OPEN ACCESS

ORIGINAL RESEARCH

Reliability of a beef cattle locomotion scoring system for use in clinical practice

Jay Tunstall, Karin Mueller, Oscar Sinfield, Helen Mary Higgins

Abstract

Background Locomotion (lameness) scoring has been used and studied in the dairy industry; however, to the authors' knowledge, there are no studies assessing the reliability of locomotion scoring systems when used with beef cattle.

Methods A four-point scoring system was developed and beef cattle filmed walking on a firm surface. Eight veterinary researchers, eight clinicians and eight veterinary students were shown written descriptors of the scoring system and four video clips for training purposes, before being asked to score 40 video clips in a random order. Participants repeated this task 4 days later.

Results The intra-observer agreement (the same person scoring on different days) was acceptable with weighted mean Kappa values of 0.84, 0.81 and 0.84 respectively for researchers, clinicians and students. The inter-observer agreement (different people scoring the same animal) was acceptable with weighted Gwet's Agreement Coefficient values of 0.70, 0.69 and 0.64 for researchers, clinicians and students. Most disagreement occurred over scores one (not lame but imperfect locomotion) and two (lame, but not severe).

Conclusion This scoring system has the potential to reliably score lameness in beef cattle and help facilitate lameness treatment and control; however, some disagreements will occur especially over scores one and two.

Introduction

Lameness in cattle is considered to be a critical welfare issue,[1 2] with lame beef cattle being a specific issue due to the risks of being left untreated for a long period of time.[3] Identification of lame animals is considered to be an important step in dealing with individual animals, but also in acknowledging and understanding the scale of the problem. As such, the UK dairy industry is encouraging farmers to locomotion score cattle,[1 4] and a sheep locomotion scoring tool is available.[5] Locomotion scoring also allows benchmarking, meaning that farmers can understand how they compare with others, and allows improvements or deteriorations, to be measured over time. However, this requires scorers to be able to give an animal with any given locomotion the same score on any given day. Furthermore, it requires different scorers to also give an animal with any given locomotion the same score in order for the results to be consistent. In particular, a practical and easy-to-use scoring system is needed that can be used by veterinary surgeons in clinical practice. This is different to scoring systems designed specifically for research purposes with typically considerable detail and a large number of possible scores. While necessary for research reasons, it makes them more complex and hence less practical for use by clinicians and is not essential for the routine monitoring and control of lameness in clinical practice.

Any locomotion scoring system would ideally have been proven to be *valid* in the sense that it accurately measures lameness, and also *reliable* which encapsulates the extent to which there is consistency (repeatability) in scores when independent measurements are performed. Although assessing validity of a locomotion scoring system can be challenging, reliability can be assessed in two regards. Inter-observer reliability relates to multiple people scoring the same animal and asks the question: how consistent are the scores they assign? In other words, do different people agree with each other over the same animal? Intra-observer reliability relates to the same person scoring the same animal on different occasions (with degree of lameness

Veterinary Record (2020) doi: 10.1136/vr.105781

E-mail for correspondence: Dr Helen Mary Higgins, Department of Livestock and One Health, Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Cheshire, UK; h.higgins@liverpool.ac.uk

Provenance and peer review Not commissioned; externally peer reviewed.

Received November 12, 2019
Revised July 21, 2020
Accepted July 26, 2020

Table 1 Proposed locomotion scoring system designed for use in beef cattle (adapted from Sprecher *et al*[7] and AHDB[4] scoring systems)

Locomotion score		
0	Normal	Even weight-bearing and rhythm on all four feet. The back is level
1	Imperfect	Uneven steps or shortened strides, but affected limb not identifiable. The back may show minimal arching while walking
2	Impaired	Uneven weight-bearing or shortened strides. Affected limb is identifiable (unless multiple limbs affected). The back may show arching while walking
3	Severely impaired	Slower pace—unable to keep up with the healthy herd. Affected limb easily identifiable (unless multiple limbs affected). An arched back may be noted while standing and walking.

unchanged) and asks the question: to what extent does a person agree with themselves?

In dairy cattle, several scoring systems have been developed and reported in the literature, typically based on a combination of subjective visual observations such as back arching, stride length, weight-bearing and so forth.[4 6 7] However, to the authors' knowledge, none of these have been assessed for reliability when scoring beef cattle.

The aim of this study was to assess the inter-observer and intra-observer reliability of a locomotion scoring system for use with beef cattle in clinical practice by veterinary surgeons.

