison of modeled sampling strategies for estimation of dairy herd lameness prevalence and cow-level variables associated with lameness. J. Dairy Sci. 96:5746-5755.

Ito, K., M. A. G. von Keyserlingk, S. J. LeBlanc, and D. M. Weary. 2010. Lying behavior as an indicator of lameness in dairy cows. J. Dairy Sci. 93:3553-3560.

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.

March, S., J. Brinkmann, and C. Winkler. 2007. Effect of training on the inter-observer reliability of lameness scoring in dairy cattle. Anim. Welfare 16:131-133.

Martin, P. and P. Bateson. 1993. Measuring behaviour: an introductory guide. 2nd edition ed. Cambridge University Press, Cambridge.

McCullagh, P. and J. A. Nelder. 1989. Generalized linear models. 2nd. Ed. ed. Chapman & Hall, London.

Nordlund, K. V., N. B. Cook, and G. R. Oetzel. 2004. Investigation Strategies for Laminitis Problem Herds. J. Dairy Sci. 87:E27-E35.

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. Welfare 12:605-610.

Rogowitz, B. E., T. N. Pappas, and J. P. Allebach. 2001. Human vision and electronic imaging. J. Electron. Imaging 10:10-19.

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.

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.

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.

Vasseur, E., J. Gibbons, J. Rushen, and A. M. de Passille. 2013. Development and implementation of a training program to ensure high repeatability of body condition scoring of dairy cows. J. Dairy Sci. 96:4725-4737.

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.

von Keyserlingk, M. A. G., A. Barrientos, K. Ito, E. Galo, and D. M. Weary. 2012. Benchmarking cow comfort on North American freestall dairies: Lameness, leg injuries, lying time, facility design, and management for high-producing Holstein dairy cows. J. Dairy Sci. 95:7399-7408.

Walker, S. L., R. F. Smith, D. N. Jones, J. E. Routly, M. J. Morris, and H. Dobson. 2010. The Effect of a Chronic Stressor, Lameness, on Detailed Sexual Behaviour and Hormonal Profiles in Milk and Plasma of Dairy Cattle. Reprod. Domest. Anim. 45:109-117.

Walker, S. L., R. F. Smith, J. E. Routly, D. N. Jones, M. J. Morris, and H. Dobson. 2008. Lameness, Activity Time-Budgets, and Estrus Expression in Dairy Cattle. J. Dairy Sci. 91:4552-4559.

Warnick, L. D., D. Janssen, C. L. Guard, and Y. T. Grohn. 2001. The effect of lameness on milk production in dairy cows. J. Dairy Sci. 84:1988-1997.

Warrens, M. J. 2013. Conditional inequalities between Cohen's kappa and weighted kappas. Stat. Methodol. 10:14-22.

Whay, H. 2002. Locomotion scoring and lameness detection in dairy cattle. In Practice 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 Agr. Scand. a-An. 30:103-107.

Chapter 4

Effect of merging levels of locomotion scores for dairy cows on intra- and interrater reliability and agreement

Published in Journal of Dairy Science (2014) 97:5533-5542

A. Schlageter Tello, 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

Abstract

Locomotion scores are used for lameness detection in dairy cows. In research, locomotion scores with 5 levels are used most often. Analysis of scores, however, is done after transformation of the original 5-level scale into a 4-, 3-, or 2-level scale to improve reliability and agreement. The objective of this study was to evaluate different ways of merging levels to optimize resolution, reliability, and agreement of locomotion scores for dairy cows. Locomotion scoring was done by using a 5-level scale and 10 experienced raters in 2 different scoring sessions from videos from 58 cows. Intra- and interrater reliability and agreement were calculated as weighted kappa coefficient (κw) and percentage of agreement (PA), respectively. Overall intra- and interrater reliability and agreement and specific intra- and interrater agreement were determined for the 5-level scale and after transformation into 4-, 3-, and 2-level scales by merging different combinations of adjacent levels. Intrarater reliability (κw) ranged from 0.63 to 0.86, whereas intrarater agreement (PA) ranged from 60.3 to 82.8% for the 5-level scale. Interrater κw = 0.28 to 0.84 and interrater PA = 22.6 to 81.8% for the 5-level scale. The specific intrarater agreement was 76.4% for locomotion level 1, 68.5% for level 2, 65% for level 3, 77.2% for level 4, and 80% for level 5. Specific interrater agreement
was 64.7% for locomotion level 1, 57.5% for level 2, 50.8% for level 3, 60% for level 4, and 45.2% for level 5. Specific intra- and interrater agreement suggested that levels 2 and 3 were more difficult to score consistently compared with other levels in the 5-level scale. The acceptance threshold for overall intra- and interrater reliability (κw and κ ≥0.6) and agreement (PA ≥75%) and specific intra- and interrater agreement (≥75% for all levels within locomotion score) was exceeded only for the 2-level scale when the 5 levels were merged as (12)(345) or (123)(45). In conclusion, when locomotion scoring is performed by experienced raters without further training together, the lowest specific intra- and interrater agreement was obtained in levels 2 and 3 of the 5-level scale. Acceptance thresholds for overall intra- and interrater reliability and agreement and specific intra- and interrater agreement were exceeded only in the 2-level scale.

4.1. Introduction

Locomotion scoring is a procedure used to indicate the quality of locomotion of cows. Raters assess gait and posture traits of cows and assign a locomotion score according to their judgment. Locomotion scores are often used to detect lameness in dairy cows (Whay, 2002; Flower and Weary, 2009). A cow is classified as lame when a predefined threshold on the scale is exceeded (Sprecher et al., 1997; Winckler and Willen, 2001; Chapinal et al., 2009).

Locomotion scores are sensitive to variation for intra- and interrater comparisons (Engel et al., 2003; O'Callaghan et al., 2003; Thomsen et al., 2008). Following Kottner et al. (2011),

reliability is defined as the capability of raters to differentiate between levels within the score (e.g., lame and non-lame), whereas agreement indicates the capability of raters to assign identical scores to the same cow. Reliability and agreement are important indicators of consistency and reproducibility of measurements (Martin and Bateson, 1993; Kottner et al., 2011). It is also stated that measurements with low reliability and agreement cannot be valid (Franzen, 2000). Reliability and agreement can be calculated by comparing data scores assigned to a cow by the same rater under similar conditions at different times (intra- rater 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, 1993). From a practical standpoint, high reliability and agreement for locomotion scores are important for generating consistent and comparable data for lameness control programs (DairyCo, 2007; Welfare Quality, 2009). In addition, locomotion scores are used as reference for calibration and validation in the development of different types of automatic locomotion scoring systems (Chapinal et al., 2010; de Mol et al., 2013; Viazzi et al., 2013).

Resolution is defined as the smallest change in locomotion that can be detected by the locomotion score and it is expressed in the number of levels of the scale (Martin and Bateson, 1993). A locomotion score with a multiple-level scale (and high resolution) is desirable because it would allow a better description of locomotion quality. In addition, a multiple-level locomotion score would allow users to take different actions with cows scored in different levels, as suggested for some locomotion scores (DairyCo, 2007). A large number of levels in a scale would provide more freedom to researchers and decision makers for data handling.

It is common practice to decrease the number of levels within a scale by merging adjacent levels to improve reliability or agreement (e.g., percentage of agreement). From a practical point of view, locomotion scores are also merged to create a binary classification of cows as lame or non-lame (Winckler and Willen, 2001; Channon et al., 2009; Main et al., 2010). However, no standard method yet exists for merging levels. Therefore, the decision as to which levels should be merged depends mainly on the criteria of the user of the locomotion score. When merging levels, resolution is lost from the locomotion score, a loss that tends to increase as fewer levels are used in the scale (Engel et al., 2003). To optimize reliability, agreement, and resolution of locomotion scores when levels are merged, it is important to understand the agreement in specific levels within the scale of a locomotion score. Thus, by knowing agreement of raters at each specific level, the level at which raters perform worst could be identified and merged.

To increase the practical value of locomotion scores and to support further development of automatic lameness detection systems, insight is needed in the reliability, agreement, and resolution of locomotion scores for dairy cows. Therefore, the objective of this study was

to evaluate different ways of merging levels to optimize resolution, reliability, and agreement of locomotion scores for dairy cows.

4.2. Materials and methods

4.2.1. Video recording

Video recording was performed at a dairy farm with 1,100 milking cows located in Israel and 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 NikonD7000 camera (Nikon Corp., Tokyo, Japan) equipped with a Nikkor DX AF-S 18–105 mm G ED lens (Nikon Corp.). The walking alley was situated at the exit of the milking area. 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 (.mov file format) had a resolution of 1,920 × 1,080 pixels at a frame rate of 25 frames per second. Camera settings were as follows: focal length = 18 mm, shutter speed = 1/40, aperture value = 3.5, and ISO speed: 5000. Because the video recordings were performed at night, external light sources were used to allow a clear observation of cows. To obtain individual video records of each cow, the video records were edited with Quick Time 7 Pro (Apple Inc., Cupertino, CA).

4.2.2. Locomotion score

Locomotion scoring was performed using a 5-level scale that was based on judging 5 gait and posture traits: asymmetric gait, arched back, reluctance to bear weight, tracking up, and head bob, as described by Flower and Weary (2006). In short, cows scored in level 1 had a smooth and fluid movement; cows in level 2 had an imperfect locomotion but were able to move freely; cows in level 3 had a compromised ability to move freely; for cows in level 4, the ability to move freely was obviously diminished; and for cows in level 5, the ability to move was severely restricted.

4.2.3. Video selection

Video records of all individual cows in the herd were stored in a video data set. Each video record was scored for locomotion according the previously described 5-level scale by 1 experienced rater [intrarater reliability/agreement: weighted kappa (κw) = 0.86/percentage of agreement = 84.5%] who did not participate in the experiment. Video records for each level within the 5-level scale were selected randomly from the video data set. A video record was included in the experiment only if the cow made at least 4 steps without stopping and sufficient contrast existed 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 a predetermined number of 12 video records per level was

reached. For level 5, only 8 video records were available that met the criteria. Therefore, 2 extra video records were added for level 3 because this level appeared to be the most difficult to assess consistently. The 58 video records selected were from 58 different cows. The number of video records used in the present experiment was determined using reporting reliability and agreement for locomotion scoring in dairy cows from the literature as reference (Flower and Weary, 2006; Thomsen et al., 2008; Channon et al., 2009) and to avoid fatigue of raters for scoring a large number of video records. In addition, because the similar number of videos in each level of the scale was an important part of the experimental design, the lack of video records classified as level 5 limited the total number of videos that could be included in the experiment.

4.2.4. Raters and scoring sessions

Locomotion scoring was performed by 10 experienced raters with different backgrounds and originally trained using different locomotion scores (Table 4.1). Raters were not informed about the objectives of the study, the number of different video records used, or the randomizations performed during the experiment.

Table 4.1. Background and experience of 10 raters participating in the study.

Rater	Group[a]	Background	Trained in locomotion score described by	Last scoring[b]
1	1	Researcher	Sprecher et al., 1997	Less than one year ago
2	1	Researcher	Manson and Leaver, 1988	Two years ago
3	1	Researcher	Manson and Leaver, 1988	Two years ago
4	1	Farmer/Researcher	Manson and Leaver, 1988	Two years ago
5	2	Veterinarian	Sprecher et al., 1997	Regularly
6	2	Technician	Winckler and Willen, 2001	Four years ago
7	3	Veterinarian	Welfare Quality, 2009	Regularly
8	3	Veterinarian	Welfare Quality, 2009	Regularly
9	2	Researcher	Winckler and Willen, 2001	Four years ago
10	4	Researcher	Sprecher et al., 1997	Regularly

[a] Same number indicates that raters performed scoring sessions together

[b] Indicates how long ago the raters performed regularly locomotion scoring in relation to the start of the experiment.

The 58 video records were shown to the 10 raters in 2 scoring sessions separated by at least 4 d. Each scoring session was split in 6 parts, in which raters scored the 58 video records each time again for either locomotion score or 1 of the 5 gait and posture traits independently. Each part lasted approximately 30 min, including 10 min for instruction and 20 min for scoring. In both sessions, the raters received a short instruction on locomotion scoring or scoring one of the gait and posture traits at the start of each part. The instruction consisted of showing 2 video records per level of the 5-level scale. The

instruction was done by the experienced rater, who was also responsible for the selection of the video records. Video records used for instruction were not included in the experiment. The instruction was the only time during which raters were allowed to discuss scoring.

For instruction and scoring, video records were shown with a projector on a white screen. Every video record was shown twice. The scoring was performed using an online interface that stored scores from raters directly in a database. The order in which locomotion and gait and posture traits were shown was randomly chosen in every session. In addition, to avoid cow recognition, video records were shown in a different random order in each part. All randomizations were done using an online random number generator (www.random.org). For practical reasons, it was not possible to have all raters in the same room at the same time; therefore, the experiment was conducted with 4 groups (Table 4.1).

4.2.5. Statistical analysis

In the present study, only data from locomotion scores were analysed and presented. Data related to scoring of individual gait and posture traits will be presented in another article. Intra- and interrater reliability and agreement were calculated for the original 5-level scale and after merging different combinations of adjacent levels to create 4-, 3-, and 2-level scales. Intrarater reliability and agreement were calculated by comparing the scores from the same cow in 2 different sessions. For both sessions, interrater reliability and agreement were calculated by comparing the scores of the same cow assigned to 2 different raters.

Intra- and interrater reliability was calculated as κw (Cohen, 1968) for the 5-, 4-, and 3-level scales; the kappa coefficient (**κ**) was calculated for the 2-level scale (Cohen, 1960). Intra- and interrater agreement was expressed as exact percentage of agreement (**PA**) for 5-, 4-, 3-, and 2-level scales; 95% CI were calculated for κw and κ (Fleiss et al., 1969) and Clopper-Pearson CI (Brown et al., 2001) were calculated for PA. When expressed as κw and κ, reliability can be classified as follows: poor (κw and κ <0.00), slight (κw and κ = 0.00–0.19), fair (κw and κ = 0.20–0.39), moderate (κw and κ = 0.4–0.59), substantial (κw and κ = 0.6–0.79), or excellent (κw and κ = 0.8–1) (Landis and Koch, 1977). The commonly accepted threshold for good reliability is indicated at κw and κ ≥0.6 (Gibbons et al., 2012). The commonly accepted threshold for agreement is ≥75% (Burn and Weir, 2011).

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 a cross table including all pairwise comparisons for raters and sessions.

Cross tables used to calculate overall intra- and inter- rater reliability and agreement were used to calculate the percentage of specific agreement. Percentage of specific agreement is based on the concept of positive and negative agreement (Cicchetti and Feinstein, 1990). Specific agreement indicates the capability of raters to agree on a specific level of the scale. The specific intrarater agreement indicates the average in which a single rater agrees in scoring a cow in the same level in 2 sessions. Specific interrater agreement indicates the average in which 2 raters agree in scoring a cow in the same level in 2 sessions. The confidence limits for the specific agreement were calculated with the delta method as proposed by Graham and Bull (1998). No established acceptance threshold exists for specific intra- and interrater agreement. Therefore, the same acceptance threshold as for inter- and intrarater agreement (PA ≥75%) was used.

4.3. Results

4.3.1. Distribution of locomotion scores

The distribution of scores for 10 raters using the 5-level scale is shown in Table 4.2. All raters scored all 58 video records. However, because of practical issues, some data were missed. Thus, rater 1 scored 56 video records in session 2, rater 4 scored 56 video records in session 1 and 53 in session 2, and rater 6 scored 57 video records in session 1. We observed large variation between raters in the distribution of scores. Three to 18 video records were scored as level 1; between 13 and 24 were scored as level 2; between 8 and 18 were scored as level 3; between 6 and 15 were scored as level 4; and between zero and 9 were scored as level 5 (Table 4.2).

Table 4.2. Distribution of locomotion scores assigned with a 5-level scale by 10 raters

		Rater[a]									
Score	**0[b]**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Level 1	12	11	9	18	3	15	9	12	14	8	11
Level 2	12	15	22	17	16	16	18	21	24	20	13
Level 3	14	14	13	11	18	8	16	16	14	18	17
Level 4	12	13	13	13	15	11	12	9	6	12	12
Level 5	8	3	2	0	1	9	3	2	1	2	6

[a] Values are averages from sessions 1 and 2.

[b] Distribution of locomotion scores according the experienced rater selecting video records

4.3.2. Intra- and interrater reliability and agreement for five-level scale

Intra- and interrater reliability and agreement are shown in Table 4.3. Intrarater reliability ranged from 0.63 to 0.86; therefore, all raters exceeded the acceptance threshold for κw. Intrarater agreement ranged from 60.3 to 82.8%; the acceptance threshold for intrarater agreement was exceeded for raters 3, 8, 9, and 10.

Interrater reliability ranged from 0.51 to 0.84 in session 1 and from 0.28 to 0.82 in session 2. The acceptance threshold for interrater reliability was exceeded in 39 of 45 pairwise comparisons in session 1 and in 29 of 45 pairwise comparisons in session 2. Interrater agreement ranged from 43.1 to 81.8% in session 1 and from 22.6 to 75.8% in session 2. The acceptance threshold for interrater agreement was exceeded in 3 of 45 pairwise comparisons in session 1 and in 1 of 45 pairwise comparisons in session 2. Some pairwise comparisons exceeded the acceptance threshold for κw, even with PA values below 50% (e.g., comparison rater 4 and rater 5 in session 1).

Although each video record in the experiment was shown 12 times in each session (24 times in total), raters indicated no cow memorization when asked at the end of session 2.

Table 4.3. Intrarater reliability and agreement (in the diagonal) and interrater reliability and agreement for session 1 (over the diagonal) and session 2 (under the diagonal) for pairwise comparison of 10 raters for a locomotion score with a 5-level scale.

			Session 1									
	Rater	**Parameter[a]**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Session 2	1	κw	**0.73**	0.75	0.72	0.74	0.71	0.82	0.75	0.67	0.73	0.71
		PA	**66.1**	68.9	63.7	67.9	58.6	77.2	72.4	58.6	68.9	60.3
	2	κw	0.72	**0.77**	0.73	0.72	0.64	0.78	0.69	0.62	0.71	0.61
		PA	66.1	**72.4**	67.2	69.6	50.0	75.4	65.1	55.2	67.2	50.0
	3	κw	0.70	0.71	**0.82**	0.70	0.60	0.72	0.75	0.75	0.60	0.60
		PA	64.3	62.1	**77.6**	64.3	43.1	66.7	70.7	70.7	53.4	44.8
	4	κw	0.45	0.43	0.38	**0.63**	0.62	0.84	0.74	0.58	0.72	0.66
		PA	43.1	37.7	32.1	**64.7**	46.4	81.8	71.4	59.0	71.4	57.2
	5	κw	0.69	0.70	0.75	0.42	**0.78**	0.63	0.57	0.61	0.58	0.72
		PA	57.1	56.9	63.8	28.3	**67.2**	47.4	43.1	50.0	44.8	60.3
	6	κw	0.67	0.71	0.57	0.57	0.66	**0.70**	0.75	0.70	0.79	0.70
		PA	60.7	63.8	41.4	54.7	50.0	**63.2**	71.9	64.9	75.4	61.4
	7	κw	0.60	0.57	0.65	0.39	0.68	0.62	**0.66**	0.66	0.65	0.58
		PA	51.8	48.3	58.6	39.6	62.1	51.7	**60.3**	62.1	62.1	48.3
	8	κw	0.58	0.66	0.74	0.28	0.64	0.52	0.63	**0.79**	0.51	0.55
		PA	51.8	60.3	70.7	22.6	56.9	41.1	58.6	**77.6**	43.1	43.1
	9	κw	0.70	0.65	0.66	0.53	0.63	0.76	0.60	0.58	**0.85**	0.63
		PA	62.5	58.6	56.9	52.8	48.3	70.7	51.7	51.7	**82.8**	53.5
	10	κw	0.66	0.69	0.63	0.52	0.74	0.82	0.58	0.53	0.69	**0.86**
		PA	53.6	58.6	48.3	50.9	62.1	75.8	44.8	41.4	60.4	**81.0**

[a] Reliability is expressed as weighted kappa (κw) and agreement is expressed as percentage of agreement (PA).

4.3.3. Overall intra- and interrater reliability and agreement

Overall intrarater reliability and agreement for different 5-, 4-, 3-, and 2-level scales are shown in Table 4.4. Overall intrarater reliability exceeded the acceptance threshold for the 5- level scale and all different combinations for 4-, 3-, and 2-level scales. The overall intra-rater agreement acceptance threshold was exceeded in most of the 4-level scales, except for the combination 123(45), and all 3- and 2-level scales (Table 4.4). The CI for intrarater reliability showed no differences for most of the 5-, 4-, 3-, and 2-level scales (Table 4.4). Overall intrarater agreement tended to increase (by approximately 7 percentage points) every time the scale decreased by 1 level.

Table 4.4. Overall intra- and interrater reliability and agreement for the original 5-level scale (5L) and for transformation into 4- (4L), 3- (3L), and 2-level scales (2L)[a]

		Intrarater		Interrater	
Scale	Combination[a]	κw / κ (CI)	PA (CI)	κw / κ (CI)	PA (CI)
5L	-	0.77 (0.74 – 0.80)	71.4 (67.7 – 75.1)	0.65 (0.64 – 0.66)	57.1 (55.7 – 58.4)
4L	(12)345	0.79 (0.75 – 0.83)	80.2 (76.9 – 83.5)	0.67 (0.66 – 0.69)	69.7 (68.3 – 70.9)
	1(23)45	0.77 (0.73 – 0.81)	82.1 (78.9 - 85.2)	0.61 (0.60 – 0.63)	70.6 (69.3 – 71.8)
	12(34)5	0.75 (0.72 – 0.80)	78.1 (74.7 – 81.5)	0.63 (0.61 – 0.64)	67.4 (66.1 – 68.7)
	123(45)	0.77 (0.75 – 0.80)	73.3 (69.7 - 76.9)	0.67 (0.65 – 0.68)	62.1 (60.7 – 63.4)
3L	(12)3(45)	0.79 (0.75 – 0.82)	82.1 (78.9 – 85.2)	0.70 (0.69 – 0.72)	74.6 (73.4 – 75.8)
	1(23)(45)	0.76 (0.72 – 0.81)	84.0 (81.0 – 87.0)	0.64 (0.62 – 0.66)	75.5 (74.4 – 76.7)
	(12)(34)5	0.77 (0.73 – 0.83)	86.8 (84.1 - 89.6)	0.66 (0.64 – 0.68)	80.0 (79.0 – 81.1)
	12(345)	0.75 (0.71 – 0.79)	80.0 (76.7 – 83.3)	0.65 (0.64 – 0.67)	72.6 (71.4 – 73.8)
	1(234)5	0.73 (0.67 – 0.79)	89.3 (86.7 – 91.8)	0.53 (0.51 – 0.56)	81.7 (80.7 – 82.8)
	(123)45	0.80 (0.75 – 0.85)	90.9 (88.5 – 93.2)	0.64 (0.62 – 0.66)	83.6 (82.6 – 84.6)
2L	1(2345)	0.71 (0.64 – 0.79)	91.2 (88.9 – 93.5)	0.57 (0.54 – 0.60)	86.9 (86.0 – 87.8)
	(12)(345)	0.78 (0.72 – 0.83)	88.7 (86.1 – 91.2)	0.70 (0.68 – 0.72)	85.2 (84.2 – 86.2)
	(123)(45)	0.81 (0.75 – 0.86)	92.8 (90.7 – 94.9)	0.70 (0.67 – 0.72)	88.6 (87.7 – 89.5)
	(1234)5	0.79 (0.67 – 0.91)	98.1 (96.9 – 99.2)	0.42 (0.37 – 0.48)	94.7 (94.2 – 95.5)

[a] Reliability was expressed as weighted kappa (κw) or kappa (κ, for 2-level scale) coefficients, and agreement was expressed as percentage of agreement (PA) and 95% CI.

