J. Dairy Sci. 97:5533–5542
http://dx.doi.org/10.3168/jds.2014-8129

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

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

*Wageningen UR Livestock Research, PO Box 65, 8200 AB Lelystad, the Netherlands
†Animal Production Systems Group, Wageningen University, PO Box 338, 6700 AH, Wageningen, the Netherlands
‡Farm Technology Group, Wageningen University, PO Box 317, 6700 AH, Wageningen, the Netherlands
§Division Measure, Model & Manage Bioresponses, KU Leuven, PO Box 2456, 3001 Heverlee, Belgium
#Institute of Agricultural Engineering–Agricultural Research Organization, PO Box 6, Bet-Dagan 50250, Israel

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.

Key words: cattle, lameness, observer, resolution

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 nonlame), 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-

Received March 12, 2014.
Accepted June 6, 2014.
[1]Corresponding author: andres.schlagetertello@wur.nl

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 nonlame (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.

MATERIALS AND METHODS

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 secondCamera 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).

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.

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