Methods

Locomotion scoring system

A four-point locomotion scoring system was developed following consultation between the authors based around two current dairy cattle scoring systems, but with due consideration for the practicalities and specific attributes of beef cattle.[4 7] Of these two dairy cattle scoring systems used to develop the new one, the AHDB system is one commonly used in practice the UK, and the Sprecher system is well publicised and cited internationally. The descriptors are given in table 1.

Video clips and online completion

Video clips were created by filming both suckler cows and finishing cattle walking on a firm surface, either from the rear, the side or a transition from side to rear. Where necessary, the animal intended as the focus of the video was identified with an arrow to avoid confusion and any audio was removed. These clips were examined by three experienced researchers to ensure a sufficient range of scores were present (approximately ten of each score, zero, one, two and three) and yielded a total of 40 video clips for assessment by participants. The researchers also selected four additional video clips to be used for training purposes (one clip for each score) that they agreed were typical of each score.

The 40 assessment video clips and 4 training clips were uploaded onto the University of Liverpool's virtual learning environment (VITAL—Virtual Interactive Teaching at Liverpool), which uses Blackboard 2018 (Blackboard, Washington, USA). This platform enabled participants to view the training videos alongside the descriptors for each score at the start. They could re-play the training clips as many times as they wanted. Thereafter, they were asked to watch each of the 40 assessment videos and assign a score to each. Videos lasted between 1 and 18 s. Each assessment video could also be re-played as many times as the observer wanted. The order of the assessment videos was randomised for each participant. After four days, the observers were asked to repeat the entire task, that is, to watch the training videos and read the descriptors again and re-score the 40 assessment videos, which were presented again in a randomised order.

Observers

Observers were a convenience (non-random) sample of eight private practice veterinary surgeons ('clinicians' or 'C') involved with livestock work and undertaking postgraduate livestock courses alongside their clinical role, eight veterinary researchers/lecturers involved with livestock research/teaching ('researchers' or 'R') and eight veterinary students, in years three to five of a five-year course ('students' or 'S'). Observers were coded 1–8 for each group ordered by their intra-observer exact agreement percentage.

Data analysis

The data were exported from VITAL into Microsoft Excel 2016 (Microsoft, Redmond, Washington, USA). Statistical tests were conducted in Minitab V.18.1 (Minitab Statistical Software, State College, Pennsylvania, USA) and R (R Core Team, 2019), including Computing Chance-Corrected Agreement Coefficients R Package (irrCAC, Gwet 2019). P values are reported as continuous values and without setting any arbitrary threshold.[8 9] Quadratic weightings were used to produce weighted Kappa values and AC2 values.

Intra-observer agreement

Percent exact agreement (and ±1 and ±2 scores) was calculated for each observer across the 40 videos and mean values for the three different groups (ie, the researchers clinicians and students) were compared with paired t-tests. Differences between the same observer at the first and second scoring (intra-observer agreement) were examined using weighted Cohen's Kappa values,[10] and the difference between mean values for researchers, clinicians and for students was compared using paired t-tests. Systematic bias between attempts for each scorer was investigated by subtracting each observer's second score from their first, and performing a one-sample t-test on the resulting value (null hypothesis: the mean value equals zero, alternative hypothesis: the mean value is not equal to zero).

Inter-observer agreement

Inter-observer scores were investigated using each observer's first attempt at scoring the videos.

The percentage of video clips that an observer agreed on with each individual observer in their group (ie, the researchers, clinicians and students) was calculated to produce seven scores. The mean of these scores produced the mean exact agreement for that observer. This was repeated for each of the 24 observers to initially assess the agreement within groups. Agreement was formally analysed using quadratic weighted Gwet's Agreement Coefficient 2 (AC2). An AC2 value was produced for each group of observers (researcher, clinician or student) and overall for all observers. AC2 values were adjusted using Critical Values provided by Gwet.[11 12]

For each video, the mode score was determined and considered to be the correct score. One video was bimodal, and the mean score was used to determine which mode to consider correct. All videos of each score were then grouped and an AC2 value generated for each score to show the agreement of observers for each individual locomotion score. This was performed for each group of observers, and overall.

Ethical approval was granted by the University of Liverpool Ethics committee. It is reported in accordance with the guidelines for reporting reliability and agreement studies.[13]

Results

The distribution of scores, as determined by the mode score for each video, were score 0: 12 clips, score 1: 10 clips, score 2: 9 clips and score 3: 9 clips. The results for one video were bimodal; therefore, the mean score was used to determine which mode to consider the correct score.