[b] Parentheses indicate levels merged from the original 5-level scale

Overall interrater reliability and agreement for 5-, 4-, 3-, and 2-level scales are shown in Table 4.4. The overall interrater reliability acceptance threshold was exceeded for the 5-level scale and most of the combinations for 4-, 3, and 2-level scales (Table 4.4). The interrater agreement acceptance threshold was exceeded for most of the 3- and 2-level scales (Table 4.4). The CI for interrater reliability indicated no differences for 5-level locomotion score and the different 4-, 3-, and 2-level scales, with 2 exceptions: for 3-level scale combination 1(234)5 and 2-level scale combination (1234)5 (Table 4.4). Overall

interrater agreement increased (by approximately 10 percentage points) every time the scale of the locomotion score decreased by 1 level.

4.3.4. Specific intra- and interrater agreement

The specific intra- and interrater agreement for the 5-level scale and the specific levels for different 4-, 3-, and 2-level scales are shown in Tables 4.5 and 4.6. The specific intrarater agreement for the original 5-level scale were 76% for level 1, 69% for level 2, 65% for level 3, 77% for level 4, and 80% for level 5 (Table 4.5). The CI indicated that level 3 presented lower specific intrarater agreement than level 4. Scales exceeding the acceptance threshold for specific intrarater agreement in all levels were a 4-level scale combination [1(23)45]; 3-level scale combinations [1(23)(45), (12)(34)5, 1(234)5, and (123)45]; and all 2-level scales.

The specific interrater agreement for the 5-level scale and different 4-, 3-, and 2-level scales are shown in Table 4.6. The specific interrater agreements for 5-level scales were 64.7, 57.5, 50.8, 60.0, and 45.2% for levels 1, 2, 3, 4, and 5, respectively (Table 4.6). The CI indicated that specific interrater agreement was lower for levels 3 and 5 than for levels 1, 2, and 4. Scales exceeding the acceptance threshold for specific intrarater agreement in all levels were the 2-level scale combinations (12) (345) and (123)(45).

Specific interrater agreement had similar values for level 1 (session 1 = 63.4%, session 2 = 65.8%) and level 4 (session 1 = 60.8%, session 2 = 59.2%). The CI suggest that specific interrater agreements for level 2 in session 1 (64.3%) and session 2 (49.8%) were different (Figure 1). The CI for specific interrater agreements for level 3 of session 1 (56.5%) and session 2 (45.1%) were different. Although specific interrater agreement for level 5 showed large variation for session 1 (41.9%) and session 2 (49.1%), the CI suggest no differences between sessions (Figure 4.1)

4.4. Discussion

In literature, reliability or agreement are usually reported briefly to indicate the level of training of the raters assessing locomotion scores (Rutherford et al., 2009; Barker et al., 2010; Ito et al., 2010). However, important information about the experimental methodology for correct interpretation of reliability and agreement estimates is commonly omitted; for example, the total number of cows and the number of cows assigned to each level of the scale, the communication allowed among raters, or randomizations performed during the experiment (Danscher et al., 2009; Katsoulos and Christodoulopoulos, 2009; Main et al., 2010). Reliability or agreement is usually estimated using the total number of animals on one or more farms, where the total number of cows in levels 1 and 2 is greater than the number of cows in levels 3, 4, and 5 (Winckler and Willen, 2001; Thomsen et al., .

Table 4.5. Specific intrarater agreement for individual levels and 95% confidence intervals (CI) of a five-level (5L) scale and the individual levels of a four- (4L), three- (3L), and two-level (2L) scales.

Scale	Combination[a]	Level 1 (CI)	Level 2 (CI)	Level 3 (CI)	Level 4 (CI)	Level 5 (CI)
5L	-	76.4 (70.1 – 82.7)	68.5 (63.0 – 74.0)	65.0 (58.6 – 71.5)	77.2 (71.2 – 83.2)	80.0 (68.4 – 91.6)
4L	(12)345	88.8 (86.1 – 91.6)	65.0 (58.6 – 71.5)	77.2 (71.2 – 83.2)	80.0 (68.4 – 91.6)	
	1(23)45	76.4 (70.1 – 82.7)	85.9 (83.0 – 88.8)	77.2 (71.2 – 83.2)	80.0 (68.4 – 91.6)	
	12(34)5	76.4 (70.1 – 82.7)	68.5 (63.0 – 74.0)	85.3 (82.0 – 88.6)	80.0 (68.4 – 91.6)	
	123(45)	76.4 (70.1 – 82.7)	68.5 (63.0 – 74.0)	65.0 (58.6 – 71.5)	85.5 (81.1 – 89.9)	
3L	(12)3(45)	88.8 (86.1 – 91.6)	65.0 (58.6 – 71.5)	85.5 (81.1 – 89.9)		
	1(23)(45)	76.4 (70.1 – 82.7)	85.9 (83.0 – 88.8)	85.5 (81.1 – 89.9)		
	(12)(34)5	88.8 (86.1 – 91.6)	85.3 (82.0 – 88.6)	80.0 (68.4 – 91.6)		
	12(345)	76.4 (70.1 – 82.7)	68.5 (63.0 – 74.0)	88.6 (85.9 – 91.4)		
	1(234)5	76.4 (70.1 – 82.7)	93.0 (91.3 – 94.8)	80.0 (68.4 – 91.6)		
	(123)45	95.2 (93.7 – 96.7)	77.2 (71.2 – 83.2)	80.0 (68.4 – 91.6)		
2L	1(2345)	76.4 (70.1 – 82.7)	94.6 (93.1 – 96.1)			
	(12)(345)	88.8 (86.1 – 91.6)	88.6 (85.9 – 91.4)			
	(123)(45)	95.2 (93.7 – 96.7)	85.5 (81.2 – 89.9)			
	(1234)5	98.9 (98.3 – 99.6)	80.0 (68.4 – 91.6)			

[a] Parenthesis indicates levels merged from the original five-level scale.

Table 4.6. Specific interrater agreement for individual levels and 95% confidence intervals (CI) of a five-level (5L) scale and the individual levels of a four- (4L), three- (3L), and two-level (2L) scales.

Scale	Combination[a]	Level 1 (CI)	Level 2 (CI)	Level 3 (CI)	Level 4 (CI)	Level 5 (CI)
5L	-	64.7 (62.2 - 67.2)	57.5 (55.4 – 59.5)	50.8 (48.5 – 53.2)	60.0 (57.3 – 62.5)	45.2 (39.6 – 50.7)
4L	(12)345	85.2 (84.2 – 86.3)	50.8 (48.5 – 53.2)	60.0 (57.5 – 62.5)	45.2 (39.6 – 50.7)	
	1(23)45	64.7 (62.2 – 67.2)	78.4 (77.2 – 79.6)	60.0 (57.3 – 62.5)	45.2 (39.6 – 50.7)	
	12(34)5	64.7 (62.2 – 67.2)	57.5 (55.4 – 59.5)	77.8 (76.5 – 79.1)	45.2 (39.6 – 50.7)	
	123(45)	64.7 (62.2 – 67.2)	57.5 (55.4 – 59.5)	50.8 (48.5 – 53.2)	77.1 (75.3 – 78.9)	
3L	(12)3(45)	85.2 (84.2 – 86.3)	50.8 (48.5 – 53.2)	77.1 (75.3 – 78.9)		
	1(23)(45)	64.7 (62.2 – 67.2)	78.4 (77.2 – 79.6)	77.1 (75.3 – 78.9)		
	(12)(34)5	85.2 (84.2 – 86.3)	77.8 (76.5 – 79.1)	45.2 (39.6 – 50.7)		
	12(345)	64.7 (62.2 – 67.2)	57.5 (55.4 – 59.5)	85.1 (84.1 – 86.2)		
	1(234)5	64.7 (62.2 – 67.2)	88.1 (87.3 – 88.8)	45.2 (39.6 – 50.7)		
	(123)45	92.4 (91.8 – 93.0)	60.0 (57.3 – 62.5)	45.2 (39.6 – 50.7)		
2L	1(2345)	64.7 (62.2 – 67.2)	91.9 (91.4 – 92.6)			
	(12)(345)	85.2 (84.2 – 86.3)	85.1 (84.1 – 86.2)			
	(123)(45)	92.4 (91.8 – 93.0)	77.1 (75.3 –78.9)			
	(1234)5	97.3 (96.9 – 97.6)	45.2 (39.6 – 50.7)			

[a] Parenthesis indicates levels merged from the original five-level scale.

2008; Channon et al., 2009). Reliability strongly depends on the distribution in the population sample (de Vet et al., 2006). Thus, when the total number of animals within the farm is used, reliability estimates may present values under the acceptance threshold, not due to an effect of rater but by the effect of the sample distribution (Hoehler, 2000; de Vet et al., 2006). A solution for this problem is to estimate reliability and agreement using similar numbers of individuals in each level of the scale or to report distribution of the population sample for a better interpretation of reliability estimates (Burn and Weir, 2011). In this regard, the methodology described herein may be used as a guideline for future studies using locomotion scoring or other indicators measured with visual scores. Further details for methodologies for reporting agreement and reliability were described by Kottner et al. (2011).

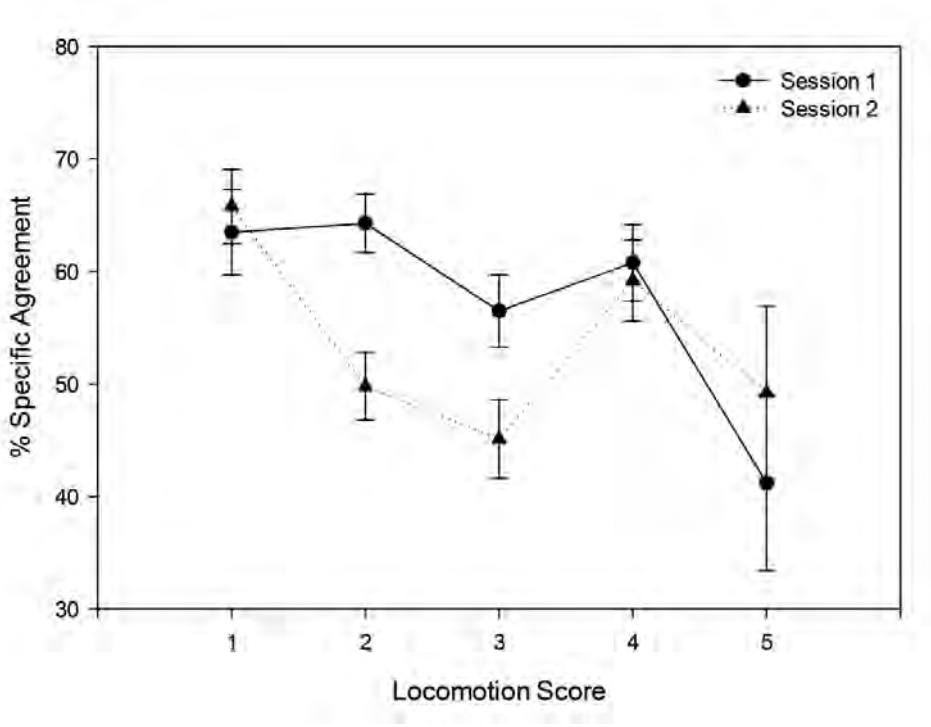


Figure 4.1. Specific interrater agreement in two sessions for a locomotion score with a five-level scale (bars indicate 95% confidence interval).

Different raters showed large variation in scoring the same cow for locomotion on a 5-level scale. In this regard, different backgrounds and initial training of raters might be factors

explaining the variation in scores. The 10 raters exceeded the acceptance threshold for intrarater reliability. This indicates that disagreements were mainly due to 1 level, whereas disagreements for 2 or 3 levels are less common (Winckler and Willen, 2001; O'Callaghan et al., 2003; Channon et al., 2009). Therefore, raters were able to differentiate properly between levels of the 5-level ordinal scale. Intrarater agreement was, in most cases, below the acceptance threshold, indicating that it is difficult even for experienced raters to obtain exact agreement in a 5-level scale. Values for intrarater reliability and agreement suggest that raters in the current experiment were experienced. Large ranges in values for interrater reliability and agreement indicate that, although raters were experienced, they did not have further training together, which is thought to be an important factor in interrater reliability and agreement (Kazdin, 1977; March et al., 2007).

Intrarater reliability values in the present experiment were higher than values reported for a similar locomotion score with a 5-level scale performed by experienced raters after a short training, with κ_W ranging from 0.38 to 0.64 (Thomsen et al., 2008). Intrarater agreement values in the present study were higher than values reported for a similar 5-level scale, where PA = 56% (O'Callaghan et al., 2003). Differences from other studies for intrarater reliability and agreement may be explained by the different levels of experience of the raters participating in different experiments (March et al., 2007; Gibbons et al., 2012) and the practical conditions in which the locomotion scoring was performed: scoring from video in the present study versus live scoring in other studies (O'Callaghan et al., 2003; Thomsen et al., 2008). Results obtained in the present experiment were similar to others reported in the literature for interrater reliability (Thomsen et al., 2008; Hoffman et al., 2013) and interrater agreement (Winckler and Willen, 2001; Katsoulos and Christodoulopoulos, 2009) for similar 5-level locomotion scores

Merging levels had no effect on the overall intra- and interrater reliability for most combinations, with some exceptions for the 3-level scale combination 1(234)5 and 2-level scale combinations 1(2345) and (1234)5, which presented κ_W and κ estimates lower than the acceptance threshold. This may be explained because merging 3 or 4 levels within the 5-level scale affected the distribution of the population sample, which also affected the reliability. Other authors reported an increment in interrater reliability estimates when expressed as κ coefficient when levels were merged from a 5- to a 2-level scale (March et al., 2007; Channon et al., 2009). However, κ coefficient is an inappropriate statistic to estimate reliability in ordinal scales (Kottner et al., 2011). Increment in interrater agreement has been re- ported previously when a locomotion score with 4-level scale was merged into a 2-level scale (Rutherford et al., 2009; Barker et al., 2010).

Relatively low values for specific intra- and interrater agreement in levels 2 and 3 for the original 5-level scale suggest that scoring of these 2 levels is difficult for experienced raters. This means that cows with slight locomotion alterations (or early stage lameness)

are difficult to identify, even by experienced raters. Winckler and Willen (2001) reported that the greatest variation (in a similar 5-level scale as that used in the present study) was between levels 1 and 2, which were also the levels in which most of the cows were scored, whereas in the present study, similar numbers of cows were present in all 5 levels of the scale. More uncertain is the explanation for the low specific interrater agreement in level 5, which might be due to the smaller number of video records in this level of the scale. However, the specific intrarater agreement for level 5 (80.0%) was almost twice as high as the specific interrater agreement for level 5 (45.2%), which indicates a disagreement between raters scoring level 5 of the scale, probably due to the lack of training of raters together. Low specific interrater agreement in level 5, however, has minor practical implications because the prevalence of cows scored as level 5 is commonly low in farms because cows are treated or culled before cows reach this severe level of alteration in locomotion (Engel et al., 2003; Thomsen et al., 2008; Channon et al., 2009).

In the current study, raters were not part of a strong training program to reach acceptable reliability and agreement. Training increases reliability and agreement of raters (March et al., 2007; Gibbons et al., 2012; Vasseur et al., 2013). Results in the present experiment suggest that evaluation of training on raters should be performed not only for PA and κ-like statistics but also for agreement in specific levels of the scale. No training program is available for locomotion scoring such as is available for other visual scores such as body condition score (Vasseur et al., 2013) or injury scoring (Gibbons et al., 2012). In training programs, it is important to consider that the response of raters to the training may vary, with raters performing better or worse after training (Engel et al., 2003). After being trained, raters should also have periodical additional training sessions to avoid the "drift effect," which is an unconscious drift from the original definitions of the observed characteristics (Kazdin, 1977) and to ensure acceptable reliability and agreement values over time. Under practical conditions, however, periodic training sessions are not always feasible because of cost or geographical distance. Therefore, the use of experienced raters without further training together is a realistic situation that may be faced in different programs using locomotion scoring for lameness control.

The selection of the best combination of levels to produce consistent and reproducible results for locomotion scoring should be based on acceptable reliability and agreement values but also on minimizing the loss of resolution associated with merging levels. Acceptance thresholds for intrarater reliability and agreement and specific intrarater agreement for all levels in the locomotion score was met in the 4-level scale with combination 1(23)45, suggesting that experienced raters were able to score locomotion consistently without an excessive loss of resolution. However, moderate overall interrater reliability and agreement and specific interrater agreement for locomotion score 1(23)45 acted as a limiting factor for the selection of this combination. Overall intra- and interrater reliability and agreement and specific intra- and interrater agreement were met only in the

2-level scales (12)(345) and (123)(45). Acceptable reliability and agreement values, however, were reached at maximum loss of resolution (2-level scale).

Because 2-level scales with combinations (12)(345) and (123)(45) had acceptable reliability, agreement, and specific agreement values, the selection of one combination would depend on different factors. One factor is related to the description of the lameness status of cows. In the literature, the 2-level combination (12)(345) is the most used to classify cow as non-lame (levels 1 and 2) and lame (levels 3, 4, and 5; Winckler and Willen, 2001; Katsoulos and Christodoulopoulos, 2009; Hoffman et al., 2013). The 2-level combination (123)(45) is also commonly used to classify cows as lame (levels 4 and 5; Bicalho et al., 2007a,b; Ito et al., 2010). It is common practice to use both locomotion scores [(12)(345) and (123)(45)] to describe lameness and severe lameness (Bicalho et al., 2007b; Ito et al., 2010). Another criterion to select the best combination [(12)(345) or (123)(45)] may be the capability to detect hoof lesions. Locomotion score (12)(345) presented the best sensitivity–specificity trade-off for the detection of painful lesions (defined as a reaction to pressure; Bicalho et al., 2007a).

A limitation of this study is the selection of arbitrary acceptance thresholds to classify reliability and agreement values as good. In this regard, κ-like statistics present a large range of acceptance thresholds from 0.4 (March et al., 2007; Burn and Weir, 2011) to 0.8 (Vasseur et al., 2013). Performing locomotion scoring under different practical conditions with an actual 4-, 3-, or 2-level scale might result in different agreement and reliability values than those obtained in the present study. In this regard, agreement and reliability reported in the present study for locomotion scores with 4-, 3-, and 2-level scales may be used only as guidelines.

4.5. Conclusions

When locomotion scoring was performed by experienced raters without further training together, specific intra- and interrater agreement had lower values for levels 2 and 3, suggesting that experienced raters had difficulties differentiating among these 2 levels. Accept- able overall intrarater reliability and agreement and specific intrarater agreement were achieved when the 5-level scale was transformed into a 4-level scale (levels 2 and 3 merged). However, acceptable overall interrater reliability and agreement and specific interrater agreement were exceeded only when the 5-level scale was transformed into a 2-level scale when levels were merged as (12)(345) or as (123)(45). Therefore, acceptable reliability and agreement values were obtained only with an important loss of resolution of locomotion scores.

Acknowledgments

This study is part of the Marie Curie Initial Training Network BioBusiness project (Marie Curie, Leuven, Belgium; FP7-PEOPLE-ITN-2008). The authors are very grateful to Jos Metz (Wageningen University, Wageningen, the Netherlands) and the reviewers for their valuable comments.

References

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.

Bicalho, R. C., S. H. Cheong, G. Cramer, and C. L. Guard. 2007a. Association between a visual and an automated locomotion score in lactating Holstein cows. J. Dairy Sci. 90:3294–3300.

Bicalho, R. C., F. Vokey, H. N. Erb, and C. L. Guard. 2007b. Visual locomotion scoring in the first seventy days in milk: Impact on pregnancy and survival. J. Dairy Sci. 90:4586–4591.

Brown, L. D., T. T. Cai, and A. DasGupta. 2001. Interval estimation for a proportion. Stat. Sci. 16:101–133.

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 Passillé, J. Rushen, and S. Wagner. 2010. Automated methods for detecting lameness and measuring analgesia in dairy cattle. J. Dairy Sci. 93:2007–2013.

Chapinal, N., A. M. de Passillé, 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 A. R. Feinstein. 1990. High agreement but low kappa: II. Resolving the paradoxes. J. Clin. Epidemiol. 43:551– 558.

Cohen, J. 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20:37–46.

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

DairyCo. 2007. DairyCo Mobility Score. Vol. 2011. DairyCo, Kenilworth, UK.

Danscher, A. M., J. M. D. Enemark, E. Telezhenko, N. Capion, C.T. Ekstrom, and M. B. Thoefner. 2009. Oligofructose overload induces lameness in cattle. J. Dairy Sci. 92:607–616.

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.

Fleiss, J. L., J. Cohen, and B. S. Everitt. 1969. Large-sample standard errors of kappa and weighted kappa. Psychol. Bull. 72:323–327.

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.

Franzen, M. D. 2000. Reliability and validity in neuropsychological assessment. Kluwer Academic/Plenum Publisher, New York, NY.

Gibbons, J., E. Vasseur, J. Rushen, and A. M. de Passillé. 2012. A training programme to ensure high repeatability of injury scoring of dairy cows. Anim. Welf. 21:379–388.

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. R. Wenz, and J. Vanegas. 2013. Comparison of modeled sampling strategies for estimation of dairy herd lameness prevalence and cow-level variables associated with lameness. J. Dairy Sci. 96:5746–5755.

Ito, K., M. A. G. von Keyserlingk, S. J. LeBlanc, and D. M. Weary. 2010. Lying behavior as an indicator of lameness in dairy cows. J. Dairy Sci. 93:3553–3560.

Katsoulos, P. D., and G. Christodoulopoulos. 2009. Prevalence of lameness and of associated claw disorders in Greek dairy cattle industry. Livest. Sci. 122:354–358.