Intra-observer agreement

Three observers did not provide a score for one clip on their second scoring session (all differed on the clips not scored). The individual's scoring for that clip were not included in the analysis for intra-observer agreement.

For all 24 observers, the mean exact agreement between first and second observation was 66.0 per cent with a 95 per cent CI of 61.9–70.1 per cent; it was 68.0 per cent (61.7–74.3 per cent) for researchers, 63.3 per cent (51.7–74.9 per cent) for clinicians and 66.8 per cent (60.9–71.7 per cent) for students (table 2). Agreement within one score (with 95 per cent confidence in brackets) was achieved as follows for researchers, clinicians and students: 98.4 per cent (96.8–100 per cent), 97.5 per cent (95.0–100 per cent) and 98.7 per cent (97.1–100 per cent). The clinicians achieved 99.7 per cent agreement within two scores; the researchers and students achieved 100 per cent agreement within two scores. The clinicians achieved 100 per cent agreement within three scores.

Table 2 Per cent exact agreement between locomotion scores given during sessions 1 and 2 (and within 1 and 2 points) for each observer

	Intra-observer agreement (%)		
Observer	Exact agreement	±1 score agreement	±2 score agreement
Researcher 1	56.4	94.9	100.0
Researcher 2	60.0	97.5	100.0
Researcher 3	65.0	100.0	100.0
Researcher 4	67.5	100.0	100.0
Researcher 5	67.5	97.5	100.0
Researcher 6	75.0	100.0	100.0
Researcher 7	77.5	97.5	100.0
Researcher 8	77.5	97.5	100.0
Mean (SD)	68.0 (7.5)	98.4 (1.9)	
Clinician 1	40.0	92.5	100.0
Clinician 2	45.0	95.0	97.5
Clinician 3	64.1	100.0	100.0
Clinician 4	65.0	95.0	100.0
Clinician 5	67.5	100.0	100.0
Clinician 6	70.0	100.0	100.0
Clinician 7	75.0	97.5	100.0
Clinician 8	80.0	100.0	100.0
Mean (SD)	63.3 (13.9)	97.5 (3.0)	99.7 (0.9)
Student 1	57.5	100.0	100.0
Student 2	60.0	100.0	100.0
Student 3	61.5	94.9	100.0
Student 4	62.5	100.0	100.0
Student 5	70.0	100.0	100.0
Student 6	72.5	100.0	100.0
Student 7	75.0	97.5	100.0
Student 8	75.0	97.5	100.0
Mean (SD)	66.8 (7.1)	98.7 (1.9)	
Mean of all observers (SD)	66.0 (9.8)	98.2 (2.3)	99.9 (0.5)
Difference in means between two groups (95% CI)			
Researcher–clinician	4.7 (−11.8 to 21.1)	0.9 (−2.4 to 4.3)	0.3 (−0.4 to 1.1)
Researcher–student	1.2 (−10.4 to 12.9)	−0.3 (−2.1 to 1.5)	
Clinician–student	−3.4 (−13.6 to 6.7)	−1.23 (−4.0 to 1.5)	−0.3 (−1.1 to 0.4)

Means and SD presented. Means of each group compared with paired t-tests and presented with 95% CIs.

The mean weighted Kappa value for agreement between first and second observation was 0.84 with a 95 per cent CI of 0.78–0.89 for researchers, 0.81 (0.73–0.89) for clinicians and 0.84 (0.82–0.86) for students (see also table 3). As shown in table 4, there may be some systematic bias between observations for some observers (examples could include researchers 1 and 2, clinician 6 and students 3, 6 and 7).

Inter-observer agreement

The mean exact agreement percent was 61.6 (95 per cent CI 59.5 to 63.7) for researchers, 57.6 (50.3 to 64.9) for clinicians and 54.6 (51.6 to 57.7) for students (see also table 5). The AC2 values were 0.70 (unadjusted 0.81, 95 per cent CI 0.76 to 0.86), 0.69 (unadjusted 0.80, 95 per cent CI 0.77 to 0.84) and 0.64 (unadjusted 0.75, 95 per cent CI 0.69 to 0.81) for researchers, clinicians and students, respectively (table 5). The overall adjusted

Table 3 Weighted Kappa values for each observer's agreement between sessions 1 and 2

Observer	Intra-observer weighted Kappa (95% CI)	Classification
Researcher 1	0.75 (0.61 to 0.90)	Substantial
Researcher 2	0.78 (0.66 to 0.91)	Substantial
Researcher 3	0.77 (0.66 to 0.88)	Substantial
Researcher 4	0.87 (0.80 to 0.95)	Almost perfect
Researcher 5	0.83 (0.72 to 0.94)	Almost perfect
Researcher 6	0.90 (0.83 to 0.97)	Almost perfect
Researcher 7	0.91 (0.83 to 0.98)	Almost perfect
Researcher 8	0.90 (0.84 to 0.97)	Almost perfect
Mean (SD)	0.84 (0.07)	Almost perfect
Clinician 1	0.63 (0.46 to 0.80)	Substantial
Clinician 2	0.69 (0.50 to 0.88)	Substantial
Clinician 3	0.83 (0.74 to 0.93)	Almost perfect
Clinician 4	0.80 (0.68 to 0.93)	Almost perfect
Clinician 5	0.85 (0.77 to 0.94)	Almost perfect
Clinician 6	0.88 (0.82 to 0.95)	Almost perfect
Clinician 7	0.88 (0.79 to 0.97)	Almost perfect
Clinician 8	0.90 (0.83 to 0.98)	Almost perfect
Mean (SD)	0.81 (0.10)	Almost perfect
Student 1	0.83 (0.74 to 0.92)	Almost perfect
Student 2	0.82 (0.74 to 0.91)	Almost perfect
Student 3	0.81 (0.69 to 0.92)	Almost perfect
Student 4	0.83 (0.74 to 0.92)	Almost perfect
Student 5	0.85 (0.77 to 0.93)	Almost perfect
Student 6	0.88 (0.81 to 0.96)	Almost perfect
Student 7	0.86 (0.76 to 0.97)	Almost perfect
Student 8	0.84 (0.71 to 0.96)	Almost perfect
Mean (SD)	0.84 (0.02)	Almost perfect
Difference in group means of weighted Kappa between two groups (95% CI)		
Researcher–clinician		0.03 (−0.02 to 0.08)
Researcher–student		−0.00 (−0.04 to 0.004)
Clinician–student		−0.03 (−0.11 to 0.04)

Means of each group compared with paired t-tests. Classification based on Landis and Koch.[14]

Table 4 Mean difference between locomotion scores given during first and second sessions and results of one-sample t-tests

Observer	Intra-observer difference	
	Mean difference between first and second observation	P value of one-sample t-test of mean difference between observations and zero
Researcher 1	−0.28	0.02
Researcher 2	−0.28	0.01
Researcher 3	0.15	0.11
Researcher 4	−0.08	0.41
Researcher 5	0.00	1.00
Researcher 6	−0.05	0.53
Researcher 7	0.10	0.25
Researcher 8	0.05	0.53
Mean (SD)	−0.05 (0.16)	
Clinician 1	−0.13	0.39
Clinician 2	−0.08	0.61
Clinician 3	0.05	0.60
Clinician 4	0.00	1.00
Clinician 5	−0.13	0.17
Clinician 6	0.20	0.02
Clinician 7	0.03	0.79
Clinician 8	−0.10	0.16
Mean (SD)	−0.02 (0.11)	
Student 1	−0.03	0.81
Student 2	0.10	0.32
Student 3	0.33	0.00
Student 4	−0.03	0.80
Student 5	0.10	0.25
Student 6	−0.18	0.03
Student 7	−0.18	0.05
Student 8	−0.08	0.41
Mean (SD)	0.00 (0.17)	

AC2 value for all observers was 0.72 (unadjusted 0.75, 95 per cent CI 0.69 to 0.81).

The adjusted AC2 values created for each locomotion score are displayed in table 6. They show almost perfect or substantial agreement for videos scoring either zero or three (as determined by the mode score). There was substantial or moderate agreement for videos scoring two, and substantial agreement for videos scoring one according to the interpretations determined by Landis and Koch[14]: <0.00=poor; 0.00–0.20=slight; 0.21–0.40=fair; 0.41–0.60=moderate; 0.61–0.80=substantial; 0.81–1.00=almost perfect.