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.

Main, D. C. J., Z. E. Barker, K. A. Leach, N. J. Bell, H. R. Whay, and W. J. Browne. 2010. Sampling strategies for monitoring lameness in dairy cattle. J. Dairy Sci. 93:1970–1978.

Manson, F. J., and J. D. Leaver. 1988. The influence of dietary protein intake and of hoof trimming on lameness in dairy cattle. Anim. Prod. 47:191–199.

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. 1993. Measuring Behaviour: An Introductory Guide. 2nd ed. Cambridge University Press, Cambridge, UK.

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.

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.

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., 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.

Vasseur, E., J. Gibbons, J. Rushen, and A. M. de Passillé. 2013. Development and implementation of a training program to ensure high repeatability of body condition scoring of dairy cows. J. Dairy Sci. 96:4725–4737.

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.

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 Practice. 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. 51:103–107.

Chapter 5

Relation between observed locomotion traits and locomotion score in dairy cows

Submitted

A. Schlageter Tello, 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

Abstract

Lameness is still an important factor in modern dairy farming. Human observation of locomotion is still used in practice. The objectives were to study relations between observed locomotion traits and locomotion scores in dairy cows, and if experienced raters are capable to score consistently individual traits that are used in locomotion scoring of cows. Locomotion and five traits (arched back, asymmetric gait, head bobbing, reluctance to bear weight and tracking up) were scored on a five-level scale for 58 videos of different cows. Videos were shown to ten experienced raters in two different scoring sessions. Relation between locomotion score and traits were estimated by two logistic regressions aiming to calculate the size of the fixed effects on a) the probability of scoring a cow in one of the five levels of the scale, and b) the probability of classifying a cow as lame (locomotion score ≥ 3). Fixed effects were rater, session, traits and interactions among fixed effects. Odd ratios were calculated to estimate the relative probability to classify a cow as lame when an altered trait was present (altered trait score ≥ 3). Overall intrarater and interrater reliability and agreement were calculated as weighted kappa coefficient (κw) and percentage of agreement (PA), respectively. Specific intrarater and interrater agreement for individual levels within a five-level scale were calculated. All traits were significantly related with locomotion when scored with a five-level scale and when classified in lame/non-lame. Odd ratios were 10.8 for reluctance to bear weight, 6.5 for asymmetric gait, 4.8 for arched back and head bobbing. Acceptance threshold for overall intrarater reliability (κw ≥ 0.60) was exceeded by locomotion scoring and all traits. Overall interrater reliability values ranged from κw = 0.53 for tracking up to κw = 0.61 for reluctance to bear weight. Intrarater and interrater agreement were below the acceptance threshold (PA < 75%). Most traits, however, tended to have lower specific intrarater and interrater agreement in level 3 and 5 of the scale. Considering the level of relation with locomotion scoring, intrarater and interrater reliability and agreement, traits to be used in practical conditions are reluctance to bear weight, asymmetric gait and arched back. Slight alterations in specific traits are difficult to detect, even by experienced raters.

5.1. 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 non-lame. 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, twelve traits have been described in locomotion scoring methods (Schlageter-Tello et al., 2014b). Depending on which locomotion scoring method is being used, raters have to evaluate between zero and seven traits. Most locomotion

scoring methods use between three and five 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 locomotion would allow to develop guidelines about which traits should have priority to assess by raters or to use individual traits instead of locomotion scores for an easy on-farm utilization (Chapinal et al., 2009; Thomsen, 2009). Another practical use of traits is related to the development of automatic locomotion scoring systems. 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). However, most of current automatic locomotion scoring systems focus on the measurement and analysis of only one trait (Schlageter-Tello et al., 2014b). 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). A different approach measures time and distance of variables associated to limb movement and some specific posture characteristics, such as tracking up (Song et al., 2008; Pluk et al., 2010), touch and release angle of hooves (Pluk et al., 2012), back curvature (Viazzi et al., 2013; Van Hertem et al., 2014) or gait variables such as asymmetry of step length, asymmetry of step time, asymmetry of step width, stance time, stride length (Maertens et al., 2011).

It is also important to know if human raters can identify and score locomotion traits consistently. Consistency is expressed as the reliability and agreement within and between raters (Martin and Bateson, 1993; Kottner et al., 2011). Reliability indicates the capability of raters to differentiate among levels within the score, 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 two or more raters assigned to the same cow under similar conditions (interrater reliability and agreement) (Martin and Bateson, 1993).

Taking the previously stated into consideration, the objectives were to study relations between observed locomotion traits and locomotion scores in dairy cows, and if experienced raters are capable to score consistently individual traits used in locomotion scoring of cows.

5.2. Materials and methods

Locomotion and five selected traits were scored using videos of cows walking through an alley. Video recording was performed at a dairy farm with 1100 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 five-level scale (described later) by one experienced rater who did not participate in the experiment. From the 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 four 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 a predetermined number of twelve video records per level were reached. For level 5, only eight 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 to select for the present experiment was determined using other articles reporting reliability and agreement for locomotion scoring in dairy cows (Flower and Weary, 2006; Thomsen et al., 2008; Channon et al., 2009) as reference, and taking into account that a too large number of video records would exhaust raters which would negatively affect the outcomes.

Locomotion scoring was performed using a five-level ordinal scale. Locomotion scoring was based on the judgment of five 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 nearly move. The five traits used to evaluate locomotion were: 1) asymmetric gait defined as differences of distance or time in the imprints between two 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 limb(s); 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 affected limb is lifted from the ground. The five traits were scored separately from locomotion using a similar five-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 5.1.

Locomotion scoring was performed by ten experienced raters with different backgrounds (six researchers, three veterinarians and one technician). They were originally trained using different locomotion scores. A detailed description of the raters can be found in Schlageter-Tello et al. (2014a). 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 ten raters in two scoring sessions in a different random order every time. Each scoring session was split in six parts in which raters scored the 58 video records each time again for either locomotion or one of the five traits separately. In both sessions, at the start of each part the raters received a short instruction on scoring locomotion or one of the traits. The instruction consisted on a description of the locomotion or trait to be scoring while showing two videos per level of the five-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 about scoring. A further explanation of the experimental design is described by Schlageter-Tello et al. (2014a).

5.2.1. Statistical Analysis

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

To establish the relation of traits with locomotion two generalized linear mixed models on a logistic scale were developed. The first model calculated probability of scoring a cow in each of the five levels of the locomotion score. This model comprised the fixed effects of session, traits scored as a five-level scale (arched back, asymmetric gait, head bobbing, reluctance to bear weight and tracking up) and interactions between raters*session, rater*trait and trait*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 model calculated the probability of classifying a cow as lame (locomotion score ≥ 3). The model included the same fixed effects as the first model. In the second model, traits were transformed into a binary scale.

Table 5.1. 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/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 hoof do not track up perfectly (approx. one hoof distance between track of front and hind hooves).
3	Capable locomotion but ability to move freely is compromised.	Arched back.	Short strides/step.	Head bobbing (associated to release of the affected limb from the floor).	Limp can be discerned.	Hind hooves do not track-up (approx. two hooves distance between track of front and hind hooves).
4	Ability to move freely is obvious diminished.	Obvious arched back.	Short and hesitant strides/step.	Obvious head bobbing (associated to 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 (approx. three 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/step.	Exaggerated head bobbing (associated to 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 (approx. four or more hooves distance between track of front and hind hooves).

The threshold to classify a cow with altered or no altered trait was score ≥ 3. 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 model, odd ratios were calculated to estimate the relative probability of classifying a cow as lame when cows show an altered trait when compared with a non-altered trait. Rater and cow were included as random effect in both models. In a logistic regression, the F-test value (F) indicates the relative size of the fixed effect on explaining the dependent variable (McCullagh and Nelder, 1989). The level of significance was established at $P < 0.05$. 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 five-level scales. Intrarater reliability and agreement were calculated comparing the scores from the same cow in two different sessions for ten raters. Interrater reliability and agreement were calculated comparing the scores of the same cow assigned for pairwise comparisons for two different raters. Interrater reliability and agreement were calculated for the two scoring sessions (N = 90).

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 (**κw**) (Cohen, 1968) using linear weighting (Cicchetti and Allison, 1971). Intrarater and interrater agreement was expressed as percentage of agreement (**PA**). The acceptance threshold for reliability values was stated at $\kappa w \geq 0.6$ and $\kappa w \geq 0.8$ indicating excellent reliability (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 traits score. Percentage of specific agreement is based on the 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 five-level scale in two 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). Since it has not been stated an acceptance threshold for specific intra and interrater agreement, it was set at ≥ 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).

5.3. Results

5.3.1. Distribution of scores

Relative distributions of scores assigned to locomotion and traits on a five-level scale are presented in Table 5.2. Distribution for each of the five levels had high variation for different raters (Table 5.2).

Table 5.2 Percentage of cows (N = 58) scored in each level of a five-level scale for locomotion (LS) and five traits: arched back (AB), asymmetric gait (AG), head bobbing (HB), reluctance to bear weight (RB), and tracking up (TU). Values were based on the scoring of ten raters in two sessions. Ranges of individual raters are given in parenthesis.

Trait	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)

5.3.2. Relation between traits and locomotion

For both, the probability to score a cow in one of the five levels of the scale or the probability of classify a cow as lame the biggest and significant effects were the five traits used to assess locomotion (Table 5.3 and 5.4). Significant interactions between rater and tracking up, and between rater and session were found when the probability of scoring a cow along five level locomotion score was used as dependent variable (Table 5.3). When the probability of classifying a cow as lame was used as dependent variable only the interaction rater and tracking up was significant. The highest odd ratio for traits was for reluctance to bear weight and the lowest for tracking up (Table 5.4).

Table 5.3. Size of the fixed effect rater (F-value), traits (arched back (AB), asymmetric gait (AG), head bobbing (HB), reluctance to bear weight (RB) and tracking up (TU)), and interactions between rater*TU and rater*session on the probability to score a cow in one of the five levels for locomotion scoring. Level of significance was established at $P < 0.05$.

Effect	F-value	p-values
AB	23.5	< 0.001
AG	15.7	< 0.001
HB	15.6	< 0.001
RB	30.4	< 0.001
TU	17.1	< 0.001
Rater*TU	1.9	0.03
Rater*Session	2.6	0.04

Table 5.4. Size of the fixed effect (F-value) and odd ratios for five traits namely arched back (AB), asymmetric gait (AG), head bobbing (HB), reluctance to bear weight (RB) and tracking up (TU) on the probability to classify a cow as lame. Level of significance was established at $P < 0.05$. CI indicates 95% confidence intervals.

Effect	F-value	P-values	Odd ratios[a] (CI)
AB	43.3	< 0.001	4.8 (3.0 – 7.7)
AG	49.8	< 0.001	6.5 (3.9 – 10.9)
HB	26.6	< 0.001	4.8 (2.6 – 8.7)
RB	77.8	< 0.001	10.8 (6.3 – 18.2)
TU	3.9	0.05	8.5 (1.0 – 71.9)
Rater*TU	2.4	0.01	-

[a] Indicates the relative probability for classifying a cow as lame when a trait is indicated to be altered (altered trait indicated when trait score was ≥ 3 on a five-level scale)

5.3.3. Overall reliability and agreement

Overall intra and interrater reliability and agreement for locomotion and the five traits are shown in Table 5.5. Overall intrarater reliability for locomotion scoring, tracking up, and head bobbing exceeded the acceptance threshold for κw (Table 5.5). However, when reliabilities are shown for individual raters, the acceptance reliability threshold was exceeded by all raters only for locomotion and head bobbing scoring, whereas 9, 7, 6 and 6 raters exceeded the substantial threshold for scoring arched back, reluctance to bear weight and asymmetric gait and tracking up, respectively. On the other hand, neither locomotion nor traits exceeded the acceptance threshold for overall interrater agreement (Table 5.5).

Table 5.5. Overall intrarater and interrater reliability (expressed as weighted kappa, κw) and agreement (expressed as percentage of agreement, PA) 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 five-level scale by ten raters in two sessions. Range indicates the values of κw and PA for each rater (intrarater, N = 10) and each pairwise comparison among all raters in two sessions (interrater, N = 90)

	Intrarater		Interrater	
	κw (Range)	PA (Range)	κw (Range)	PA (Range)
LS	0.77 (0.63 – 0.86)	71.4 (60.3 – 82.2)	0.65 (0.28 – 0.84)	57.1 (22.6 – 81.8)
AB	0.71 (0.59 – 0.83)	66.2 (55.1 – 79.3)	0.59 (0.40 - 0.77)	48.7 (31.0 - 70.7)
AG	0.67 (0.43 – 0.83)	61.8 (41.3 – 75.8)	0.58 (0.40 - 0.74)	51.5 (31.5 - 70.7)
HB	0.72 (0.65 – 0.81)	68.3 (63.7 – 75.4)	0.61 (0.35 - 0.75)	56.6 (41.3 - 71.9)
RB	0.69 (0.54 – 0.84)	60.8 (43.6 – 77.6)	0.61 (0.48 - 0.79)	53.9 (41.7 - 70.7)
TU	0.65 (0.43 – 0.91)	58.8 (39.6 – 86.2)	0.53 (0.25 - 0.77)	45.7 (22.8 - 67.2)

The acceptance threshold for overall interrater reliability was exceeded only for locomotion and the traits head bobbing and reluctance to bear weight (Table 5.5). For locomotion, 75% of pairwise comparisons among raters were above the acceptance threshold for reliability. 60% of κw values were above the substantial threshold for asymmetric gait, head bobbing and reluctance to bear weight. Of the κw values, 50% and 35% were above the acceptance threshold for asymmetric gait and tracking up, respectively. The acceptance threshold for overall interrater agreement was not exceeded by any of the five traits with PA values ranging from 45.7% (tracking up) to 57% (locomotion) (Table 5.5).

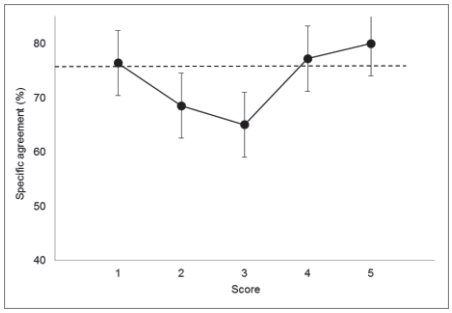
5.3.4. Specific agreement

Specific intrarater agreement and specific interrater agreement for individual levels of the five-level scale for locomotion and the five traits are shown in Figure 5.1 and 5.2. Acceptance threshold for specific intrarater agreement was exceeded by locomotion in level 1, 4 and 5 (Figure 5.1a).

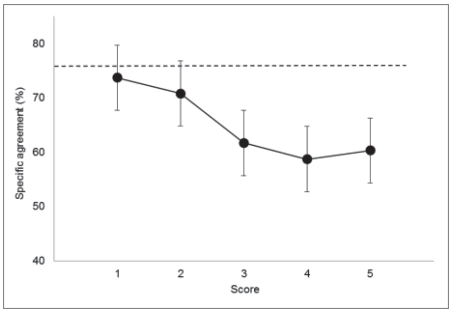
Specific intrarater and interrater agreement for specific levels of the scale resulted in different patterns for locomotion and the five traits assessed during the experiment. The pattern for specific intrarater agreement for locomotion, asymmetric gait, reluctance to bear weight and tracking up showed the highest values in level 1, 4 or 5, while 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 five levels of the scale (Figure 5.1). 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 (Figure 5.2).

5.4. Discussion

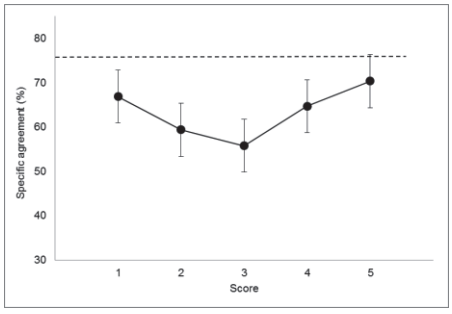
In the current study, all five traits had a significant effect on the probability of scoring a cow in one of the five levels of the locomotion score. The relationship between traits and locomotion in literature has been established by others using different approaches not directly comparable with the methodology used in the current experiment. Based on the assessment of two 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 nine-level scale. For head bobbing, tracking up, joint flexibility these correlation coefficients ranged from 0.70 - 0.80, and for arched back from 0.41 - 0.68 (Borderas et al., 2008; Chapinal et al., 2009). Abduction/adduction, defined as a tendency to rotate the limb outwards/inwards, 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 analyse relations between ten traits and a three level locomotion score assessed by 39 raters with different levels of experience. Asymmetric gait, reluctance to bear weight, arched back, and abduction/adduction had an effect for predicting locomotion score. Taking into account the current study and other studies all five traits, but especially reluctance to bear weight and asymmetric gait, are related with locomotion scoring. The strength of the relation between traits such as arched back, head bobbing, and tracking up and locomotion 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



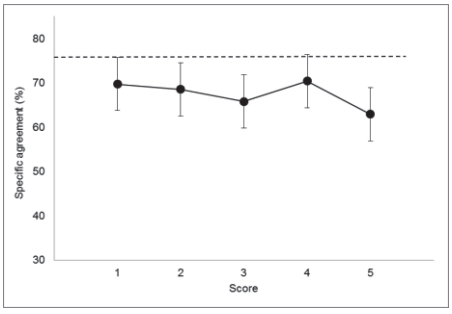
a. Locomotion



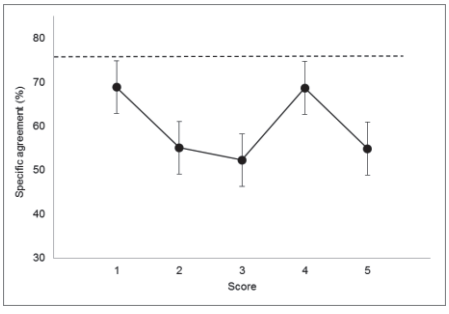
b. Arched back



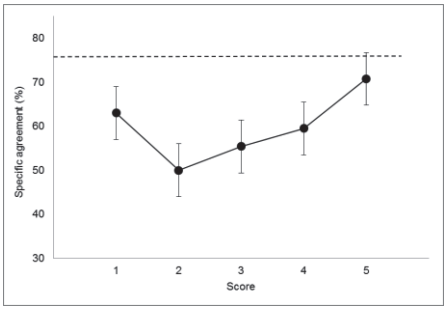
c. Asymmetric gait



d. Head bobbing

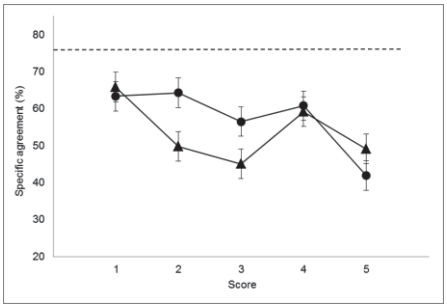


e. Reluctance to bear weight

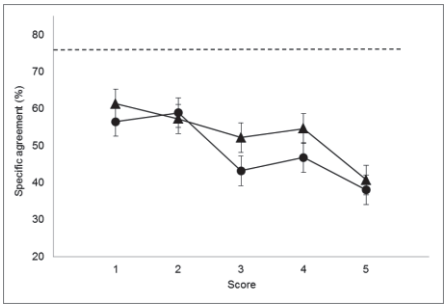


f. Tracking up

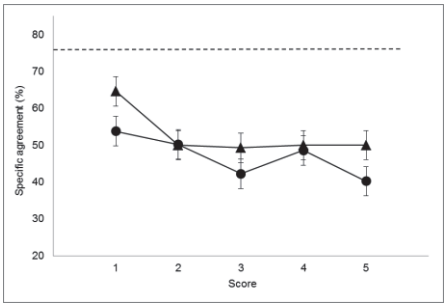
Figure 5.1. Specific intrarater agreement for locomotion (a) and five specific traits (b-f) used in locomotion scoring using a five-level scale. Bars indicate 95% confidence interval. Dotted lines indicate the level of the commonly accepted threshold for good agreement (75%).



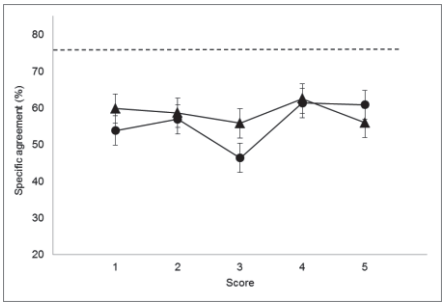
a. Locomotion



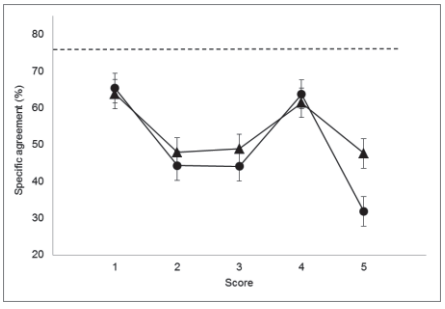
b. Arched back



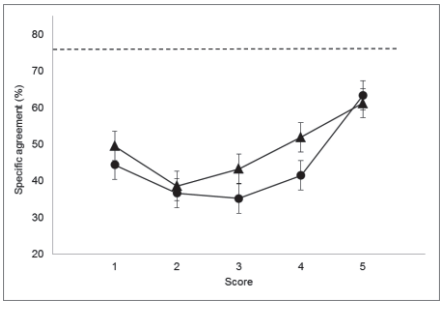
c. Asymmetric gait



d. Head bobbing



e. Reluctance to bear weight



f. Tracking up

Figure 5.2. Specific interrater agreement for locomotion (a) and five specific traits (b-f) used in locomotion scoring using a five-level scale in session 1 (●) and session 2 (▲). Bars indicate 95% confidence interval. Dotted lines indicate the level of the commonly accepted threshold for good agreement (75%).

estimate the relation between traits and locomotion. Although we gave the raters the task to score each trait and locomotion separately and independently, there were no ways of controlling for this in the current experiment.

In the current experiment, raters had different probabilities for scoring the locomotion of a cow in one of the five levels of the scale in different sessions, which is indicated by the significant interaction between rater and session. The interaction effect between raters and session, however, was smaller than the effect of traits indicating that effect of rater is minor on the probability for scoring a cow in one of the five levels of the scale. The significant effect for the interaction between rater and the trait tracking up indicates that different raters give different importance to tracking up on the probability of scoring a cow in one of the five levels of the locomotion score. Similarly, the interaction between raters and 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.

Odd ratios were included in the study to provide an alternative and probably easier explanation to the relation between traits and locomotion than F-values. Odd ratios indicate that when a trait is classified as altered the probability to classify a cow as lame increases. The trait which is mostly related with lameness seems to be reluctance to bear weight. Raters had about 11 times higher chance to classify a cow as lame when reluctance to bear weight was classified as altered than when it was not altered. When altered, asymmetric gait, arched back and head bobbing also increased the probabilities to classify a cow as lame. Odd ratio for tracking up indicates that cows showing altered tracking up increase their probability of being classified as lame. However, confidence intervals for tracking up were wide and contain value for odd ratio = 1. An odd ratio of 1 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).

Acceptance thresholds for overall intrarater reliability were exceeded for locomotion and all traits, whereas for overall interrater reliability the acceptance thresholds were exceeded by locomotion, head bobbing and reluctance to bear weight. Exceeding the acceptance threshold for κw indicates that disagreements amongst raters were mainly by one level difference and that disagreements with two or three levels were less frequent. This suggests that raters had an acceptable capability to differentiate among levels on a five-level ordinal scale. Acceptance thresholds for overall intrarater and interrater agreement were not exceeded by locomotion or traits. This indicates that even for experienced raters obtaining acceptable agreement values is difficult when using a five-level scale which is in agreement with previous studies (Winckler and Willen, 2001; Rutherford et al., 2009; Schlageter-Tello et al., 2014a). Few and not directly comparable results have been reported in the literature for reliability and agreement for specific traits. Flower and Weary (2006) and Borderas et al. (2008) reported acceptable inter and

intrarater reliability values (expressed as Pearson correlation coefficient, r ≥ 0.7) for tracking up, arched back, head bobbing and reluctance to bear weight. 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 inter- and intrarater reliability and agreement for the different traits.

Different patterns for the values of specific intrarater and interrater agreement for traits suggest that traits indeed were scored separately from locomotion and from other traits. 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 indicates 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.

Based on the relation between locomotion and traits, reliability and agreement and specific agreement we recommend to include the locomotion traits reluctance to bear weight, asymmetric gait and arched back in a locomotion score. Tracking up had acceptable overall reliability values and an effect on the probability of scoring a cow in each level of the five-level scale, but had low specific agreement in almost all levels of the scale (levels 1, 2, 3 and 4). In addition, the effect of the interaction rater*tracking up indicates that different raters assign different importance to tracking up in relation to locomotion scoring. Head bobbing was related with locomotion scoring, it showed acceptable overall reliability values and had similar specific agreement values across the five levels of the scale, indicating that raters were scoring head bobbing consistently. However, it has been reported that head bobbing was not frequently shown in cows classified as lame (Chapinal et al., 2009).

Most automatic locomotion scoring systems mimic locomotion scoring performed by humans. Due to 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 which is most related with 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 the results of the current study, automatic locomotion scoring 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). So far,

the performances of automatic locomotion scoring systems for lameness detection varied independent of the trait being measured (Schlageter-Tello et al., 2014b).

5.5. Conclusions

The five locomotion traits assessed in the current experiment (arched back, asymmetric gait, head bobbing, reluctance to bear weight and tracking up) were significantly related with the locomotion score. Raters had acceptable values for overall intrarater and interrater reliability and agreement for all five locomotion traits. Best traits to assess locomotion under practical conditions are reluctance to bear weight, asymmetric gait and arched back. Specific agreement for each level indicates that slight alterations in specific traits are difficult to detect by experienced raters.

Acknowledgement

This study is part of the Marie Curie Initial Training Network BioBusiness project (FP7-PEOPLE-ITN-2008). Thanks to Bas Engel for his valuable advice on the statistical analysis.

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 Passillé. 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 Passillé, 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.

Cook, T. D. 2002. Advanced Statistics: Up with Odds Ratios! A Case for Odds Ratios When Outcomes Are Common. Academic Emergency Medicine 9:1430-1434.

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.

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.

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. Biosyst. Eng. 110:29-39.

Martin, P. and P. Bateson. 1993. Measuring behaviour: an introductory guide. 2nd edition ed. Cambridge University Press, Cambridge.

McCullagh, P. and J. A. Nelder. 1989. Generalized linear models. 2nd. Ed. ed. Chapman & Hall, London.

Neveux, S., D. M. Weary, J. Rushen, M. A. G. von Keyserlingk, and A. M. de Passillé. 2006. Hoof discomfort changes how dairy cattle distribute their body weight. J. Dairy Sci. 89:2503-2509.

Pastell, M., L. Hanninen, A. M. de Passillé, 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.

Pluk, A., C. Bahr, A. Poursaberi, W. Maertens, A. van Nuffel, and D. Berckmans. 2012. Automatic measurement of touch and release angles of the fetlock joint for lameness detection in dairy cattle using vision techniques. J. Dairy Sci. 95:1738-1748.

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 Passillé. 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.

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. Agr. 64:39-44.

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. Biosyst. Eng. 119:108-116.

Van Nuffel, A., M. Sprenger, F. A. M. Tuyttens, and W. Maertens. 2009. Cow gait scores and kinematic gait data: can people see gait irregularities? Anim. Welfare 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.

Whay, H. 2002. Locomotion scoring and lameness detection in dairy cattle. In Practice 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 Agr. Scand. a-An. 30:103-107

.

Chapter 6

General discussion

A. Schlageter Tello

6.1. Introduction

Lameness is considered a major issue in dairy farming both from an economic and from an animal welfare perspective (Kossaibati and Esslemont, 1997; Flower and Weary, 2009; Bruijnis et al., 2010). Identifying lame cows, therefore, is important for both farmer and dairy cattle. Up until now, locomotion scoring of cows has been the preferred procedure for classifying cows as lame or non-lame. This is reflected in the fact that several locomotion scoring systems have been developed and incorporated into animal welfare assessment protocols and hoof health programs, and these could be used for example in, certification processes to assure animal welfare to consumers (Knierim and Winckler, 2009). Locomotion scoring can be useful as an on-farm management tool, allowing farmers to monitor and control lameness and identify related causal factors such as hoof or other limb lesions. Finally, locomotion scoring systems have been used to support the development of automatic locomotion scoring systems (e.g. the BioBusiness project).

The main objective of this thesis was to evaluate the performance of raters for assessing locomotion in dairy cattle in terms of reliability and agreement (Step 1, Figure 1.1). In order to achieve this goal a literature review was made to provide knowledge of the available locomotion scoring systems, and information concerning reliability and agreement values reported in the literature (Chapter 2). Experiments were performed to estimate reliability and agreement of raters when assessing locomotion and gait and posture traits (Chapters 3, 4 and 5).

Chapter 2, provides details of the different locomotion scoring systems. Several articles reported reliability, agreement and other consistency estimators for evaluating the performance of raters assessing locomotion. However, many articles applied inappropriate statistics or provided insufficient information for a sound interpretation of the findings. Chapter 3, established that experienced raters obtained better intrarater reliability and agreement values when performing locomotion scoring from video recordings than from live observation. However, the probability of classifying a cow as lame was the same for observations based on video recordings or live. Some other studies, described in Chapter 2, indicate an increase in consistency after the original multilevel scale was merged into a two-level scale for lame and non-lame classification. This was especially the case when percentage of agreement (PA) and prevalence adjusted bias adjusted kappa (PABAK) were used. Results presented in Chapter 4 showed an increment in agreement but not in reliability values when multilevel scales were merged into fewer levels, indicating that increment in agreement is due to chance. In Chapter 5, it was concluded that traits that were mainly associated to the locomotion scores assigned to a cow included: reluctance to bear weight, arched back and asymmetric gait. Finally, raters showed a limited performance for consistently scoring cows with slightly impaired locomotion (Chapter 4) and slightly impaired locomotion traits (Chapter 5).

The objective of the current chapter is to provide a deeper discussion of important topics that were briefly or not discussed in the previous chapters of this thesis. A general finding, from this thesis was the large variation in reliability, agreement and other consistency estimators obtained by raters when assessing locomotion. Several aspects that explain the variation of reliability and agreement have still to be discussed, e.g. the effect of the different procedures used to perform locomotion scoring, the factors affecting reliability and agreement associated to raters (e.g. training, experience and motivation), and the characteristic of the population sample (homogeneous and heterogeneous). This discussion also explores possibilities for the practical application of locomotion scoring systems as welfare indicators and their diagnostic value for hoof or other limb lesions. Finally, since this research has been conducted within the framework of the BioBusiness project, the usefulness of automatic locomotion scoring systems for classifying cows as lame and detection of hoof lesions and the possibilities of using them on-farm is also discussed.

This chapter will conclude with the general conclusions of this thesis.

6.2. Variation in reliability and agreement

One important issue of locomotion scoring systems is the variation in the scores assigned to a cow by different raters in the same session and by the same rater in repeated sessions. The variation is reflected in the wide range of values obtained for reliability and agreement and other estimators of consistency commonly used, such as correlation coefficients and PABAK. In cattle, variation in reliability and agreement values is not exclusive to locomotion scoring systems, but can also be found for injuries scoring (Gibbons et al., 2012), body condition scoring (Vasseur et al., 2013), scoring of qualitative behaviour assessment (Bokkers et al., 2012), and for behavioural indicators (Bokkers et al., 2009). In other species, similar problems for reliability and agreement occur, for example for locomotion scoring in sheep (Kaler et al., 2009), pigs (Dalmau et al., 2010; D'Eath, 2012), horses (Hewetson et al., 2006), and chickens (Garner et al., 2002), and body condition scoring (Phythian et al., 2012b), and different behavioural indicators in sheep (Phythian et al., 2012a), pigs (Dalmau et al., 2010) and horses (Burn et al., 2009).

Some of the most important factors affecting reliability and agreement when performing locomotion scoring are: the different procedures used to perform locomotion scoring, the characteristics of raters, and the characteristics of the population sample.

6.2.1. Effect of locomotion scoring procedure on reliability and agreement

The wide range of reliability and agreement values for locomotion scoring reported in different studies can be partially explained by the different procedures used for

locomotion scoring. For instance, Thomsen et al. (2008b) scored cows live on a five-level scale, walking on a dry slatted concrete floor, judging seven traits, observed from all possible angles for approximately one minute. Channon et al. (2009) scored cows live and from video recordings using a nine-level locomotion scoring system, judging five traits. During live observation cows were observed for between 30 and 60 seconds. Cows were observed laterally, caudally and when turning. Raters were allowed to move around during scoring. During video observations, two simultaneous video recordings were synchronized to facilitate observation of the cows laterally and caudally. Raters could view the videos as often as required. In Chapter 3, locomotion scoring was performed on cows walking on a solid concrete floor, covered with dry manure. Observations were performed live and from video recordings from a lateral view. Assessment was based on a five-level locomotion scoring system judging four traits. During live observation raters remained at the same point of observation throughout scoring. During video observations, the raters were only allowed to view the images once. These three examples serve to indicate the differences in locomotion scoring procedures. Different factors associated to different procedures affect the consistency of raters for scoring. For instance, it is well-known that the number of levels in the scale has an effect on agreement (Chapter 4). In Chapter 3, raters had a higher intrarater reliability and agreement when locomotion was assessed from video. The surface material on which the cows walk also has an effect on some kinematic gait characteristics (e.g. stride length and walking speed) which are observed during locomotion scoring (Telezhenko and Bergsten, 2005). Finally, time and observation perspective may influence rater performance for locomotion scoring.

Results from Chapters 4 and 5 indicate that even when locomotion scoring is performed using the same procedure, reliability and agreement values for scoring locomotion and locomotion traits varied considerably. Table 6.1, for example, shows values for interrater reliability, agreement and specific agreement for the three worst and three best pairwise comparisons (based on data from Chapter 4). Note that Rater 4 is present in both, the best and the worst pairwise comparisons (Table 6.1) which demonstrates once more the variation between raters. This confirms that regardeless of the impact that any selected procedure may have, the wide range in reliability and agreement values is mainly determined by factors associated to the raters.

Table 6.1. The three best and worst interrater reliability (expressed as weighted kappa, κw), agreement (expressed as percentage of agreement, PA) and specific agreement per rater using a five-level scale for locomotion scoring. (Based on data from Chapter 4).

		Reliability	Agreement					
	Comparison	κw	PA	Level 1	Level 2	Level 3	Level 4	Level 5
		(-)	(%)	(%)	(%)	(%)	(%)	(%)
Best	Raters 4 – 6	0.84	81.8	66.6	85.0	83.8	83.3	66.6
	Raters 1 – 6	0.82	77.2	73.6	83.3	75.8	75.0	80.0
	Raters 6 – 10	0.82	75.8	81.8	69.2	74.2	75.0	88.8
Overall	-	0.65	57.1	64.7	57.5	50.8	60.0	45.2
Worst	Raters 3 – 4	0.38	31.1	11.7	20.6	13.7	70.9	-
	Raters 4 – 5	0.42	28.3	14.2	19.3	29.6	57.1	0
	Raters 4 – 8	0.28	22.6	16.6	21.6	12.1	41.6	-

6.2.2. Factors associated to raters that influence reliability and agreement

There are several factors associated to raters that can affect reliability and agreement. One of the best ways to improve reliability and agreement of raters is training (Kaufman and Rosenthal, 2009). Several articles studying locomotion involved trained raters (Winckler and Willen, 2001; Flower and Weary, 2006; Flower et al., 2008). Usually, to become trained, raters are requested to participate in a training programme. The objective of the training programme is to establish similar opinion amongst raters about which locomotion score to assign to a cattle in order to improve reliability and agreement at acceptable levels (Martin and Bateson, 1993). Most training programmes comprise two parts (March et al., 2007; Gibbons et al., 2012; Vasseur et al., 2013). In the first part of the programme, raters are introduced to the chosen locomotion scoring system (e.g. scale and which traits to observe). Additionally, several examples of cows to be scored in different levels of the scale are presented. In the second part, raters are allowed to perform locomotion scoring under practical conditions, usually raters are allowed to question and comment on the scores assigned. An effective training program should aim to reach and maintain acceptable levels of reliability and agreement within and between raters and maintain this over time.

An important issue in a training programme is related to the selection of a statistical estimator to calculate the performance (consistency) of raters. As shown in Chapter 2, several studies use controversial consistency estimators such as PABAK and correlation coefficients (Flower and Weary, 2006; Brenninkmeyer et al., 2007). PABAK, for instance, corrects for the effect of prevalence and bias of raters on the κ coefficient. Low prevalence is a characteristic of a population samples that tends to be homogeneous. In this type of population sample, the probability of obtaining agreement by chance is high. Since the κ coefficient was designed to correct the effect of the expected agreement by chance on PA,

PABAK is actually corrects a useful characteristic of the κ coefficient (Vach, 2005). Additionally, PABAK corrects the effect of rater bias on the κ coefficient. Bias, in this case, is related to asymmetry in rater disagreement displayed in a cross-table. Bias, however, is an important indicator of disagreement, hence it should not be corrected (Hoehler, 2000). Correlation coefficients do not estimate reliability or agreement, but indicate linear associations among raters (Gallagher et al., 2003). Although PABAK and correlation coefficient values may increase as the level of training of raters also increases, the interpretation of such estimators as indicators for reliability and agreement is questionable and, therefore, should not be used to assess the level of training of raters.

In addition to using inappropriate reliability or agreement estimators, many training programmes base conclusions on the effectiveness of the training exclusively on reliability estimators, i.e. weighted kappa (κw) (Gibbons et al., 2012; Vasseur et al., 2013). Using only reliability estimators to assess training quality may be misleading because such values are influenced by both rater performance and the characteristics of the population sample (de Vet et al., 2006; Kottner et al., 2011). Agreement estimators (e.g. PA and specific agreement) are complementary to reliability estimators and therefore should be included for a better assessment of rater performance to score locomotion. For instance, high agreement and reliability indicates excellent performance for assigning the same score to the same cow in repeated measurements and level differentiation within scale. A low reliability and a high agreement indicates good agreement at some levels within the scale but poor agreement in others, resulting in a poor differentiation between levels within the scale. Specific agreement provides a useful indication of those levels at which raters performance is worst. It is possible, that reliability has acceptable levels with low agreement (see Table 4.3 for a comparison between raters 5 and 6). This combination appears to be an artefact of κw coefficient (Graham and Jackson, 1993). An additional issue of using only reliability estimators to assess training quality is related to the weighting used to calculate κw. When a quadratic weighting is applied, κw values tend to be higher than when a linear weighting is used. In the current thesis, linear weighting was applied because it is considered to be a more conservative estimator, whereas a quadratic weighting tend to overestimate the performance of raters (Warrens, 2013). Thus, as discussed in Chapter 4, decisions for the level of training of raters should be done taking into account both, reliability (κw and κ coefficients) and agreement estimators (PA and specific agreement) with a clear description of how the estimators were calculated.

The decision concerning when a rater is considered to be trained is commonly taken when a certain threshold for reliability and agreement is achieved (Martin and Bateson, 1993). There is no commonly accepted threshold for reliability estimators. In Chapters 3, 4 and 5, two different thresholds were used for considering reliability values as acceptable (κw and $\kappa \geq 0.4$ in Chapter 3 and ≥ 0.6 in Chapters 4 and 5). An acceptance threshold for κw and $\kappa \geq 0.8$ was proposed by Vasseur et al. (2013). The acceptance threshold usually

recommended for agreement estimators is ≥ 75%. Throughout the chapters of this thesis (Chapters 3, 4 and 5), the acceptance thresholds for reliability estimators were often exceeded, whereas the acceptance thresholds for agreement were rarely exceeded, especially for interrater comparison. In order to provide a guideline for acceptance thresholds for reliability and agreement estimators, two linear regressions were performed to calculate the equivalence between κw (using linear weighting) and PA considering two different scenarios: a) population sample tending to be homogeneous (data interrater comparisons from **Chapter 3**) and b) population sample tending to be heterogeneous (data interrater comparisons from **Chapter 4**) (Figure 6.1).

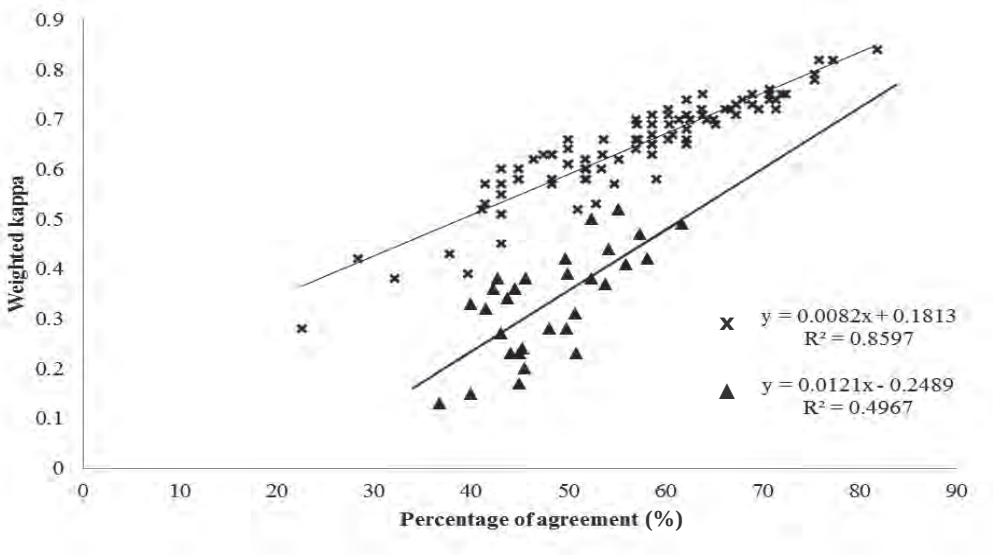


Figure 6.1. Linear regression between percentage of agreement and weighted kappa values for different pairwise comparisons for between raters for locomotion scoring on a five-level scale by experienced raters for two scenarios a) Population sample tending to be homogeneous (▲), data from Chapter 3 (Table 3.6), and b) Population sample tending to be heterogeneous (**x**). Data obtained from Chapter 4 (interrater pairwise comparisons from Table 4.3)

Table 6.2 shows the equivalence between κw and PA in population samples tending to be homogeneous and heterogeneous based on the linear regressions from Figure 6.1. When a threshold of 75% is used for PA, the recommended acceptance threshold for a κw should be ≥ 0.65, whereas in a heterogeneous population sample, recommended acceptance threshold for κw should be ≥ 0.8. Probably a good acceptance threshold for κw to be used in both population samples should be about 0.70 - 0.75. The acceptance threshold of κw ≥ 0.4 should not be used because it is associated with low agreement in both scenarios.

However, it was chosen in Chapter 3 due to it being applied in most studies using κw and κ (Brenninkmeyer et al., 2007; March et al., 2007; Burn and Weir, 2011) at the time that our experiment was performed. Although a PA ≥ 75% is considered the acceptance threshold for agreement, it is also important to estimate specific agreement in order to indicate rater performance at specific levels within the scale. The decision on whether or not these commonly used acceptance thresholds for reliability and agreement estimators suffice for practical application of locomotion scoring (e.g. welfare certification purpose, lameness management plan) remains in question.

Table 6.2. Equivalence between weighted kappa (κw) and percentage of agreement (PA) for different acceptance thresholds commonly referred to in literature for heterogeneous and homogeneous population samples. Equivalence calculated using linear regressions shown in Figure 6.1.

	Homogeneous	Heterogeneous
When PA =	κw =	κw =
75%	0.66	0.80
When κw =	PA =	PA =
0.4	54 %	27 %
0.6	70 %	51 %
0.8	87 %	75 %

Training alone is not enough to obtain acceptable levels of reliability and agreement. In Chapter 3, for instance**,** raters attended three training sessions and the experiment was conducted over three weeks, but no improvements were observed in reliability and agreement. Similarly, Engel et al. (2003) and Thomsen et al. (2008b) reported different response 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., 2008b). March et al. (2007) reported that during the training period (about 140 cows scored) raters obtained moderate reliability (κw = 0.52) and agreement (PA = 52%) values. Acceptable levels of reliability (κw = 0.69) and agreement (PA = 73%) were only achieved after 650 cows were assessed (March et al., 2007). This result indicates that besides training, experience of raters is a fundamental aspect in attaining acceptable reliability and agreement. Even after training and with experience, the maintenance of acceptable reliability and agreement levels 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). Therefore, regular update training sessions are required to prevent this from happening. The requirement for regular training affects the feasibility of an effective training programme, because it might demand too much effort from the raters to accomplish. Generally, under practical conditions, raters scoring locomotion are experienced, but do not get any refresher training. Performing studies using experienced

raters without further training (as in Chapters 3, 4 and 5) may provide reliability and agreement similar to those obtained by farmers, veterinarians or technicians applying locomotion scoring during their daily work.

Although not commonly mentioned in scientific literature, motivation is a factor that has an effect on the performance of raters to assess locomotion. Lack of motivation provides an explanation for the large variation in the reliability and agreement of Rater 4, discussed previously (Table 6.1). Additionally, lack of motivation may also explain incidences of lower reliability and agreement in session 2 in comparison to session 1 in the experiment described in Chapter 4 (Table 4.3). Thus, although training and experience increase the probability of better reliability and agreement, good performance of raters cannot be assured.