Discussion

Locomotion scoring is currently relied on in the livestock sector, both to identify lame animals and to determine a herd level prevalence, including enabling benchmarking. Although locomotion scoring is criticised for being subjective, this subjectivity can be reduced by using a scoring system with good reliability, both by the same scorer when scoring on different occasions, and by different scorers scoring the same cattle. Lack of knowledge of the reliability of a scoring system makes it difficult to fully acknowledge its subjectivity.

This study has assessed the reliability of the proposed beef locomotion scoring system, that is, its consistency. However, it should be emphasised that it has not assessed the validity of the scoring system, which still needs testing. Neither inter-reliability or intra-reliability addresses the issue of accuracy because observers can consistently agree with each other, and themselves on different occasions, and still be wrong.

When using this locomotion scoring system, researchers, clinicians and students achieved at least substantial agreement in both the intra-observer and inter-observer assessments with all results greater than 0.61 (classed as 'substantial' according to Landis and Koch[14]). This suggests that if the same observer scores the clips on different occasions, or if different observers score the clips, over the 40 clips they could expect to achieve substantial agreement. However, at the level of each score (table 6), scores zero and three show almost perfect or substantial agreement, with score one showing substantial agreement and score two showing moderate or substantial agreement. This indicates that there is less agreement between observers over the actual locomotion score categories. This also shows that most disagreement is likely to be around score one

Table 5 Mean exact agreement and Gwet's AC2 for each group of observers (researchers, clinicians and students) and for all observers combined (AC2 values adjusted for critical values*)

Observer	Inter-observer	
	Mean % exact agreement	Gwet's AC2/ classification
Researcher 1	61.1	
Researcher 2	65.0	
Researcher 3	58.2	
Researcher 4	62.1	
Researcher 5	58.2	
Researcher 6	61.4	
Researcher 7	62.5	
Researcher 8	64.3	
Mean^1 (SD or 95% CI)	61.6 (2.5)	0.81 (0.76 to 0.86)
Adjusted AC2		0.70/Substantial
Clinician 1	38.6	
Clinician 2	55.0	
Clinician 3	61.8	
Clinician 4	53.6	
Clinician 5	60.4	
Clinician 6	63.9	
Clinician 7	65.4	
Clinician 8	62.1	
Mean^2 (SD or 95% CI)	57.6 (8.7)	0.80 (0.77 to 0.84)
Adjusted AC2		0.69/Substantial
Student 1	56.8	
Student 2	60.0	
Student 3	47.1	
Student 4	54.3	
Student 5	56.4	
Student 6	53.6	
Student 7	54.3	
Student 8	54.6	
Mean^3 (SD or 95% CI)	54.6 (3.7)	0.75 (0.69 to 0.81)
Adjusted AC2		0.64/Substantial
All observers mean (95% CI)		0.79 (0.75 to 0.82)
Adjusted AC2		0.72/Substantial
Difference in means between two groups with 95% CI in brackets		
Researcher^1–clinician^2	4.0 (–3.7 to 11.7)	
Researcher^1–student^3	7.0 (4.4 to 9.5)	
Clinician^2–student^3	3.0 (–6.0 to 11.9)	

Means of each group compared with paired t-tests. Classification based on Landis and Koch.[14]
*Critical value for all 24 observers=0.07, critical value for 8 observers=0.11.[10]

and two, and as such care should be given when scoring animals believed to be in these categories. In veterinary practice, it is generally considered important to lift the feet of animals equivalent to the score two and three descriptors and treat them appropriately. Therefore, on an individual animal basis, where an observer is unsure if an animal is a score one or score two, we suggest that it may be worthwhile to take one of two options, with an aim to reduce the risk of missing lame animals: (1) score these unsure animals as a two, ensuring that they have their feet lifted and are treated if appropriate, or (2) create a new category of 'unsure', requiring a timely re-score.

The observers were all provided with training before watching the scoring videos. Although some evidence suggests that training can improve agreement for on-farm scoring systems,[15] there is also some evidence to suggest that training may not lead to much improvement in intra-observer or inter-observer agreement for locomotion scores,[16 17] but more scoring sessions, that is, more experience, may lead to improvements in inter-observer agreement.[18] If further experience of using the system, for example, a number of practice clips that could be scored (with answers being shown afterwards) had been provided, it may have led to improved inter-observer agreement. This is also demonstrated by evidence indicating that experienced observers perform better than inexperienced observers.[19]

This scoring system has not been studied with farmer observers. This would be worthwhile future work. The observers used for this study were not a random sample, and this may be a limitation of the study. Due caution should therefore be taken when extrapolating results to the wider population. In particular, the clinicians selected were all experienced veterinary surgeons undertaking further qualifications. It may be that less experienced clinicians (eg, new graduate clinicians) may not be as reliable. However, the veterinary students studied showed almost perfect intra-observer agreement, and only slightly lower inter-observer agreement than the researcher group (AC2: researcher value of 0.70 compared with a student value of 0.64), yet still substantial agreement with each other. However, when looking at the level of individual locomotion scores (table 6), there was a slight trend towards lower AC2 values than the researchers and clinicians suggesting that experience may lead to improved agreement on each specific locomotion score category.