6.2.3. Characteristics of the population sample on reliability and agreement

As stated in the introduction, estimating reliability and agreement in a heterogeneous population sample allows a better evaluation of the performance of raters assessing locomotion by minimizing the expected agreement by chance. However, under practical conditions, population samples tend to be homogeneous with most cows scored in levels 1 and 2 within a five-level scale (Thomsen et al., 2008b). In this regard, it is relevant to evaluate the performance of raters assessing locomotion in both homogeneous and heterogeneous population samples. In this thesis, no specific study was done to evaluate the performance of raters assessing locomotion in different population samples. However, based on data reported within the chapters of this thesis and in the literature, some insight can be given on this issue. Table 6.3 shows reliability and agreement values from three different studies performing locomotion scoring. Within these studies four cases can be characterized as: A) a population sample tending to be heterogeneous with locomotion scoring performed by experienced untrained raters (results from Chapter 4); B) a population sample tending to be homogeneous and locomotion scoring performed by experienced and untrained raters (results from Chapter 4); C) a population sample tending to be homogeneous with locomotion scoring performed by experienced raters and trained raters (Winckler and Willen, 2001); and D) a homogeneous population sample with locomotion scoring performed by inexperienced untrained raters (results from Chapter 4).

When locomotion scoring was performed by experienced untrained raters, agreement had similar values for a heterogeneous (PA = 57%) and homogeneous (PA = 59%) population sample (Studies A and B, Table 6.3). Reliability, expressed as κ_W, for experienced untrained raters had a lower value in case B (homogeneous population sample, $\kappa_W = 0.52$) than in case A (heterogeneous population sample, $\kappa_W = 0.65$) (Table 6.3). Cases A and B in Table 6.3, confirm the findings stated in Chapter 1, a) reliability estimators are affected by the characteristic of the population sample and b) agreement is not affected by the

characteristics of the population sample. Although, case C shows that experienced trained raters displayed greater agreement (PA = 73.5%) when compared to cases A and B, part of this agreement may have been obtained by chance as indicated by the reliability (κw = 0.68), which had a similar value to case A (κw = 0.65, PA = 57.1 %, heterogeneous population sample). Note that for cases A and B (Table 6.3) reliability estimators are indeed affected by the characteristics of the population sample. However, cases B (κw = 0.52) and C (κw = 0.68) displayed different reliability in population samples tending to be homogeneous, indicating that raters are the most important factor reliability estimators (Table 6.3). The importance of raters in the value of reliability estimators is confirmed by the wide range of κw values obtained in cases B, C and D from Table 6.3. In these cases population samples tended to be homogeneous. Thus although the characteristics of the population sample indeed affect reliability estimators, the most important factor is the raters.

The range in κw values (0.52 to 0.68, Table 6.3) indicates that, regardless of the characteristic of the population sample, experienced raters have a moderate to good performance on the differentiation between levels of the scale. Specific agreements from case A show that raters had the most problems differentiating between levels 2 and 3 of the five-level scale, whereas in studies B and D raters worsened for differentiation between levels 3, 4 and 5 of the five-level scale (levels with fewer individuals). Specific agreement in case C shows that obtaining high specific agreements in those levels with fewer individuals is possible; however, experience and trained raters are required for this achievement. Regardless of the characteristics of the population sample and the level of experience and training of raters, the lowest specific agreement is commonly found in level 3 of the scale, indicating that this level is the most difficult to score consistently by raters.

Results from literature and chapters in this thesis show that raters have a large variation in reliability and agreement when scoring locomotion. This variation can be explained by different factors such as the lack of a standard procedure or the characteristics of the population sample (i.e. population samples tending to be homogeneous or heterogeneous) in which locomotion scoring is performed. However, the most important factor explaining the large variation in reliability and agreement is the raters themselves. Although the probability for obtaining acceptable reliability and agreement levels increases with training and experience, it is impossible to assure a consistent locomotion scoring in every session. Given the large variation in reliability and agreement, it is possible to conclude that raters only achieve a moderate performance in assessing locomotion in dairy cattle.

Table 6.3 Interrater reliability (expressed as κw) and agreement (expressed as PA and specific agreement) for locomotion scoring system with a five-level scale for four different cases using raters with different experience (Exp: experienced; Inexp: Inexperienced) and training level in population samples tending to be heterogeneous (Het) or homogeneous (Hom).

Case	Raters	Sample	Reliability	Agreement					
			κw (-)	PA (%)	Level 1 (%)	Level 2 (%)	Level 3 (%)	Level 4 (%)	Level 5 (%)
A[a]	Exp-untrained	Het	0.65	57.1	64.7	57.5	50.8	60.0	45.2
B[b]	Exp-untrained	Hom	0.52	59.0	64.8	68.4	43.7	48.0	53.8
C[c]	Exp-Trained	Hom	0.68	73.5	76.3	73.6	61.1	83.3	66.6
D[d]	Inexp-untrained	Hom	0.32	48.5	52.3	71.4	27.5	27.3	58.8

[a] Data from Chapter 4
[b] Calculated from data reported in Chapter 3, based on data from three experienced raters performing locomotion scoring from video
[c] Calculated from data reported by Winckler and Willen (2001)
[d] Calculated from data reported in Chapter 3, based on data from two inexperienced raters performing locomotion scoring from video

6.3. Practical application of locomotion scoring systems

The second part of this chapter will focus on the discussion of the practical application of locomotion scoring systems. Topics to be discussed include the applicability of locomotion scoring systems to a) consistently classify cows as lame (Step 2, Figure 1.1), and b) detect cows with hoof or other limb lesions (Step 3, Figure 1.1). This section will also discuss the performance of automatic locomotion scoring systems for lameness classification and hoof or other limb lesions detection and the potential for application of automatic locomotion scoring systems in dairy farming.

6.3.1. Usefulness of locomotion scoring systems to classify cows as lame

Locomotion scoring systems are mainly used to classify cows as lame (Step 2, Figure 1.1). A cow is classified as lame when the locomotion score assigned to a cow exceeds a predetermined threshold on a certain scale, commonly the middle level (Chapter 2).

A method to determine the usefulness of locomotion scoring systems as tool for lameness classification is the comparison of the lameness prevalence computed from locomotion scores assigned by the same raters to the same cows. This was attempted based on data from Chapter 4. Table 6.4 contains percentage of cows classified as lame based on locomotion scoring with a five-level scale using a threshold ≥ 3 for ten experienced raters. The ranges illustrate that there is large variation between and within raters for classifying cows as lame (Table 6.4).

A second method for determining the usefulness of locomotion scoring systems as a lameness classification tool for dairy cows is to estimate reliability and agreement of raters when the five-level scale is merged into a two-level scale.

Table 6.4. Percentage of cows classified as lame based on locomotion scoring of 58 cows from video recordings by ten experienced raters in two sessions. (Based on data from Chapter 4).

Rater	Session 1	Session 2
1	53.4	51.8
2	50.0	44.8
3	43.1	37.9
4	53.6	73.6
5	53.4	41.4
6	50.9	58.6
7	46.6	43.1
8	39.7	32.8
9	53.4	53.4
10	55.2	58.6

Table 6.5, presents reliability and agreement for the four cases shown in Table 6.3 for lame and non-lame classification. Table 6.5 shows variation between different studies to be dependent upon the raters and the characteristics of the population sample. Although the PA was > 80% in all four studies (higher than for those reported for the five-level scale, Table 6.3), it is likely that the increment in the agreement was due to chance. This is reflected by the similarity in κw values reported in Table 6.3 for the five-level scale and κ values two-level scale (Table 6.5) for the same cases (shown also in Chapter 4). Other studies displayed variation in reliability and agreement estimates for classification lame and non-lame, for example κ = 0.67 – 0.93, PA = 83.9 – 96.8% (Barker et al., 2010); κ = 0.79, PA = 88.3% (Channon et al., 2009); and κ = 0.72 – 0.80, PA = 96% (Hoffman et al., 2013). Most of these studies, however, do not report the characteristics of the population sample and the procedure used to perform locomotion scoring. Therefore, it is difficult to interpret these data accurately.

The range in κ coefficient values (range, κ = 0.38 – 0.70, Table 6.5) indicates that when a lame and non-lame classification is used, raters have a variable performance to differentiate between lame and non-lame cows. When a lame and non-lame classification is given in population samples with a lameness prevalence of 25% or lower (homogeneous population sample), the specific agreement for lame cows is lower than when calculated for a population sample with lameness prevalence of approximately 50% (heterogeneous population sample). This data suggests that when lameness prevalence is within the range of what can possible be found in practical farm conditions (a lameness prevalence of 15% - 25%) a consistent classification of lame cows is difficult. When locomotion scoring is done

by trained and experienced raters, it is possible to achieve acceptable specific agreement values (case C, Table 6.5).

Table 6.5. Reliability (expressed as κ), agreement (expressed as PA and specific agreement) for cows classified as lame and non-lame based on locomotion scores reported in four cases using raters with different experience (Exp: experienced; Inexp: Inexperienced) and training in population samples with different lameness prevalence.

Case	Rater	Lameness Prevalence	Reliability	Agreement		
			κ (-)	PA (%)	Non-lame (%)	Lame (%)
A[a]	Exp-untrained	≈ 50%	0.70	85.2	88.8	88.6
B[b]	Exp-untrained	≈ 24%	0.52	82.1	88.1	63.4
C[c]	Exp-Trained	≈ 17%	0.69	91.1	94.7	74.5
D[d]	Inexp-untrained	≈ 15%	0.38	83.7	90.3	47.5

[a] Data from Chapter 4
[b] Calculated from data reported in Chapter 3, based on data from three experienced raters performing locomotion scoring from video
[c] Calculated from data reported by Winckler and Willen (2001)
[d] Calculated from data reported in Chapter 3, based on data from two inexperienced raters performing locomotion scoring from video

Since lameness is commonly associated with impaired production (Green et al., 2002; Archer et al., 2010) and reproduction (Barkema et al., 1994; Walker et al., 2008), it has been stated that the detection of lameness in an early stage is an important task to minimize the negative impact on welfare and production of cows (Almeida et al., 2007; Van Nuffel et al., 2013). There is no clear definition of what can be considered early stage lameness. Since lameness is defined as impaired locomotion, mild or slightly impaired locomotion can be recognized as early stage lameness. Since cows are classified as lame when scored at 3 level or higher, early stage lameness can be found in cows that are scored with a 2. Performance of raters for classifying cows in an early stage may be estimated by calculating the intra- and interrater specific agreement of levels 2 and 3. As shown in Chapter 4, raters on average showed lower specific agreement for levels 2 and 3 than for the other levels in a heterogeneous population sample. A similar trend was found in data reported by Winckler and Willen (2001), in which lower specific agreements were found for levels 2 and 3 (Table 6.3) when compared with other levels. Lower agreement in level 2 and 3 suggests that raters have most difficulties differentiating between these two levels within the scale. From this it can be interpreted that raters have difficulties to detect early signs of lameness. As discussed in Chapter 5, raters also showed moderate performance for detecting slight alterations in different locomotion traits. It is important to note, however,

that there is a substantial variation in the values of specific agreements between raters as shown in Table 6.1.

Considering the findings presented in this section related to the variable consistency for classifying cows as lame, it can be said that locomotion scoring systems have a limited utility as tool for classifying lameness in cows consistently.

6.3.2. Locomotion scoring systems and hoof or other limb lesions

Although no specific part of this thesis has been dedicated to investigation of the relationship between locomotion scoring and lesions, it is a topic relevant to the practical application of locomotion scoring systems (Step 3, Figure 1.1). Based on data from literature some conclusions can be drawn concerning the relationship between locomotion scoring systems and lesions of hoof or other limbs lesions. Generally, hoof lesions refer to horn disruptions (e.g. white line disease and sole ulcer) or lesions in the skin surrounding the hoof (e.g. digital dermatitis) (Thomsen et al., 2012). Other limb lesions refer to lesions in other parts of the limbs, excluding the hooves.

Table 6.6 shows percentages of cows with hoof lesions at each level of a five-level locomotion scoring system. Schlageter-Tello et al. (2014) recorded the severity of hoof lesions on a four-level scale (Level 0: No lesion; Level 3: severe lesion) as described by Winckler and Willen (2001). Thomsen et al. (2012) recorded only severe hoof lesions based on the criteria of the rater. Bicalho et al. (2007) recorded painful lesions (defined as retraction of limb when digital pressure was applied to the lesion). Cows with higher locomotion scores (i.e. scores 3, 4 and 5) are more likely to have hoof lesions than cows with lower locomotion scores (i.e. scores 1 and 2), which confirms results of several previous studies (Sogstad et al., 2005; Frankena et al., 2009; Sogstad et al., 2012; Thomsen et al., 2012).

Table 6.6. Percentage of cow with a lesion in each level of a five-level locomotion scoring system.

	Level 1	Level 2	Level 3	Level 4	Level 5	Reference
Painful lesions	6	20	56	80	100	Bicalho et al. (2007)
Severe hoof lesions	19	28	45	65	85	Thomsen et al. (2012)[a]
Hoof lesions	34	52	62	82	50[b]	Schlageter-Tello et al. (2014)

[a] Data extracted from a graphic. Data reported in this table are approximated values
[b] Two cows were scored in level 5

Although cows scored with high locomotion scores (e.g. 3, 4 and 5) are more likely to have hoof lesions, locomotion scoring systems show a only moderate performance for detecting hoof lesions. This is reflected in the low to moderate sensitivity values obtained for detecting painful lesions (67%, Bicalho et al., 2007), sole ulcers (54%, Chapinal et al.,

2009) and hoof lesions in general (42.5%, Schlageter-Tello et al., 2014) when cows are classified as lame. An explanation for this fact is that sensitivity is not only affected by the capability for detecting true positives (i.e. cows classified as lame with hoof lesions), but also by the number of false negatives (i.e. cows classified as non-lame with hoof lesions). In this regard, a large number of false negatives can explain the moderate sensitivity of locomotion scores for detecting hoof lesions. A rough estimation for these false negatives may be obtained from Table 6.6 where 6% - 34% of the cows scored at level 1 and 20% - 52% of the cows scored at level 2 had at least one type of hoof lesion. Although the risk of having a hoof lesion increases as the locomotion score increases, the actual relationship of locomotion scores with hoof lesions is only moderate (Bicalho et al., 2007; Chapinal et al., 2009).

Other types of limb lesions may also be responsible for lameness in dairy cows. Although limbs may suffer different type of lesions, the most common are probably lesions in the tarsal or hock joint (Brenninkmeyer et al., 2013; Chapinal et al., 2014b; Heyerhoff et al., 2014). Hock lesions usually include: hairless zones, scabs, ulceration or swelling in the tarsal joint (Gibbons et al., 2012). Although hock lesions have been related to lameness in dairy cows (Brenninkmeyer et al., 2013), this relationship appears to be weaker than for hoof lesions (Potterton et al., 2011; Thomsen et al., 2012; Heyerhoff et al., 2014).

6.3.3. Locomotion scoring systems as a tool for animal welfare and hoof health protocols

Since lameness is considered to be an animal welfare problem, locomotion scoring systems are usually included in on-farm animal welfare assessment programmes (University of Bristol, 2004; Welfare Quality, 2009). Such animal welfare protocols provide some form of standardization for on-farm welfare assessment. Animal welfare protocols are used as certification tools for assuring animal welfare status of farms (Knierim and Winckler, 2009). Animal welfare assessment protocols should contain reliable and valid measurements. Using locomotion scoring systems to determine whether or not a cow is lame is open to discussion because of the variation in reliability and agreement values of raters. With the current average lameness prevalence commonly reported (about 20 - 25%), experienced raters appear to show only a moderate capability for classifying consistently cows as lame (Table 6.5, cases B and C). Although an efficient training programme may improve the reliability and agreement, it does not guarantee acceptable levels of consistency in all the locomotion scoring sessions due to the different procedures used perform locomotion scores (as discussed in **Chapter 3**) or due to factors associated to raters (training, experience, motivation). It should also be taken into account that lameness is only a visual sign of a possible underlying problem and not the problem itself (hoof or other limb lesions). Hence, lameness is useful as an animal welfare indicator if it is possible to detect accurately the cows with hoof or other limb lesions. However, as

discussed in the previous section, lameness has a moderate association with lesions. Given the variable performance of raters when performing locomotion scoring and the moderate relationship of locomotion scores with hoof and other limb lesions, locomotion scoring systems should not be included in protocols aiming to certify and assure animal welfare on farms.

Locomotion scoring systems are also included in several programs aimed at improving hoof health (DairyCo., 2007; Alberta Dairy Hoof Health Project, 2014). The fact that most cows classified as lame do indeed have a hoof lesion indicates that all cows classified as lame must receive treatment. The positive impact of a treatment for lame cows has been reported previously (Leach et al., 2012; Groenevelt et al., 2014). However, the large number of false negatives indicates that an effective strategy for control of lesions should not be exclusively based on lameness classification with locomotion scoring systems, but a combination of different actions aimed at preventing the occurrence of hoof and hock lesions.

There are several actions that can be performed along with locomotion scoring to reduce the prevalence of hoof or other limb lesions. Some preventive actions are related with farm design and may include the avoidance of cows walking constantly on hard surfaces, especially on slatted concrete floors (Barker et al., 2009; Fjeldaas et al., 2011), the provision of access to pastures (if possible) (Vermunt and Greenough, 1996; de Vries et al., 2015), and the provision of comfortable bedding to lie down. Deep sand bedding appears to be the best for minimizing the occurrence of both hoof and hock lesions (van Gastelen et al., 2011; Andreasen and Forkman, 2012; Chapinal et al., 2014a). Preventive actions related to herd management are: fixed-time hoof trimming (Manske et al., 2002; van der Tol et al., 2004), reducing incidences of ruminal acidosis by optimizing feed supply (Nordlund et al., 2004; Lean et al., 2013) and adding biotin to the ration (Hedges et al., 2001; Potzsch et al., 2003; Randhawa et al., 2008). From a genetic point of view, selecting sires based on hoof angles (van der Waaij et al., 2005; Onyiro et al., 2008) or claw health index (van der Linde et al., 2010) may decrease the prevalence of hoof lesions. Other helpful actions may include performing periodic footbaths (Teixeira et al., 2010; Fjeldaas et al., 2014) and examination for lesions during milking (Thomsen et al., 2008a; Relun et al., 2011).

Inclusion of a locomotion scoring system in hoof health programmes for dairy cows is recommended in combination with other actions aimed at controlling hoof and other limb lesions.

6.3.4. Automatic locomotion scoring systems

Another practical application of locomotion scoring systems is to serve as a "golden standard" for model calibration and validation for developers of automatic locomotion

scorings systems. However, variable reliability and agreement of locomotion scoring systems make a clear definition of a lameness case difficult, which affects the validity of automatic locomotion scoring systems for classifying cows as lame (Chapter 2).

Despite several limitations associated with automatic locomotion scoring systems, they still could be useful for practical on-farm utilization. Therefore, it is important to perform a deeper analysis on the performance of automatic locomotion scoring systems for classifying cows as lame (Step 2, Figure 1.1) and lesion detection (Step 3, Figure 1.1) when compared to raters performing locomotion scoring.

Most studies found in literature determine the performance of automatic locomotion scoring systems for classifying cows as lame or non-lame by calculating sensitivity and specificity using locomotion scoring systems as the "golden standard". Specific agreements for lame and non-lame are comparable to the terms sensitivity and specificity, respectively (Cicchetti and Feinstein, 1990). Therefore, specific agreement for lame and non-lame obtained by raters performing locomotion scoring (Table 6.5) can be compared with sensitivity and specificity obtained by automatic locomotion scoring systems for lame and non-lame classification when locomotion scoring is used as the "golden standard" (Table 2.6). Both manual and automatic locomotion scoring systems show variable performance for classifying cows as lame. As shown in Table 6.5, raters had variable values for specific agreement for lame cows ranging from 48% to 89% (equivalent to sensitivity) and a specific agreement for non-lame cows ranging from 89% to 95% (equivalent to specificity, Table 6.5). Similar results were obtained by automatic locomotion scoring systems when lameness classification was used as the golden standard with variable values for sensitivity (range = 40% – 86%) that tended to be lower than for specificity (range = 80% – 91%) (Table 2.6).

In Chapter 5, it was stated that most important locomotion traits when assessing locomotion scoring were reluctance to bear weight followed by arched back and asymmetric gait. In this regard it is expected that automatic locomotion scoring systems using a kinetic approach (comparable assessment of reluctance to bear weight) may perform better than automatic locomotion scoring systems using a kinematic approach (comparable to assessment of asymmetric gait and arched back) or the indirect approach (based on measurement of production and behaviour data) for lameness classification. Using the kinetic approach and force plates measuring force exerted on the floor, sensitivity was 51.9% and specificity was 88.4% (Liu et al., 2011). Using the kinematic approach and computer vision techniques for measuring arched back, sensitivity was 76% and specificity was 91% (Viazzi et al., 2013). Using the indirect approach, variable results were obtained by different authors using different behaviour and production data. Alsaaod et al. (2012) reported a sensitivity of 72% and a specificity of 81%. Kamphuis et al. (2013) showed sensitivities of 40% and 57% when specificities were fixed at 80% and 90%,

respectively. De Mol et al. (2013) and Van Hertem et al. (2013) reported a sensitivity and specificity > 85%. These results indicate that automatic locomotion scoring systems show a large variation for sensitivity and specificity regardless of the approach used. It is important to consider, however, that validation of automatic locomotion scoring systems was performed under controlled conditions and on a single farm.

Few articles reported a direct comparison between automatic locomotion scoring systems and manual locomotion scoring systems for detecting hoof or other limb lesions (Chapter 2). Bicalho et al. (2007) reported that raters had a higher sensitivity (automatic = 33.3%; raters = 67.2%) and similar specificity (automatic = 89.9%; raters = 84.6%) when compared to automatic locomotion scoring system based on the measurement of forces exerted on the floor by hooves as described by Rajkondawar et al. (2002) for detection of painful lesions (defined as retreatment of the limb when digital pressure was applied to the lesion). Recently, Schlageter-Tello et al. (unpublished data) reported that an automatic locomotion scoring systems based on measurement of back curvature using computer vision techniques, showed higher sensitivity (automatic = 58%; Rater = 43%) and lower specificity (automatic = 63%; rater = 78%) when compared with locomotion scoring for detecting hoof lesions (defined as lesions with a severity score ≥ 2 in a four-level scale).

Results indicated that automatic locomotion scoring systems show a similar performance as raters for classifying cows as lame and lesion detection. Since lack of time is the main factor that farmers give to have a proper management and control of lameness (Leach et al., 2010), automatic locomotion scoring systems could be a useful tool to help farmers in their task of monitoring lameness in dairy cows. Further research is required to estimate the usefulness of automatic locomotion scoring systems. Some recommended studies include estimating the performance of automatic locomotion scoring systems for lameness classification and hoof lesion detection on different farms and under practical conditions or studies related with the economic benefit of having an automatic locomotion scoring system instead of periodic locomotion scoring performed by humans.

Conclusions

In conclusion, this thesis shows that raters have a large variation in reliability and agreement when assessing locomotion. The variation is explained by different factors such as the lack of a standard procedure for assessing locomotion or the characteristics of the population sample that is assessed (i.e. population samples tending to be homogeneous or heterogeneous). The factor affecting reliability and agreement most, however, is the rater him/herself. Although the probability for obtaining acceptable reliability and agreement levels increases with training and experience, it is not possible to assure that raters score cows consistently in every scoring session. Given the large variation in reliability and

agreement, it can be concluded that raters demonstrate a moderate performance for assessing locomotion consistently in dairy cows.

Specific conclusions from this thesis are:

- Raters show variable performance for assessing locomotion in dairy cattle when expressed in terms of reliability and agreement.
- The complementary concepts reliability and agreement, as proposed by Kottner et al. (2011), are useful for a better interpretation of different statistical estimators for consistency. The weighted kappa and kappa coefficient are the preferred reliability estimators for ordinal and binary scales, respectively. Percentage of agreement and specific agreement are the preferred agreement estimators. Interpretation of reliability and agreement must be done taking into account the characteristics of the population sample (e.g. population samples tending to be homogeneous or heterogeneous).
- Experienced raters had better intrarater reliability and agreement when locomotion scoring is performed from video than by live observations. Video observations did not show any important influence on the probability of classifying a cow as lame. Video observations seem to be an acceptable method for assessing locomotion in dairy cows.
- Acceptance threshold for overall intrarater and interrater reliability ($\kappa w \geq 0.6$) and agreement and specific intrarater and interrater agreement ($\geq 75\%$) were exceeded only when a five-level scale is merged into a two-level scale. This increase in agreement, however, was due to chance. Raters showed moderate agreement when scoring slightly impaired locomotion.
- Raters showed variable performance when scoring gait and posture traits. Traits which are most related to locomotion scores are reluctance to bear weight, arched back and asymmetric gait. Raters showed moderate agreement when scoring slightly impaired traits.
- Locomotion scoring systems have a limited utility as a tool for classifying cows as lame.
- Locomotion scoring have a moderate relationship with hoof or other limb lesions. Poor performance for detecting hoof lesions is related to the large number of false negatives (e.g. cows classified as non-lame which have hoof lesions).
- Given the variability in the classification of cows as lame and the moderate association with hoof and other limb lesions, locomotion scoring systems should not be included in protocols aiming to certificate and assure animal welfare on farms. Although, it is recommended to include locomotion scoring systems in programs aiming to improve hoof health, locomotion scoring should never be used as a unique strategy for the management of hoof or other limb lesions.

- Automatic locomotion scoring systems have a similar variable performance as locomotion scoring systems for lameness classification and hoof or other limb lesions.

Recommendations

Automatic locomotion scoring systems have the potential to be a useful tool for helping farmers in management of lameness and hoof lesions in dairy cows. Most of the automatic locomotion scoring systems reported in the current thesis are in an experimental phase, and therefore not tested under practical conditions. Thus, it would be useful to evaluate the performance of automatic locomotion scoring systems in different farms under different practical conditions. This evaluation should include a comparison between manual and automatic locomotion scoring for lameness classification and hoof lesions detection. Additionally, it would be beneficial to investigate the economic impact of the inclusion of such technological tools in the dairy production chain.

Although manual and automatic locomotion scoring systems are useful tools for the control of hoof lesions, detection of some types of lesions is still an unsolved problem. Detection of certain type of lesion such as claw disruptions (e.g. sole ulcers) requires hoof trimming which is expensive and stresses the animals when performed periodically. Development of new technological tools, different from automatic locomotion scoring systems, could provide a solution, for example, for an accurate detection of claw disruptions. Claw disruptions trigger an inflammatory response in the affected hoof. The inflammatory response releases several biomarkers to the bloodstream that may be used for detecting claw disruptions. Determining specific biomarkers for claw disruptions and finding adequate sensors to detect these would be a first step for the creation of a technological tool aimed at improving the detection of this type of lesions.

References

Alberta Dairy Hoof Health Project. 2014. Dairy claw lesion identification. Access date: Jul 15, 2014.Web page:http://www.hoofhealth.ca/Section5/DD_Summer_2009_Lesion_ID.pdf.

Almeida, P. E., D. R. Mullineaux, W. Raphael, C. Wickens, and A. J. Zanella. 2007. Early detection of lameness in heifers with hairy heel warts using a pressure plate. Anim. Welfare 16:135-137.

Alsaaod, M., C. Romer, J. Kleinmanns, K. Hendriksen, S. Rose-Meierhofer, L. Plumer, and W. Buscher. 2012. Electronic detection of lameness in dairy cows through measuring pedometric activity and lying behavior. Appl. Anim. Behav. Sci. 142:134-141.

Andreasen, S. N. and B. Forkman. 2012. The welfare of dairy cows is improved in relation to cleanliness and integument alterations on the hocks and lameness when sand is used as stall surface. J. Dairy Sci. 95:4961-4967.

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.

Barkema, H. W., J. D. Westrik, K. A. S. Vankeulen, Y. H. Schukken, and A. Brand. 1994. The Effects of Lameness on Reproductive-Performance, Milk-Production and Culling in Dutch Dairy Farms. Prev. Vet. Med. 20:249-259.

Barker, Z. E., J. R. Amory, J. L. Wright, S. A. Mason, R. W. Blowey, and L. E. Green. 2009. Risk factors for increased rates of sole ulcers, white line disease, and digital dermatitis in dairy cattle from twenty-seven farms in England and Wales. J. Dairy Sci. 92:1971-1978.

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.

Bicalho, R. C., S. H. Cheong, G. Cramer, and C. L. Guard. 2007. Association between a visual and an automated locomotion score in lactating holstein cows. J. Dairy Sci. 90:3294-3300.

Bokkers, E. A. M., M. de Vries, I. Antonissen, and I. J. M. de Boer. 2012. Inter- and intra-observer reliability of experienced and inexperienced observers for the Qualitative Behaviour Assessment in dairy cattle. Anim. Welfare 21:307-318.

Bokkers, E. A. M., H. Leruste, L. F. M. Heutinck, M. Wolthuis-Fillerup, J. T. N. van der Werf, B. J. Lensink, and C. G. van Reenen. 2009. Inter-observer and test-retest reliability of on-farm behavioural observations in veal calves. Anim. Welfare 18:381-390.

Brenninkmeyer, C., S. Dippel, J. Brinkmann, S. March, C. Winckler, and U. Knierim. 2013. Hock lesion epidemiology in cubicle housed dairy cows across two breeds, farming systems and countries. Prev. Vet. Med. 109:236-245.

Brenninkmeyer, C., S. Dippel, S. March, J. Brinkmann, C. Winckler, and U. Knierim. 2007. Reliability of a subjective lameness scoring system for dairy cows. Anim. Welfare 16:127-129.

Bruijnis, M. R. N., H. Hogeveen, and E. N. Stassen. 2010. Assessing economic consequences of foot disorders in dairy cattle using a dynamic stochastic simulation model. J. Dairy Sci. 93:2419-2432.

Burn, C. C., J. C. Pritchard, and H. R. Whay. 2009. Observer reliability for working equine welfare assessment: problems with high prevalences of certain results. Anim. Welfare 18:177-187.

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.

Chapinal, N., Y. Liang, D. M. Weary, Y. Wang, and M. A. G. Von Keyserlingk. 2014a. Risk factors for lameness and hock injuries in Holstein herds in China. 97:4309-4316.

Chapinal, N., D. M. Weary, L. Collings, and M. A. G. von Keyserlingk. 2014b. Lameness and hock injuries improve on farms participating in an assessment program. Vet. J. 202:646-648.

Cicchetti, D. V. and A. R. Feinstein. 1990. High agreement but low kappa: II. Resolving the paradoxes. J. Clin. Epidemiol. 43:551-558.

D'Eath, R. B. 2012. Repeated locomotion scoring of a sow herd to measure lameness: consistency over time, the effect of sow characteristics and inter-observer reliability. Anim. Welfare 21:219-231.

DairyCo. 2007. DairyCo mobility score. Access date: May 10, 2011.Web page:http://www.dairyco.org.uk/.

Dalmau, A., N. A. Geverink, A. Van Nuffel, L. van Steenbergen, K. Van Reenen, V. Hautekiet, K. Vermeulen, A. Velarde, and F. A. M. Tuyttens. 2010. Repeatability of lameness, fear and slipping scores to assess animal welfare upon arrival in pig slaughterhouses. Animal 4:804-809.

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.

de Vries, M., E. A. M. Bokkers, C. G. van Reenen, B. Engel, G. van Schaik, T. Dijkstra, and I. J. M. de Boer. 2015. Housing and management factors associated with indicators of dairy cattle welfare. Prev. Vet. Med. 118:80-92.

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.

Fjeldaas, T., M. Knappe-Poindecker, K. E. Bøe, and R. B. Larssen. 2014. Water footbath, automatic flushing, and disinfection to improve the health of bovine feet. J. Dairy Sci. 97:2835-2846.

Fjeldaas, T., A. M. Sogstad, and O. Osteras. 2011. Locomotion and claw disorders in Norwegian dairy cows housed in freestalls with slatted concrete, solid concrete, or solid rubber flooring in the alleys. J. Dairy Sci. 94:1243-1255.

Flower, F. C., M. Sedlbauer, E. Carter, M. A. G. von Keyserlingk, D. J. Sanderson, and D. M. Weary. 2008. Analgesics improve the gait of lame dairy cattle. J. Dairy Sci. 91:3010-3014.

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.

Frankena, K., J. Somers, W. G. P. Schouten, J. V. van Stek, J. H. M. Metz, E. N. Stassen, and E. A. M. Graat. 2009. The effect of digital lesions and floor type on locomotion score in Dutch dairy cows. Prev. Vet. Med. 88:150-157.

Gallagher, A. G., E. M. Ritter, and R. M. Satava. 2003. Fundamental principles of validation, and reliability: rigorous science for the assessment of surgical education and training. Surg. Endosc. 17:1525-1529.

Garner, J. P., C. Falcone, P. Wakenell, M. Martin, and J. A. Mench. 2002. Reliability and validity of a modified gait scoring system and its use in assessing tibial dyschondroplasia in broilers. Br. Poult. Sci. 43:355-363.

Gibbons, J., E. Vasseur, J. Rushen, and A. M. de Passille. 2012. A training programme to ensure high repeatability of injury scoring of dairy cows. Anim. Welfare 21:379-388.

Graham, P. and R. Jackson. 1993. The analysis of ordinal agreement data - beyond weighted kappa. J. Clin. Epidemiol. 46:1055-1062.

Green, L. E., V. J. Hedges, Y. H. Schukken, R. W. Blowey, and A. J. Packington. 2002. The impact of clinical lameness on the milk yield of dairy cows. J. Dairy Sci. 85:2250-2256.

Groenevelt, M., D. C. J. Main, D. Tisdall, T. G. Knowles, and N. J. Bell. 2014. Measuring the response to therapeutic foot trimming in dairy cows with fortnightly lameness scoring. Vet. J. 201:283-288.

Hedges, J., R. W. Blowey, A. J. Packington, C. J. O'Callaghan, and L. E. Green. 2001. A longitudinal field trial of the effect of biotin on lameness in dairy cows. J. Dairy Sci. 84:1969-1975.

Hewetson, M., R. M. Christley, I. D. Hunt, and L. C. Voute. 2006. Investigations of the reliability of observational gait analysis for the assessment of lameness in horses. Vet. Rec. 158:852-858.

Heyerhoff, J. C. Z., S. J. LeBlanc, T. J. DeVries, C. G. R. Nash, J. Gibbons, K. Orsel, H. W. Barkema, L. Solano, J. Rushen, A. M. de Passille, and D. B. Haley. 2014. Prevalence of and factors associated with hock, knee, and neck injuries on dairy cows in freestall housing in Canada. J. Dairy Sci. 97:173-184.

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. R. Wenz, and J. Vanegas. 2013. Comparison of modeled sampling strategies for estimation of dairy herd lameness prevalence and cow-level variables associated with lameness. J. Dairy Sci. 96:5746-5755.

Kaler, J., G. J. Wassink, and L. E. Green. 2009. The inter- and intra-observer reliability of a locomotion scoring scale for sheep. Vet. J. 180:189-194.

Kamphuis, C., E. Frank, J. K. Burke, G. A. Verkerk, and J. G. Jago. 2013. Applying additive logistic regression to data derived from sensors monitoring behavioral and physiological characteristics of dairy cows to detect lameness. J. Dairy Sci. 96:7043-7053.

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.

Knierim, U. and C. Winckler. 2009. On-farm welfare assessment in cattle: validity, reliability and feasibility issues and future perspectives with special regard to the Welfare Quality (R) approach. Anim. Welfare 18:451-458.

Kossaibati, M. A. and R. J. Esslemont. 1997. The costs of production diseases in dairy herds in England. Vet. J. 154:41-51.

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.

Leach, K. A., D. A. Tisdall, N. J. Bell, D. C. J. Main, and L. E. Green. 2012. The effects of early treatment for hindlimb lameness in dairy cows on four commercial UK farms. Vet. J. 193:626-632.

Leach, K. A., H. R. Whay, C. M. Maggs, Z. E. Barker, E. S. Paul, A. K. Bell, and D. C. J. Main. 2010. Working towards a reduction in cattle lameness: 1. Understanding barriers to lameness control on dairy farms. Res. Vet. Sci. 89:311-317.

Lean, I. J., C. T. Westwood, H. M. Golder, and J. J. Vermunt. 2013. Impact of nutrition on lameness and claw health in cattle. Livest. Sci. 156:71-87.

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.

Manske, T., J. Hultgren, and C. Bergsten. 2002. The effect of claw trimming on the hoof health of Swedish dairy cattle. Prev. Vet. Med. 54:113-129.

March, S., J. Brinkmann, and C. Winkler. 2007. Effect of training on the inter-observer reliability of lameness scoring in dairy cattle. Anim. Welfare 16:131-133.

Martin, P. and P. Bateson. 1993. Measuring behaviour: an introductory guide. 2nd edition ed. Cambridge University Press, Cambridge.

Nordlund, K. V., N. B. Cook, and G. R. Oetzel. 2004. Investigation Strategies for Laminitis Problem Herds. J. Dairy Sci. 87:E27-E35.

Onyiro, O. M., J. Offer, and S. Brotherstone. 2008. Risk factors and milk yield losses associated with lameness in Holstein-Friesian dairy cattle. Animal 2:1230-1237.

Phythian, C. J., P. J. Cripps, E. Michalopoulou, P. H. Jones, D. Grove-White, M. J. Clarkson, A. C. Winter, L. A. Stubbings, and J. S. Duncan. 2012a. Reliability of indicators of sheep welfare assessed by a group observation method. Vet. J. 193:257-263.

Phythian, C. J., D. Hughes, E. Michalopoulou, P. J. Cripps, and J. S. Duncan. 2012b. Reliability of body condition scoring of sheep for cross-farm assessments. Small Rumin. Res. 104:156-162.

Potterton, S. L., M. J. Green, J. Harris, K. M. Millar, H. R. Whay, and J. N. Huxley. 2011. Risk factors associated with hair loss, ulceration, and swelling at the hock in freestall-housed UK dairy herds. J. Dairy Sci. 94:2952-2963.

Potzsch, C. J., V. J. Collis, R. W. Blowey, A. J. Packington, and L. E. Green. 2003. The impact of parity and duration of biotin supplementation on white line disease lameness in dairy cattle. J. Dairy Sci. 86:2577-2582.

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.

Randhawa, S., K. Dua, C. Randhawa, and S. Munshi. 2008. Effect of biotin supplementation on hoof health and ceramide composition in dairy cattle. Vet. Res. Commun. 32:599-608.

Relun, A., R. Guatteo, P. Roussel, and N. Bareille. 2011. A simple method to score digital dermatitis in dairy cows in the milking parlor. J. Dairy Sci. 94:5424-5434.

Schlageter-Tello, A., T. Van Hertem, S. Viazzi, E. Bokkers, P. Groot Koerkamp, C. Bites Romanini, M. Steensels, C. Bahr, I. Halachmi, D. Berckmans, and K. Lokhorst. 2014. Hoof lesion detection of dairy cows with manual and automatic locomotion scores. Page 158. In Proc. 65th Annual Meeting of the European Federation of Animal Science. Wageningen Academics. Copenhagen, Denmark.

Sogstad, A. M., T. Fjeldaas, and O. Osteras. 2005. Lameness and claw lesions of the Norwegian red dairy cattle housed in free stalls in relation to environment, parity and stage of lactation. Acta Vet. Scand. 46:203-217.

Sogstad, A. M., T. Fjeldaas, and O. Osteras. 2012. Locomotion score and claw disorders in Norwegian dairy cows, assessed by claw trimmers. Livest. Sci. 144:157-162.

Teixeira, A. G. V., V. S. Machado, L. S. Caixeta, R. V. Pereira, and R. C. Bicalho. 2010. Efficacy of formalin, copper sulfate, and a commercial footbath product in the control of digital dermatitis. J. Dairy Sci. 93:3628-3634.

Telezhenko, E. and C. Bergsten. 2005. Influence of floor type on the locomotion of dairy cows. Appl. Anim. Behav. Sci. 93:183-197.

Thomsen, P. T., I. C. Klaas, and K. Bach. 2008a. Short Communication: Scoring of Digital Dermatitis During Milking as an Alternative to Scoring in a Hoof Trimming Chute. J. Dairy Sci. 91:4679-4682.

Thomsen, P. T., L. Munksgaard, and J. T. Sorensen. 2012. Locomotion scores and lying behaviour are indicators of hoof lesions in dairy cows. Vet. J. 193:644-647.

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

University of Bristol. 2004. Bristol welfare assurance program: Cattle assessment, Version 2.0. University of Bristol, Bristol, UK.

Vach, W. 2005. The dependence of Cohen's kappa on the prevalence does not matter. J. Clin. Epidemiol. 58:655-661.

van der Linde, C., G. de Jong, E. P. C. Koenen, and H. Eding. 2010. Claw health index for Dutch dairy cattle based on claw trimming and conformation data. J. Dairy Sci. 93:4883-4891.

van der Tol, P. P. J., S. S. van der Beek, J. H. M. Metz, E. N. Noordhuizen-Stassen, W. Back, C. R. Braam, and W. A. Weijs. 2004. The effect of preventive trimming on weight bearing and force balance on the claws of dairy cattle. J. Dairy Sci. 87:1732-1738.

van der Waaij, E. H., M. Holzhauer, E. Ellen, C. Kamphuis, and G. de Jong. 2005. Genetic parameters for claw disorders in dutch dairy cattle and correlations with conformation traits. J. Dairy Sci. 88:3672-3678.

van Gastelen, S., B. Westerlaan, D. J. Houwers, and F. J. C. M. van Eerdenburg. 2011. A study on cow comfort and risk for lameness and mastitis in relation to different types of bedding materials. J. Dairy Sci. 94:4878-4888.

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 Nuffel, A., J. Vangeyte, K. C. Mertens, L. Pluym, S. De Campeneere, W. Saeys, G. Opsomer, and S. Van Weyenberg. 2013. Exploration of measurement variation of gait variables for early lameness detection in cattle using the GAITWISE. Livest. Sci. 156:88-95.

Vasseur, E., J. Gibbons, J. Rushen, and A. M. de Passille. 2013. Development and implementation of a training program to ensure high repeatability of body condition scoring of dairy cows. J. Dairy Sci. 96:4725-4737.

Vermunt, J. J. and P. R. Greenough. 1996. Sole haemorrhages in dairy heifers managed under different underfoot and environmental conditions. Br. Vet. J. 152:57-73.

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.

Walker, S. L., R. F. Smith, J. E. Routly, D. N. Jones, M. J. Morris, and H. Dobson. 2008. Lameness, Activity Time-Budgets, and Estrus Expression in Dairy Cattle. J. Dairy Sci. 91:4552-4559.

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. in Welfare Quality Consortium. Lelystad, The Netherlands.

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 Agr. Scand. a-An. 30:103-107.

Summary

Lameness is considered an important problem in modern dairy farming. Locomotion scoring systems are generally used to classify cows as lame or non-lame. Locomotion scoring systems are procedures used to evaluate the quality of the locomotion of cows. When scoring locomotion, raters focus their attention on gait and posture traits that are described in the protocol. Using these traits, raters assign a locomotion score to cows according to a pre-determined scale. A preselected threshold within the scale determines whether a cow is classified as lame or non-lame. These lame cows are commonly assumed to suffer pain due to, either hoof or other limb lesions. Therefore, locomotion scoring systems are also used to detect hoof or other limb lesions.

This thesis is part of a project aiming to develop an automatic locomotion scoring system. Automatic locomotion scoring systems use sensors, instead of raters, to collect on-farm cow locomotion data. Data from these sensors are analysed using mathematical algorithms to assess the locomotion of cows and to classify whether or not cows are lame. In general, locomotion scores from raters are used as reference or "golden standard" to validate automatic locomotion scoring systems. However, much remains unknown about the performance of raters to score locomotion consistently.

Although lameness is considered an important problem in modern dairy farming, it is important to remark, that lameness is only a visual indicator of a possible underlying problem (e.g. hoof or other limb lesions) and therefore not the actual problem. Thus, lameness is only useful when: a) raters are capable of consistently scoring locomotion, and b) cows classified as lame are indeed affected by hoof or other limb lesions. Locomotion scoring is particularly important because it is the animal based measurement on which lame or non-lame classification is based. The objective of this thesis is to evaluate the performance of raters to assess locomotion in dairy cattle in terms of reliability and agreement. This thesis explores possibilities for the practical application of locomotion scoring systems related with lameness classification and hoof lesions and other limb lesions. Since the research was conducted within the framework of the BioBusiness project, this thesis will discuss the usefulness of automatic locomotion scoring systems for on-farm application.