The exact agreement between sessions was generally high (mean=66.0 per cent (SD 9.8) for all observers). However, the range is quite wide (40.0 per cent to 80.0 per cent) as there were a number of outliers that are likely to have skewed the results (eg, clinicians 1 and

Table 6 Inter-observer agreement coefficient (Gwet's AC2) for researchers, clinicians, students and all 24 observers combined (AC2 values adjusted with critical values*)

Locomotion score	AC2 for all observers combined		AC2 for researchers		AC2 for clinicians		AC2 for students	
0 (95% CI)	0.81 (0.86 to 0.94)	Almost perfect	0.73 (0.84 to 0.97)	Substantial	0.73 (0.87 to 0.95)	Substantial	0.71 (0.84 to 0.94)	Substantial
1 (95% CI)	0.72 (0.75 to 0.88)	Substantial	0.61 (0.67 to 0.91)	Substantial	0.66 (0.79 to 0.90)	Substantial	0.62 (0.66 to 0.95)	Substantial
2 (95% CI)	0.76 (0.78 to 0.93)	Substantial	0.73 (0.86 to 0.97)	Substantial	0.71 (0.81 to 0.98)	Substantial	0.53 (0.56 to 0.86)	Moderate
3 (95% CI)	0.88 (0.99 to 1)	Almost perfect	0.80 (0.94 to 1)	Substantial	0.78 (0.92 to 0.99)	Substantial	0.78 (0.91 to 1)	Substantial

Classification of adjusted values based on Landis and Koch.[14] NB. 95% CI refers to the unadjusted AC2 values, therefore adjusted AC2 point estimates may not fall within the unadjusted 95% CI.
*Critical value for all 24 observers=0.09, critical value for 8 observers=0.18.[10]

2). This suggests that some observers are not as good as others, and perhaps before individuals use this scoring system in practice, they should test their own agreement (precision). The videos used in this study can be made into a package for this use, and if individuals find that their intra-observer agreement is poor, they may want to practice and train before reattempting the package with the aim of increasing their intra-observer agreement. Systematic bias between attempts could also be identified and controlled. Inter-observer agreement could also be assessed in the same way in clinicians working across the same farms to ensure that they are scoring similarly.

On the second scoring session, there was some evidence to support the notion that some observers had systemic bias in how they scored. However, these were in different directions (some increased their mean scores, and some decreased their mean scores), and only small mean changes were made. This suggests some bias in terms of systematically increasing or decreasing the scores between sessions one and two. In the authors' opinion, this bias is small and unlikely to have a detrimental impact on the assessment of the scoring system.

The video clips used were variable in length. The authors felt that this reflected on-farm locomotion scoring, where on occasions, scorers will need to score quickly. As all observers scored the same clips, and as it was possible to watch the clips as many times as required, the authors do not believe that this negatively affects the assessment of the scoring system.

The authors have now used the system for research purposes and added a fifth point[20] to enable differentiation of severely lame animals from those who have non–weight-bearing limbs. However, this was considered not clinically relevant, as a score 3 and a score 4 would both constitute severe lameness, warranting examination and suitable treatment. For practical use, the authors would recommend using the zero to three system described in this study.

There is some disagreement regarding the categories from Landis and Koch.[14] Some suggest higher scores should be achieved before agreement is considered 'substantial' or 'almost perfect'. For this reason, all values have been provided so that readers can interpret as required. However, in the authors' opinion, the intra-observer and inter-observer agreement across the 40 video clips is considered acceptable when compared with similar studies in the literature.[5 16 21–23]

Acknowledgements The authors would like to thank the farmers who allowed their cattle to be filmed, the clinicians, researchers and students who took part in this study, as well as Denis Duret and Bob Read at The University of Liverpool who provided assistance with the online platform.