In Chapter 2, a literature review comprising 244 peer-reviewed articles was done. The objective of Chapter 2 was to describe, compare and evaluate agreement, reliability, and validity of (manual) locomotion scoring systems and automatic locomotion scoring systems used in dairy cattle lameness research. Twenty-five locomotion scoring systems were found. Locomotion scoring systems use different types of scale (ordinal or continuous) and different gait and posture traits to be observed. Fifteen automatic locomotion scoring systems were found that could be categorized into three approaches:

a) the kinetic; b) kinematic, and c) the indirect approach. Reliability and agreement estimates were scarcely reported in articles related to locomotion scoring systems. Some of the most frequently used locomotion scoring systems was poorly evaluated for reliability and agreement. Reliability and agreement estimates for the original locomotion scoring system and after lame/non-lame classification showed large ranges among and sometimes also within articles. Reliability and agreement estimates for automatic locomotion scoring systems were not reported in any article. Several automatic locomotion scoring systems use locomotion scoring systems as a reference for model calibration and validation. However, varying reliability and agreement estimates of locomotion scoring systems make a clear definition of a lameness case difficult, and thus affect the validity of automatic locomotion scoring systems. Both locomotion scoring systems and automatic locomotion scoring systems showed limited validity for hoof lesion detection. Long-term studies comparing locomotion scoring systems and automatic locomotion scoring systems while applying various strategies to prevent and control unfavourable conditions leading to impaired locomotion (e.g. hoof lesions) are required.

In Chapter 3, the objective of a practical experiment was to evaluate intrarater and interrater reliability and agreement of experienced and inexperienced raters for locomotion scoring performed live and from video, and to calculate the influence of raters and the method of observation (live or video) on the probability of classifying a cow as lame. Using a five-level locomotion score, 409 to 572 cows were scored twice live and twice from video by three experienced and two inexperienced raters. Intrarater and interrater reliability (expressed as weighted kappa, κw) and agreement (expressed as percentage of agreement, PA) for live/live, live/video and video/video comparisons were determined. A logistic regression was performed to estimate the influence of rater and method of observation on the probability of classifying a cow as lame in live and video observations. Experienced raters had higher values for intrarater reliability and agreement for video/video than for live/live and live/video comparison. Inexperienced raters, however, did not differ for intrarater and interrater reliability and agreement for live/live, live/video and video/video comparisons. The logistic regression indicated that raters were responsible for the main effect and the method of observation (live or from video) had a minor effect on the probability for classifying a cow as lame (locomotion score ≥ 3). Since scoring from video did not show any important influence on the probability of classifying a cow as lame, scoring from video seems to be an acceptable method for assessing locomotion in dairy cows.

Data from Chapter 4 and 5 were obtained from the same experiment. Ten experienced raters scored 58 video records for locomotion and five different gait and posture traits (reluctance to bear weight, arched back, asymmetric gait, head bobbing and tracking up). A similar number of cows were allocated to each level of the five-level scale for locomotion

scoring (a heterogeneous population sample). The 58 video records were scored for locomotion and gait and posture traits in two sessions.

The objective of Chapter 4 was to evaluate different ways of merging levels to optimize resolution, reliability and agreement of locomotion scores for dairy cows. Intrarater and interrater reliability and agreement had a large variation. Intrarater reliability ranged from 0.63 - 0.86 whereas intrarater agreement ranged from 60.3% - 82.8% (PA) for the five-level scale. Interrater reliability ranged from 0.28 - 0.84 (κw) and interrater agreement ranged from 22.6% - 81.8% (PA) for the five-level scale. When locomotion scoring is performed by experienced raters and in a heterogeneous population sample, the lowest specific intrarater and interrater agreement was obtained in level 2 and 3 of the five-level scale. Acceptance threshold for overall intrarater and interrater reliability and agreement and specific intrarater and interrater agreement were exceeded only in the aggregated two-level scale. This increase in agreement, however, was due to chance.

In Chapter 5, the reliability and agreement of raters evaluating five gait and posture traits (reluctance to bear weight, arched back, asymmetric gait, head bobbing and tracking up) and the relation of these traits with locomotion scores were studied. Overall, interrater reliability values ranged from κw = 0.53 for tracking up to κw = 0.61 for reluctance to bear weight. Intrarater and interrater agreement were below the acceptance threshold (PA < 75%). There was a large variation in reliability and agreement obtained by raters for the five traits assessed. Most traits tended to have lower specific intrarater and interrater agreement in level 3 and 5 of the scale. All traits were significantly related with locomotion when scored with a five-level scale and when classified in lame/non-lame. Odd ratios were 10.8 for reluctance to bear weight, 6.5 for asymmetric gait, and 4.8 for arched back and head bobbing. Considering the level of relation with locomotion scoring, intrarater and interrater reliability and agreement, traits to be used in practical conditions are reluctance to bear weight, asymmetric gait and arched back. Slight alterations in specific traits are difficult to detect, even by experienced raters.

Literature (Chapter 2) and the experiments described in this thesis (Chapters 3, 4 and 5) show that there is a large variation in reliability and agreement within and between raters when scoring locomotion or gait and posture traits. The general discussion (Chapter 6) focused on different aspects that could explain this variation such as: the effect of the different procedures used to perform locomotion scoring, the factors affecting reliability and agreement associated to raters (e.g. training, experience and motivation), and the characteristic of the population sample (homogeneous and heterogeneous). The discussion also explored possibilities for the practical application of locomotion scoring systems as welfare indicator and its diagnostic value for hoof or other limb lesions. Finally, since the research was conducted within the framework of the BioBusiness project, this chapter

discussed the usefulness of automatic locomotion scoring systems for classifying cows as lame, the detection of hoof lesions, and the possibilities of using them on-farm.

In conclusion, this thesis shows that raters have a large variation in reliability and agreement when assessing locomotion. The variation is explained by different factors such as the lack of a standard procedure for assessing locomotion or the characteristics of the population sample that is assessed (i.e. population samples tending to be homogeneous or heterogeneous). The factor affecting reliability and agreement most, however, is the rater him/herself. Although the probability for obtaining acceptable reliability and agreement levels increases with training and experience, it is not possible to assure that raters score cows consistently in every scoring session. Given the large variation in reliability and agreement, it can be concluded that raters have a moderate performance to assess consistently locomotion in dairy cows.

Specific conclusions from this thesis are:

- Raters show variable performance for assessing locomotion in dairy cattle when expressed in terms of reliability and agreement.
- The complementary concepts reliability and agreement, as proposed by Kottner et al. (2011), are useful for a better interpretation of different statistical estimators for consistency. The weighted kappa and kappa coefficient are the preferred reliability estimators for ordinal and binary scales, respectively. Percentage of agreement and specific agreement are the preferred agreement estimators. Interpretation of reliability and agreement must be done taking into account the characteristics of the population sample (e.g. population samples tending to be homogeneous or heterogeneous).
- Experienced raters had better intrarater reliability and agreement when locomotion scoring is performed from video than by live observations. Video observations did not show any important influence on the probability of classifying a cow as lame. Video observations seem to be an acceptable method for assessing locomotion in dairy cows.
- Acceptance threshold for overall intrarater and interrater reliability ($\kappa w \geq 0.6$) and agreement and specific intrarater and interrater agreement ($\geq 75\%$) were exceeded only when a five-level scale is merged into a two-level scale. This increase in agreement, however, was due to chance. Raters showed moderate agreement when scoring slightly impaired locomotion.
- Raters showed variable performance when scoring gait and posture traits. Traits which are most related to locomotion scores are reluctance to bear weight, arched

back and asymmetric gait. Raters showed moderate agreement when scoring slightly impaired traits.

- Locomotion scoring systems have a limited utility as tool for classifying cows as lame.
- Locomotion scoring outcomes have a moderate relationship with hoof or other limb lesions. Poor performance for detecting hoof lesions is related to the large number of false negatives (e.g. cows classified as non-lame but do have hoof lesions).
- Given the variability in the classification of cows as lame and the moderate association with hoof and other limb lesions, locomotion scoring systems should not be included in protocols aiming to certificate and assure animal welfare on farms. Although, it is recommended to include locomotion scoring systems in programs aiming to improve hoof health, locomotion scoring should never be used as a unique strategy for the management of hoof or other limb lesions.
- Automatic locomotion scoring systems have a similar variable performance as locomotion scoring systems for lameness classification and hoof or other limb lesions.

Samenvatting

In de huidige melkveehouderij is kreupelheid nog steeds een veelvoorkomend gezondheidsprobleem en een belangrijke reden tot afvoer van melkkoeien. Het loopgedrag of locomotie van een koe kan worden uitgedrukt in een locomotiescore welke gegeven wordt na het beoordelen van de locomotie via een vooraf vastgesteld protocol. Er zijn verschillende systemen om locomotie te beoordelen en een score toe te kennen, waarbij de waarnemers op verschillende bewegings- en houdingskenmerken letten die in het protocol beschreven zijn. In de melkveeveehouderij worden de systemen voor het beoordelen van locomoties in het algemeen alleen gebruikt om van koeien aan te geven of zij kreupel zijn of niet. Hierbij wordt vaak aangenomen dat een koe die als kreupel aangemerkt wordt ook pijn lijdt als gevolg van een klauw- of pootaandoening. Daarom worden systemen voor het beoordelen van locomotie soms ook gebruikt om klauw- en pootaandoeningen op te sporen.

Dit proefschrift maakt deel uit van een project waarin een systeem ontwikkeld wordt dat automatisch de locomotie van een koe beoordeelt en een locomotiescore vaststelt, hetgeen vertaald wordt in een classificatie of een koe wel of niet kreupel is. Automatische systemen maken gebruik van sensoren in plaats van mensen die de waarnemingen uitvoeren en de gegevens van de koeien verzamelen. De sensorwaarnemingen worden met mathematische rekenregels geanalyseerd. Bij de ontwikkeling van automatische systemen voor het vaststellen van de locomotiescore worden menselijke waarnemingen van diezelfde locomotie gebruikt voor de validatie. Het is echter de vraag of mensen in staat zijn om locomotie goed en consistent te kunnen beoordelen.

Het vaststellen van kreupelheid is alleen zinvol als: a) waarnemers in staat zijn om consistent locomotie te beoordelen en een locomotiescore toe te kennen, en b) koeien die als kreupel aangemerkt worden ook daadwerkelijk een klauw- of pootaandoening hebben. Het beoordelen van locomotie vormt de basis voor een goede kreupelheidsclassificatie die gebaseerd is op metingen aan het dier. Daarom was het doel van dit proefschrift om de capaciteit, in de vorm van betrouwbaarheid en overeenstemming, van waarnemers vast te stellen om locomotie van koeien te beoordelen. Daarnaast is onderzocht wat de praktische toepassingen zijn van het gebruik van toegekende locomotiescores voor het classificeren van kreupelheid en/of klauw- en pootaandoeningen.

In hoofdstuk 2 worden de resultaten besproken van een literatuurstudie waarin 244 wetenschappelijke artikelen zijn bestudeerd. 25 systemen met waarnemers en 15 automatische systemen worden beschreven en vergeleken. Daarnaast zijn de verschillende systemen geëvalueerd op betrouwbaarheid, overeenstemming en validiteit. De systemen met waarnemers voor het beoordelen van locomotie gebruiken verschillende type schalen (ordinaal of continu) en verschillende beweging- en houdingskenmerken. De vijftien

gevonden automatische systemen konden ingedeeld worden naar de volgende drie principes: a) kinetisch, b) kinematisch, en c) een indirecte benadering. Schattingen van betrouwbaarheid en overeenstemming van deze systemen voor het beoordelen van locomotie zijn echter beperkt gerapporteerd; dit geldt ook voor de meest gebruikte systemen. Schattingen van betrouwbaarheid en overeenstemming van de systemen voor het beoordelen van locomotie en na classificatie van kreupel of niet-kreupel vertoonden grote verschillen tussen, en soms zelfs binnen artikelen. De systemen met waarnemers worden vaak gebruikt als referentie voor modelkalibratie en –validatie bij de ontwikkeling van automatische systemen voor het beoordelen van locomotie. Echter, verschillen in schattingen van overeenstemming en betrouwbaarheid van de locomotie beoordeling maakt het moeilijk om een eenduidige afbakening te geven van wanneer een koe als kreupel aangemerkt moet worden. Dit heeft een effect op de validiteit van systemen voor automatische locomotie beoordeling. Zowel systemen met waarnemers als automatisch systemen hebben een beperkte validiteit om klauw- en pootaandoeningen op te sporen. Lange termijn studies zijn noodzakelijk om systemen met waarnemers en automatisch systemen te kunnen vergelijken in situaties met verschillende praktische omstandigheden en met preventieve en curatieve behandelingen van bijvoorbeeld klauwaandoeningen.

Het doel van het experiment dat in hoofdstuk 3 beschreven is, was om de overeenstemming en betrouwbaarheid tussen en binnen ervaren en onervaren waarnemers te evalueren. Daartoe moesten ze direct en indirect (op basis van een video) de locomotie van koeien beoordelen. Daarnaast werd ook de invloed van de waarnemers en de methode (direct vs. indirect) op het succesvol kunnen classificeren van kreupel en niet-kreupel onderzocht. Drie ervaren en twee onervaren waarnemers hebben gedurende drie weken tussen de 409 en 572 koeien beoordeeld. Zij maakten hierbij gebruik van een 5-punts scoresysteem. Per week werd twee keer direct en twee keer indirect koeien beoordeeld. De populatie was homogeen in dit experiment. De betrouwbaarheid (uitgedrukt in gewogen kappa, κw) en overeenstemming (uitgedrukt als percentage overeenstemming, PA) zijn berekend tussen en binnen waarnemers voor direct/direct, direct/indirect en indirect/indirect vergelijkingen. Betrouwbaarheid en overeenstemming binnen waarnemers was hoger voor de ervaren waarnemers in de indirect/indirect vergelijking dan voor de direct/direct en direct/indirect vergelijking, maar dit was niet het geval voor de onervaren waarnemers. Op basis van logistische regressie bleek dat de waarnemers een groot effect had, en de methode van waarnemen (direct of indirect) een klein effect had op de kans om een koe als kreupel (een locomotiescore ≥ 3) te classificeren. Omdat indirecte waarnemingen geen groot effect hadden op de kans om een koe als kreupel te classificeren, lijkt het verantwoord om video-opnames te gebruiken om locomotie te beoordelen en daarmee kreupelheid bij koeien te detecteren.

De resultaten beschreven in hoofdstuk 4 en 5 zijn afkomstig van hetzelfde experiment. Tien ervaren waarnemers beoordeelden 58 video's van lopende koeien en gaven een score

voor de locomotie en voor vijf verschillende bewegings- en houdingskenmerken (ontlasting van een poot, kromming van de rug, asymmetrie in beweging, mate van op en neer bewegen van de kop, en het optrekken van de poten). Iedere klasse van de locomotieschaal (1 tot en met 5) kreeg ongeveer evenveel video's toegewezen, waarmee de populatie als heterogeen gekenmerkt kon worden. De 58 video's zijn in twee sessies getoond, iedere keer in een willekeurige volgorde.

Het doel van het onderzoek beschreven in hoofdstuk 4 was om het effect van verschillende manieren van aggregatie van klassen op resolutie, overeenstemming en betrouwbaarheid van locomotiescores vast te stellen. Betrouwbaarheid binnen waarnemers varieerde van 0.63 tot 0.86 (κw), terwijl de overeenstemming (PA) binnen waarnemers varieerde van 60.3% tot 82.8%. Betrouwbaarheid tussen waarnemers varieerde van 0.28 tot 0.84 (κw) en overeenstemming tussen waarnemers varieerde van 22.6% tot 81.8%. De laagste overeenstemming tussen en binnen waarnemers bij deze heterogene populatie was bij klasse 2 en 3 van de 5-punts schaal. De acceptatiegrenzen voor overeenstemming en betrouwbaarheid tussen en binnen waarnemers zijn alleen gehaald voor het aggregatieniveau waarin twee klassen overbleven. Hierbij dient opgemerkt te worden dat de toename in overeenstemming bij hogere aggregatieniveaus was toe te wijzen aan toeval.

In hoofdstuk 5 zijn de betrouwbaarheid en overeenstemming van waarnemers die vijf verschillende houdings- en bewegingskenmerken (ontlasting van een poot, kromming van de rug, asymmetrie in beweging, maten van op en neer bewegen van de kop, het optrekken van de poten) scoren en de relatie met de waargenomen locomotiescore bestudeerd. De betrouwbaarheid (κw) tussen waarnemers varieerde van 0.53 voor 'het optrekken van de poten' tot 0.61 voor 'ontlasting van een poot'. De overeenstemming tussen en binnen waarnemers bleef in alle gevallen onder de acceptatiegrens van 75%. De waarnemers vertoonden grote verschillen in overeenstemming en betrouwbaarheid bij het scoren van de vijf houdings- en bewegingskenmerken. De meeste kenmerken vertoonden een lagere specifieke overeenstemming tussen en binnen waarnemers in klasse 3 en 5 van de 5-punts schaal. Alle houdings- en bewegingskenmerken hadden een significante relatie met de locomotiescore en met de classificatie van kreupel of niet-kreupel. De kans-verhoudingen (odd ratios) waren 10.8 voor 'ontlasting van een poot', 6.5 voor 'asymmetrie in beweging' en 4.8 voor 'kromming van de rug' en 'op en neer bewegen van de kop'. Het niveau van de relatie tussen de locomotiescore en de betrouwbaarheid en overeenstemming tussen en binnen waarnemers in ogenschouw nemende zijn de kenmerken 'ontlasting van een poot', 'asymmetrie in beweging' en 'kromming van de rug' bruikbaar in praktische omstandigheden. Kleine veranderingen in houdings- en bewegingskenmerken zijn zelfs voor ervaren waarnemers moeilijk waar te nemen.

Gebaseerd op het literatuuronderzoek (hoofdstuk 2) en de uitgevoerde experimenten (hoofdstuk 3, 4 en 5) kan gesteld worden dat er grote verschillen zijn in betrouwbaarheid en overeenstemming tussen waarnemers als zij locomotie of houdings- en bewegingskenmerken moeten beoordelen. De algemene discussie (hoofdstuk 6) gaat op verschillende aspecten in die deze verschillen zouden kunnen verklaren, zoals: het effect van de verschillende systemen om locomotie te beoordelen, de factoren die betrouwbaarheid en overeenstemming van waarnemers beïnvloeden (zoals training, ervaring en motivatie), en de karakteristieken (homogeen, heterogeen) van de populatie waarin gemeten wordt. In de discussie is ook verkend wat de praktische toepassingen kunnen zijn van de systemen voor het beoordelen van locomotie als welzijnsindicator of als diagnostisch systeem voor klauw- of pootaandoeningen. Omdat dit promotieonderzoek onderdeel uitmaakt van het BioBusiness project wordt in de discussie ook ingegaan op het praktisch nut van systemen voor het automatisch beoordelen van locomotie van koeien op melkveebedrijven en om dit te vertalen in een classificatie voor kreupelheid en klauw- en pootaandoeningen.

De hoofdconclusie van dit onderzoek is dat aangetoond is dat waarnemers grote verschillen laten zien in betrouwbaarheid en overeenstemming bij het beoordelen van locomotie van, en het toekennen van een locomotiescore aan koeien. De verschillen worden verklaard door onder andere een gebrek aan een standaard voor het geven van een locomotiescore en het meenemen van de populatiekarakteristiek (homogeen of heterogeen) bij het waarnemen. Bij dit alles is de waarnemer nog steeds de belangrijkste factor die de betrouwbaarheid en de overeenstemming van de waarnemingen bepaalt. Alhoewel de kans op het verkrijgen van acceptabele overeenstemmingen en betrouwbaarheden toeneemt met training en ervaring kan het niet gegarandeerd worden dat waarnemers koeien consistent beoordelen. Gegeven de grote verschillen in betrouwbaarheid en overeenstemming kan geconcludeerd worden dat waarnemers een beperkt vermogen hebben om locomotie van koeien consistent te beoordelen.

Meer specifieke conclusies uit dit onderzoek zijn:

- Betrouwbaarheid en overeenstemming, zoals voorgesteld door Kottner et al. (2011), vullen elkaar aan en zijn beter geschikt als statistische indicatoren dan consistentie. De statistische indicatoren gewogen kappa en kappa zijn geschikt om de betrouwbaarheid te duiden voor ordinale en binaire schalen. Het percentage overeenstemming en de specifieke overeenstemming zijn de voorkeursindicatoren voor overeenstemming. De interpretatie van overeenstemming en betrouwbaarheid moet gedaan worden met inachtneming van de karakteristiek (homogeen of heterogeen) van de populatie waarin gemeten wordt.

- Ervaren waarnemers hadden een hogere betrouwbaarheid en overeenstemming binnen waarnemers als de locomotie beoordeling indirect (via video) werden uitgevoerd in plaats van via directe waarnemingen. Indirecte waarnemingen hadden geen invloed op de kans om een koe als kreupel te classificeren en zijn daarmee acceptabel voor het vaststellen van locomotiescores van koeien.
- De acceptatiegrenzen voor betrouwbaarheid ($\kappa w \geq 0.6$) en overeenstemming (PA > 75%) tussen en binnen waarnemers werden alleen gehaald als de 5-punts schaal werd geaggregeerd tot een 2-punts schaal. De toename voor overeenstemming bij deze aggregatie was echter toeval. Waarnemers hadden een matige overeenstemming bij het beoordelen van koeien met een licht afwijkende locomotie.
- Waarnemers lieten veel verschillen zien tijdens het scoren van bewegings- en houdingskenmerken. Kenmerken die het meest gerelateerd zijn aan locomotiescores zijn 'ontlasting van een poot', 'asymmetrie in beweging' en 'kromming van de rug'. Waarnemers hadden een matige overeenstemming bij het beoordelen van koeien met een licht afwijkende bewegings- en houdingskenmerken.
- Systemen die een locomotiescore geven hebben een beperkt praktisch nut om koeien als kreupel te classificeren.
- Uitkomsten van het beoordelen van locomotie (locomotiescores) hebben een beperkte relatie met het voorkomen van klauw- en pootaandoeningen. Slechte resultaten om klauwaandoeningen te duiden is gerelateerd aan het grote aantal vals negatieven (bijvoorbeeld koeien die als niet-kreupel aangemerkt worden maar toch een klauwaandoening hebben).
- Gegeven de verschillen in classificatie van koeien die als kreupel aangemerkt kunnen worden en de beperkte relatie met klauw- en pootaandoeningen zouden systemen voor het beoordelen van locomotie niet opgenomen moeten worden in welzijnsprotocollen om dierenwelzijn op bedrijven in te schatten. Alhoewel het advies is om systemen voor het beoordelen van locomotie op te nemen in programma's voor verbetering van klauwgezondheid, moet ervoor gewaakt worden om dit als enige strategie toe te passen om klauwproblemen te detecteren.
- Systemen voor automatisch locomotie beoordeling vertonen vergelijkbare resultaten als systemen voor classificatie van kreupelheid en klauw- en pootaandoeningen.

About the author

Curriculum Vitae

Andrés was born in Osorno, the heart of the Chilean dairy industry, in 1980. During his childhood Andrés worked at the family's dairy farm where he got involved in this world. Given the influence of the dairy world in his life, Andrés decided to study veterinary medicine at the Universidad de Chile obtaining the degree in 2003. Afterwards, Andrés worked as consultant for installation of Precision Livestock Farming tools aiming to improve management of dairy and beef farms as well as consultant for nutrition based on grazing for dairy cows. After working in the private industry, Andrés decided pursuing a career as researcher. Andrés enrolled a program for obtaining the degree Master of Science (MSc) in animal nutrition at the Mediterranean Agronomic Institute of Zaragoza in Spain. The research of his Master thesis was related to mineral nutrition in dairy ewes, and was done at the Universitat Autonoma de Barcelona in Spain.

In 2010 Andrés started his PhD at Livestock Research Wageningen UR and Farm Technology Group of Wageningen University. Andrés' PhD work aimed to establish the performance of raters to perform locomotion scoring for lameness and hoof lesions detection in order to support the development of an automatic locomotion scoring system. This PhD work was part of the BioBusiness project which was a European project involved in the Marie Curie Initial Training Networks. By working in the BioBusiness Project, Andrés acquired experience on dealing with multicultural and multidisciplinary teams in order to achieve the project goal.

Andrés is author and co-author of quality scientific articles published in top journals in his research field (Preventive Veterinary Medicine, Journal of Dairy Science and Animal Welfare). Andrés presented his PhD work, in oral and poster presentations, at several international conferences such as ISAH, EAAP and ECPLF. Currently Andrés is writing project aiming to develop Precision Livestock Farming tools for improving health and welfare in dairy farming.

Publications

Articles in internationally reviewed academic journals

Schlageter-Tello, A., Bokkers, E. A. M., Groot Koerkamp, P. W. G.,Van Hertem, T., Viazzi, S., Bites Romanini, C., Halachmi, I., Bahr, C., Berckmans, D., Lokhorst, K. (2015). Comparison of locomotion scoring for dairy cows by experienced and inexperienced raters using live or video observation methods. Animal Welfare 24: 69-79.

Van Hertem, T., Parmet, Y., Steensels, M., Maltz, E., Antler, A., Schlageter-Tello, A., Lokhorst, K., Bites Romanini, C., Viazzi, S., Bahr, C., Berckmans, D., Halachmi, I. (2014). The effect of routine hoof trimming on locomotion score, ruminating time, activity, and milk yield of dairy cows. Journal of Dairy Science. 97: 4852-4863.

Schlageter-Tello, A., Bokkers, E. A. M., Groot Koerkamp, P. W. G., Van Hertem, T., Viazzi, S., Bites Romanini, C., Halachmi, I., Bahr, C., Berckmans, D., Lokhorst, K. (2014). Manual and automatic locomotion scoring systems in dairy cows: A review. Preventive Veterinary Medicine, 116: 12-25.

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

Viazzi, S., Bahr, C., Van Hertem, T., Schlageter-Tello, A., Bites Romanini, C., Halachmi, I., Lokhorst, C., Berckmans, D. (2014). Comparison of a three-dimensional and two-dimensional camera system for automated measurement of back posture in dairy cows. Computers and Electronics in Agriculture. 100: 139-147.

Van Hertem, T., Viazzi, S., Steensels, M., Maltz, E., Antler, A., Alchanatis, V., Schlageter-Tello, A., Lokhorst, K., Bites Romanini, C., Bahr, C., Berckmans, D., Halachmi, I. (2014).Automatic lameness detection based on consecutive 3D-video recordings. Biosystems Engineering. 119: 108-116.

Van Hertem, T., Maltz, E., Antler, A., Bites Romanini, C., Viazzi, S., Bahr, C., Schlageter-Tello, A., Lokhorst, C., Berckmans, D., Halachmi, I. (2013). Lameness detection based on multivariate continuous sensing of milk yield, rumination, and neck activity. Journal of Dairy Science. 96 (7): 4286-4298.

Van Hertem, T., Alchanatis, V., Antler, A., Maltz, E., Halachmi, I., Schlageter-Tello, A., Lokhorst, C., Viazzi, S., Romanini, E., Pluk, A., Bahr, C., Berckmans, D. (2013). Comparison of segmentation algorithms for cow contour extraction from natural barn background in side view images. Computers and Electronics in Agriculture. 91: 65-74.

Viazzi, S., Bahr, C., Schlageter-Tello, A., Van Hertem, T., Bites Romanini, C., Pluk, A., Halachmi, I., Lokhorst, C., Berckmans, D. (2013). Analysis of individual classification of lameness using automatic measurement of back posture in dairy cattle. Journal of Dairy Science. 96: 257-266.

Papers at international scientific conferences and symposia, published in full in proceedings

Bites Romanini, C., Viazzi, S., Van Hertem, T., Schlageter-Tello, A., Halachmi, I., Lokhorst, C., Bahr, C., Berckmans, D. (2013). Video pre-processing for the improvement of an automated lameness detection system for dairy cows. In Berckmans, D. (Ed.), Vandermeulen, J. (Ed.), Precision Livestock Farming 2013. European Conference on Precision Livestock Farming. Leuven, Belgium. 10-12 September 2013 (pp. 479-487).

Viazzi, S., Van Hertem, T., Schlageter-Tello, A., Bahr, C., Bites Romanini, C., Halachmi, I., Lokhorst, C., Berckmans, D. (2013). Using a 3D Camera to Evaluate the Back Posture of Dairy Cows. ASABE Annual International Meeting. ASABE Annual International Meeting. Kansas City, Missouri, USA. 21-24 July 2013 (pp. 4222-4227).

Schlageter-Tello, A., Bokkers, E. A. M., Groot Koerkamp, P. W. G., Van Hertem, T., Viazzi, S., Bites Romanini, C., Halachmi, I., Bahr, C., Berckmans, D., Lokhorst, C. (2013). Gold standards concepts for automatic lameness assessment systems in dairy cows. In Berckmans, D. (Ed.), Vandermeulen, J. (Ed.), Precision Livestock Farming 2013. European Conference on Precision Livestock Farming. Leuven, Belgium. 10-12 September 2013 (pp. 471-478).

Van Hertem, T., Maltz, E., Antler, A., Alchanatis, V., Schlageter-Tello, A., Lokhorst, C., Bites Romanini, C., Viazzi, S., Bahr, C., Berckmans, D., Halachmi, I. (2013). Automatic lameness detection based on 3D-video recordings. In Berckmans, D. (Ed.), Vandermeulen, J. (Ed.), Precision Livestock Farming 2013. European Conference on Precision Livestock Farming. Leuven, Belgium. 10-12 September 2013 (pp. 59-67).

Viazzi, S., Van Hertem, T., Bites Romanini, C., Bahr, C., Halachmi, I., Schlagater-Tello, A., Lokhorst, C., Rozen, D., Berckmans, D. (2013). automatic back posture evaluation in dairy cows using a 3D camera. Precision Livestock Farming '13: Vol. 1. oint European Conference on Precision Livestock Farming. Leuven, Belgium. 10-12 September 2013 (pp. 83-92).

Bites Romanini, C., Bahr, C., Viazzi, S., Van Hertem, T., Schlageter-Tello, A., Halachmi, I., Lokhorst, K., Berckmans, D. (2013). Application of image based filtering to improve the performance of an automated lameness detection system for dairy cows. 2013 ASABE Annual International Meeting. ASABE Annual International Meeting. Kansas City, Missouri - USA, 21-24 July 2013 (art.nr. 131620675) ASABE.

Schlageter Tello, A., Lokhorst, C., Van Hertem, T., Halachmi, I., Maltz, E., Voros, A., Bites Romanini, C., Viazzi, S., Bahr, C., Groot Koekamp, P. W. G., Berckmans, D. (2011). Selection of a golden standard for visual-based automatic lameness detection for dairy cows. In Kofer, J. (Ed.), Schobesberger, H. (Ed.), Proceedings of the XVth International Congress of the International Society for Animal Hygiene: Vol. 1. International Congress on Animal Hygiene. Vienna, Austria. 3-7 July 2011 (pp. 325-327).

Schlageter-Tello, A., Bokkers, E.A.M., Koerkamp, P.W.G., Van Hertem, T., Viazzi, S., Romanini, C.E.B., Halachmi, I., Bahr, C., Berckmans, D., And Lokhorst, C. (2013). Within and between observer agreement for specific levels in a five levels locomotion score for dairy cows. In Proceedings 17th International Symposium and 9th International Conference on Lameness in Ruminants, Bristol, England 11 -14 August 2013. (pp. 88 – 89).

Van Hertem, T., Alchanatis, V., Antler, A., Maltz, E., Halachmi, I., Schlageter Tello, A., Lokhorst, K., Voros, A., Bites Romanini, C., Bahr, C., Berckmans, D. (2011). Experimental setup for the study of a computer vision based automatic lameness detection system for dairy cows. In Lokhorst, K. (Ed.), Berckmans, D. (Ed.), Precision Livestock Farming 2011: Vol.1 (1). European Conference on Precision Livestock Farming. Czech Republic Prague. 11-14 July 2011 (pp. 113-121).

Meeting abstracts, presented at international scientific conferences and symposia, published or not published in proceedings or journals

Van Hertem, T., Bahr, C., Viazzi, S., Steensels, M., Bites Romanini, C., Lokhorst, K., Schlageter Tello, A., Halachmi, I., Maltz, E., Berckmans, D. (2014). On farm implementation of a fully automatic computer vision system for monitoring gait related measures in dairy cows. Annual International Meeting of the American Society of Agricultural and Biological Engineers. Montreal, Quebec Canada. 13-16 July 2014.

Schlageter-Tello, A., Van Hertem, T., Viazzi, S., Bokkers, E. A. M., Groot Koerkamp, P. W. G., Bites Romanini, C., Steensels, M., Bahr, C., Halachmi, I., Berckmans, D., Lokhorst, K. (2014). Hoof lesion detection of dairy cows with manual and automatic locomotion scores. Book of Abstracts of the 65th Annual Meeting of the European Federation of Animal Science: vol. 20.EAAP - Annual Meeting of the European Federation of Animal Science. Copenhagen, Denmark. 25-29 August 2014, .

Van Hertem, T., Steensels, M., Viazzi, S., Bahr, C., Bites Romanini, C., Lokhorst, K., Schlageter Tello, A., Maltz, E., Halachmi, I., Berckmans, D. (2014). Effect of cow traffic on an implemented automatic 3D vision monitor for dairy cow locomotion. Book of Abstracts of the 65th Annual Meeting of the European Federation of Animal Science: vol. 20. EAAP - Annual Meeting of the European Federation of Animal Science. Copenhagen, Denmark. 25-29 August 2014.

Van Hertem, T., Steensels, M., Viazzi, S., Bites Romanini, C., Schlageter-Tello, A., Lokhorst, K., Maltz, E., Halachmi, I., Hong, S., Bahr, C., Berckmans, D. (2014). Automatic lameness detection by computer vision and behavior and performance sensing. International Conference of Agricultural Engineering (AgEng). Zurich, Switzerland. 6-10 July 2014.

Van Hertem, T., Maltz, E., Viazzi, S., Bites Romanini, C., Bahr, C., Berckmans, D., Lokhorst, K., Schlageter Tello, A., Antler, A., Halachmi, I. (2013). The effect of hoof trimming on the locomotion score, neck activity and ruminating time of dairy cows. Proceedings of the 64th Annual Meeting of the European Association for Animal Production: vol. 134. EAAP - Annual Meeting of the European Federation of Animal Science. Nantes, France. 26-30 August 2013.

Van Hertem, T., Viazzi, S., Bites Romanini, C., Bahr, C., Berckmans, D., Schlageter Tello, A., Lokhorst, K., Rozen, D., Maltz, E., Halachmi, I. (2013). Automatic lameness detection by computer vision and behavior and performance sensing. Proceedings of The 2013 Joint ADSA-ASAS Annual Meeting. The 2013 Joint ADSA-ASAS Annual Meeting. Indianapolis, Indiana, USA. 8 - 12 July 2013.

Van Hertem, T., Maltz, E., Antler, A., Schlageter-Tello, A., Lokhorst, C., Viazzi, S., Bites Romanini, C., Bahr, C., Berckmans, D., Halachmi, I. (2012). Evaluation of potential variables for sensor-based detection of lameness in dairy cattle. Proceedings of the 63rd Annual Meeting of the European Association for Animal Production. EAAP. Bratislava, Slovakia. 27 -31 August 2012.

Schlageter-Tello, A., Lokhorst, C., Bokkers, E. A. M., Groot Koerkamp, P. W. G., Van Hertem, T., Steensels, M., Halachmi, I., Maltz, E., Viazzi, S., Bites Romanini, C., Bahr, C., Berckmans, D.(2012). Comparison between direct and video image observation for locomotion assessment in dairy cow. Proceedings of the 63rd Annual Meeting of the European Association for Animal Production. EAAP. Bratislava, Slovakia. 27 -31 August 2012.

Schlageter-Tello, A., Van Hertem, T., Viazzi, S., Bites Romanini, C., Bergoug, H., Tong, Q., Ismayilova, G., Roulston, N., Sonda, L., Rozen, D., Oczak, M., Bahr, C., Berckmans, D. (2012). BioBusiness Project: Development of Precision Livestock Farming Solutions for Animal Welfare. Proceedings of the Encuentros 2012. Encuentros 2012. Paris, France, 4-6 July 2012.

Van Hertem, T., Bites Romanini, C., Bahr, C., Schlageter-Tello, A., Lokhorst, K., Voros, A., Maltz, E., Berckmans, D., Halachmi, I. (2011). Precision Agriculture in dairy farming: Experimental setup for a computer vision based automatic lameness detection system. Abstract Book of the International Symposium on Sensing in Agriculture 2011. International Symposium on Sensing in Agriculture in Memory of Dahlia Greidinger. Technion Haifa, Israel, 21-24 February 2011.

Van Hertem, T., Alchanatis, V., Antler, A., Bites Romanini, C., Bahr, C., Schlageter-Tello, A., Lokhorst, K., Voros, A., Maltz, E., Berckmans, D., Halachmi, I. (2011). Experimental setup for the study of a computer vision based automatic lameness detection system for dairy cows. Proceedings of The Annual Meeting of the Israeli Society of Agricultural Engineers. The Annual Meeting of the Israeli Society of Agricultural Engineers. Bet Dagan, Israel, 7 July 2011.

Training and supervision plan

The Basic Package (**3 ECTS**)

- WIAS Introduction Course (2011)
- Course on philosophy of science and/or ethics (2011)

Scientific Exposure (21 ECTS)

International conferences

- 15th ISAH Congress, Vienna Austria, 3-7 Jul (2011)
- 5th ECPLF, Prague, Czech Republic, 11-14 Jul (2011)
- Encuentros 2012, Paris, France, 4-6 Jul (2012)
- 63rd Annual Meeting EAAP, Bratislava, Slovakia, 29 Aug-2 Sep (2012)
- 6th ECPLF, Leuven, Belgium, 10-12 Sep (2013)
- 9th Conference on Lameness in Ruminants, Bristol, England, 11-14 Aug (2013)
- 65th Annual Meeting EAAP, Copenhagen, Denmark, 25-29 Aug (2014)

Workshops and Seminars

- Workshop BioBusiness project. Nazareth, Israel. 26-29 Sep (2010)
- Workshop BioBusiness project Celle, Germany, 16-19 Nov (2010)
- Workshop BioBusiness project. Ploufragan, France, 14-16 Mar (2011)
- Workshop BioBusiness project. Brussels, Belgium, 5 – 9 Sep (2011)
- Workshop BioBusiness project. Paestum. Italy, 11-14 Sep (2012)
- Workshop BioBusiness project, Bet Dagan, Israel, 21-23 May (2012)

Presentations

- Theatre, 15th ISAH Congress, Vienna Austria, 3-7 Jul (2011)
- Poster, Encuentros 2012, Paris, France, 4-6 Jul (2012)
- Theatre, 63rd Annual Meeting EAAP, Bratislava, Slovakia, 29 Aug-2 Sep (2012)
- Poster, 6th ECPLF, Leuven, Belgium, 10-12 Sep (2013)
- Theatre, 9th Conference on Lameness in Ruminants, Bristol, England, 11-14 Aug (2013)
- Theatre, 65th Annual Meeting EAAP, Copenhagen, Denmark, 25-29 Aug (2014)

In-Depth Studies (9 ECTS)

- Advanced statistics course design of experiments, WIAS, Wageningen UR (2010)
- Statistics for the life Science, WIAS, Wageningen UR (2011)
- Meta-analysis, PE & RC, Wageningen UR (2012)
- Animal Pain, Aarhus University (2013)

Professional Skills Support Courses (5 ECTS)

- Dutch for Employees, Wageningen UR (2010)
- Information Literacy including and EndNote Introduction, Wageningen UR (2010)
- Effective Presentation, BioBusiness training, KU Leuven (2011)
- Making a video for marketing a project, BioBusiness training, KU Leuven, Belgium (2013)

Research Skills Training (5 ECTS)

- External training period, Agricultural Research Organization, Tel Aviv, Israel (2012)
- External training period KU Leuven, Belgium (2013)
- Reviewing scientific articles for Conferences and peer reviewed journals (2013 - 2014)

Didactic Skills Training (2 ECTS)

- Supervising MSc thesis

Management Skills Training (1 ECTS)

- Bio-Business Fellow Board (Newsletter Editor)

Total ECTS in training and supervision plan: 46 ECTS

Acknowledgement

Recently, I watched "The Castaway" starring by Tom Hanks. In one of the final scenes, Tom was speaking to his friends about his sad experience on the island where he was trapped. In that very scene, Hanks told one of the best quotes in cinema history (according to my opinion) He said: "Keep breathing, because tomorrow the sun will rise. Who knows what the tide could bring?". That quote reminded me of the story how I arrived in Wageningen. By March 2010, I was working in a lost and huge dairy farm in the south of Chile. The farm was more or less isolated and contact with people outside the farm was really limited. While working on this farm, my ex-boss sent me an advertisement for a PhD position at Wageningen UR. Without further expectation, I applied for that position. After several e-mails and two Skype interviews (in my car), I was announced that I'd got the position. And that's the story about how a guy working in a lost farm in southern Chile in early 2010, went to work to Netherlands by September 2010. Who knows what the tide could bring?

Fortunately for me, the tide brought me one of the most wonderful experiences in my life. In the beginning I thought that a PhD was only about improving researcher skills; How wrong I was. After making a PhD I can state that a PhD changes your perception of the world. Given the deep and positive impact that this experience had in my life, I think this "acknowledge words" will hardly reflect my deep gratitude towards the people who were part of this process.

First things first. Many thanks to my supervisors, Kees Lokhorst and Eddie Bokkers. You coached me, guided me, inspired me, listen to me, advise me, and kept me motivated during 4.5 years. It seems easy, but it is not. Thanks for being "the engine" behind this thesis. To my promotor, Peter Groot Koerkamp thanks for always having the precise comment to make me re-think everything. I learned a lot working with you. To the three of you, thanks for showing me the kind of researcher and professional that, right now, I am aiming to be.

This PhD thesis was done as part of EU BioBusiness Project. The Team was spread all around Europe, but I think we successfully achieved the difficult task of working together towards a common aim. Within the project, I had the opportunity to know and share experience with excellent professionals and wonderful people. Many thanks to the "Cow Group" fellows; Tom Van Hertem, Stefano Viazzi, Daniel Rozen, and to the honorary member, Machteld Steensels. It was great to work with you friends. Thanks to the "Cow Group" supervisors: Claudia Bahr, Ilan Halachmi and Uzi Birk. I also would like to mention the other BioBusiness fellows: Eduardo, Nancy, Gunel, Lilia, Hakim, Monica, Maciej, Anna, and Anna Maria; and their respective supervisors: Vasilis Exadaktylos, Pascal Garain, Marcella Guarino, Jorg Hartung, Nicolas Eterradossi, Theo Demmers and Erik Vranken. Finally, thanks to the Project Coordinator Daniel Berckmans.

About 70% of the time during my PhD, I was physically located in Netherlands, in Wageningen UR Livestock Research offices at Triton Building (Gebouw 119). Some people in Wageningen UR Livestock Research helped me in my research: Gidi Smolders, Wijbrand Ouweltjes, Klaas Blanken, Joop van der Werf, Hans van den Heuvel and Vincent Hindle. Additionally, there I knew most of the colleagues with who I shared nice moments and laughs, meetings and coffee breaks. For all these moments I would like to thank, Bert Ipema, Rudy de Mol (please add me to the Tour de France Pool, even if I'm gone), Peter Hogewerf, Wim Houwers, Johan van Riel and all the "Animal Welfare Group" within Wageningen UR Livestock Research.

I also would like to thank the people outside BioBusiness and Wageningen UR Livestock Researchers who collaborate in this thesis. To Jos Metz, you were the supervisor of some of my supervisors. It was a pleasure to receive your experience. Thanks to the raters who participate in the experiment from Chapters 4 and 5: Rik Vlemminx, Jan Hulsen (Vetvice), Thomas Dijkstra and Menno Holzhauer (GD Animal Health Service) and Fokje Steenstra (WUR). Thanks to Bas Engel and Willem Buist for their valuable statistical advice; and Mike Grossman for helping me to improve the quality of my first paper.

Thanks to people who did not collaborate in the research, but provide me things as important as research and writing skills ... I am speaking of social interaction, emotional support in bad moments and laughs in the good ones. Me gustaría agradecer a la comunidad chilena en Wageningen, a los clásicos como Rossier Miranda, Chavez Oyanedel, Marcela, Lena, Yenni, Pablo, Pamela, Denisse, Gabriel, Daniela P, Manuel, Sofia, Carlos, Alicia y el Tío Pollo; y a los actuales como: Marcia, Carter-Leal, Leo, Daniela B, Labrita, Loreto, Henk, Yelica, Daniel, Nicole, Mauricio. Muchas gracias amigos. I also would like to mention the PhD(c) fellows from "Farm Technology Group", especially to Liansun, Bastiaan and Dennis. Thanks to the people from "Genetics and Breeding group", especially to Coralia (please, do not forget the platinum rule).

Finally, and most important, I would like to thank the people who are always there for me no matter what, my family. A mi familia, en especial a mis tíos Alberto y Wilma y mis primos. A mis hermanos, Karem y Claudio. Mis sobrinos Eloísa, Max, Ines, Theo y Luis. A mi papá. A la meva xicota, Roser. Gràcies per acompanyar-me durant aquests 5 anys. I per mostrar que quan es vol es pot, t'estimo. Y a mi mamá, muchas gracias por todos tus anos de sacrificio, educarme y por hacerme el hombre que soy.

To all of you

Thank you very much
Muchas gracias
Moltes gracies
Hartelijk bedankt

Colophon

This study was part of the Marie Curie Initial Training Network BioBusiness project (FP7-PEOPLE-ITN-2008).

Cover design and pictures by Andrés Schlageter Tello

Printed by GVO drukkers & vormgevers B. V. | Ponsen & Looijen, Ede, the Netherlands