Funding The authors would like to thank the Animal Welfare Foundation who funded this research under the Norman Hayward Fund (grant ref. NHF_2016_05_KM).

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval University of Liverpool Ethics Committee (VREC533a).

Data availability statement Data are available on reasonable request. The videos created during this study are available on request.

ORCID iDs
Karin Mueller http://orcid.org/0000-0002-0674-8007
Helen Mary Higgins http://orcid.org/0000-0003-0706-1976

References

1 Farm Animal Welfare Council. Opinion on the welfare of the dairy cow [Internet], 2009. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/325044/FAWC_opinion_on_dairy_cow_welfare.pdf
2 Cattle Health and Welfare Group. GB Catle Health & Welfare Group Fourth Report [Internet], 2018. Available: http://beefandlamb.ahdb.org.uk/wp-content/uploads/2018/10/CHAWG-Fourth-Report-2018.pdf
3 Farm Animal Welfare Committee. Opinion on the welfare of cattle kept for beef production [Internet], 2019. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/777246/FAWC_Opinion_on_the_welfare_of_cattle_kept_for_beef_production.pdf
4 Agriculture and Horticulture Development Board. Dairy Mobility Score laminate [Internet]. 2015 [cited 2019 Nov 10]. p. 0–2. Available: https://dairy.ahdb.org.uk/resources-library/technical-information/health-welfare/mobility-score-instructions/#.XchiMXd2uCQ
5 Angell JW, Cripps PJ, Grove-White DH, *et al*. A practical tool for locomotion scoring in sheep: reliability when used by veterinary surgeons and sheep farmers. *Vet Rec* 2015;176:521.
6 Manson FJ, Leaver JD. The influence of concentrate amount on locomotion and clinical lameness in dairy cattle. *Anim Sci* 1988;47:185–90.
7 Sprecher DJ, Hostetler DE, Kaneene JB. A lameness scoring system that uses posture and gait to predict dairy cattle reproductive performance. *Theriogenology* 1997;47:1179–87.
8 Wasserstein RL, Lazar NA. The ASA Statement on *p*-values: context, process, and purpose. *Am Stat* 2016;70:129–33.
9 Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "p‹0.05". *Am Stat* 2019;73:1–19.
10 Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968;70:213–20.
11 Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. *Br J Math Stat Psychol* 2008;61:29–48.
12 Gwet K. Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among raters. Advanced Analytics LLC, 2010.
13 Kottner J, Audigé L, Brorson S, *et al*. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *J Clin Epidemiol* 2011;64:96–106.
14 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74.
15 Vasseur E, Gibbons J, Rushen J, *et al*. Development and implementation of a training program to ensure high repeatability of body condition scoring of dairy cows. *J Dairy Sci* 2013;96:4725–37.
16 Thomsen PT, Munksgaard L, Tøgersen FA. Evaluation of a lameness scoring system for dairy cows. *J Dairy Sci* 2008;91:119–26.
17 Garcia E, König K, Allesen-Holm BH, *et al*. Experienced and inexperienced observers achieved relatively high within-observer agreement on video mobility scoring of dairy cows. *J Dairy Sci* 2015;98:4560–71.
18 Brenninkmeyer C, Dippel S, March S, *et al*. Reliability of a subjective lameness scoring system for dairy cows. *Anim Welf* 2007;16:127–9.
19 Schlageter-Tello A, Bokkers E, Groot Koerkamp P, *et al*. Comparison of locomotion scoring for dairy cows by experienced and inexperienced raters using live or video observation methods. *Anim Welf* 2015;24:69–79.
20 Tunstall J, Mueller K, Grove White D, *et al*. Lameness in beef cattle: UK farmers' perceptions, knowledge, barriers, and approaches to treatment and control. *Front Vet Sci* 2019;6:1–14.
21 Schlageter-Tello A, Van Hertem T, Bokkers EAM, *et al*. Performance of human observers and an automatic 3-dimensional computer-vision-based locomotion scoring method to detect lameness and hoof lesions in dairy cows. *J Dairy Sci* 2018;101:6322–35.
22 Kaler J, Wassink GJ, Green LE. The inter- and intra-observer reliability of a locomotion scoring scale for sheep. *Vet J* 2009;180:189–94.
23 Vanhoudt A, Yang DA, Armstrong T, *et al*. Interobserver agreement of digital dermatitis M-scores for photographs of the hind feet of standing dairy cattle. *J Dairy Sci* 2019;102:5466–74.