J. Dairy Sci. 106:5715–5722
https://doi.org/10.3168/jds.2022-22737

Crowdsourcing remote comparative lameness assessments for dairy cattle

Kehan Sheng,[1]* Borbala Foris,[1]* Marina A. G. von Keyserlingk,[1] John Gardenier,[2] Cameron Clark,[3] and Daniel M. Weary[1]†
[1]Animal Welfare Program, Faculty of Land and Food Systems, The University of British Columbia, Vancouver, BC, V6T 1Z6, Canada
[2]Australian Centre for Field Robotics, Faculty of Engineering, The University of Sydney, Darlington, NSW 2006, Australia
[3]Livestock Production and Welfare Group, School of Life and Environmental Sciences, Faculty of Science, The University of Sydney, Camden, NSW 2570, Australia

ABSTRACT

Lameness assessments are rarely conducted routinely on dairy farms and when completed typically underestimate lameness prevalence, hampering early diagnosis and treatment. A well-known feature of many perceptual tasks is that relative assessments are more accurate than absolute assessments, suggesting that creating methods that allow for the relative scoring of which cow is more lame will allow for reliable lameness assessments. Here we developed and tested a remote comparative lameness assessment method: we recruited nonexperienced crowd workers via an online platform and asked them to watch 2 videos side-by-side, each showing a cow walking, and to identify which cow was more lame and by how much (on a scale of −3 to 3). We created 11 tasks, each with 10 video pairs for comparison, and recruited 50 workers per task. All tasks were also completed by 5 experienced cattle lameness assessors. We evaluated data filtering and clustering methods based on worker responses and determined the agreement among workers, among experienced assessors, and between these groups. A moderate to high interobserver reliability was observed (intraclass correlation coefficient, ICC = 0.46 to 0.77) for crowd workers and agreement was high among the experienced assessors (ICC = 0.87). Average crowd-worker responses showed excellent agreement with the average of experienced assessor responses (ICC = 0.89 to 0.91), regardless of data processing method. To investigate if we could use fewer workers per task while still retaining high agreement with experienced assessors, we randomly subsampled 2 to 43 (1 less than the minimum number of workers retained per task after data cleaning) workers from each task. The agreement with experienced assessors increased substantially as we increased the number of workers from 2 to 10, but little increase was observed after 10 or more workers were used (ICC > 0.80). The proposed method provides a fast and cost-effective way to assess lameness in commercial herds. In addition, this method allows for large-scale data collection useful for training computer vision algorithms that could be used to automate lameness assessments on farm.

Key words: wisdom of the crowd, dairy cow, animal welfare, click worker, gait score

INTRODUCTION

Lameness affects cattle health and welfare (Whay et al., 2003), causing pain that can last weeks or months (Green et al., 2002). Lameness is also associated with economic loss (Willshire and Bell, 2009), affecting farm profitability (Bennett et al., 1999; Bruijnis et al., 2010). Despite this, farmers and veterinarians often underestimate cattle lameness prevalence (Denis-Robichaud et al., 2020) and lameness is often considered a lower priority than other health issues (Leach et al., 2010; Ózsvári, 2017). Lameness is rarely diagnosed in the early stages (Silva et al., 2021), likely due to the lack of sensitive and cost-effective diagnosis methods, and as such early treatment and prevention are rarely implemented (Sadiq et al., 2019). Despite extensive research on the etiology, diagnosis, and treatment of lameness in the last 20 years, little progress has been made globally to reduce lameness prevalence in dairy herds (Cook, 2020), likely due in part to limitations in lameness detection.

Lameness assessments are typically conducted by observers who assign a score to each cow (locomotion score) based on specific traits including gait and body posture (Afonso et al., 2020; reviewed in van Nuffel et al., 2015a). Reliable scoring requires that the observer is trained, and scoring is time-consuming and labor-intensive (Bicalho et al., 2007). As such, locomotion scoring is conducted infrequently on most farms (Leach et al., 2010) and longitudinal lameness assessments are rare even in the scientific literature (reviewed in Alsaaod et al., 2019). Recent work has shown that more frequent assessments allow for more cases to be

Received September 7, 2022.
Accepted February 14, 2023.
*These authors contributed equally to this work.
†Corresponding author: dan.weary@ubc.ca

detected (Eriksson et al., 2020; Sahar et al., 2022), and thus likely benefit early lameness detection and management (O'Leary et al., 2020).

Research into automatic lameness detection using sensors and computer vision systems has gained interest (reviewed in Alsaaod et al., 2019; Qiao et al., 2021), but the cost of these technologies is still prohibitive for most dairy farms (Van De Gucht et al., 2018). In addition, the sensitivity and accuracy of these products are low (Bicalho et al., 2007; reviewed in Dominiak and Kristensen, 2017), and commercially available systems are lacking (reviewed in van Nuffel et al., 2015b; Kang et al., 2021). The limited performance of these technologies could be due, in part, to dependence on visual locomotion scores as the gold standard in the development and validation of these technologies (van Nuffel et al., 2015b). Low interobserver and intraobserver consistency reduces the reliability of the lameness visual locomotion scoring (Schlageter-Tello et al., 2014; Gardenier et al., 2021), especially when performed by inexperienced observers (Schlageter-Tello et al., 2015). A reliable and efficient lameness scoring method would help the collection of large, high-quality data sets needed to train automated lameness assessment methods.

One limitation of current scoring systems is that cows are assessed relative to an absolute standard that an experienced observer has been trained to recognize (e.g., Flower and Weary, 2009). A large body of literature in psychophysics shows that comparison-based assessments are typically more accurate and sensitive than absolute ones (Gulliksen, 1946; Nutter and Esker, 2006). For example, distinguishing the difference in the frequency between 2 musical notes is much easier than recognizing one specific note (Miyazaki, 1995; McDermott et al., 2008). Consistent with this larger literature, recent results suggest that comparison-based lameness assessments (by trained observers) are more reliable than absolute assessments (Gardenier et al., 2021). However, it is not known if the reliability of comparison-based lameness assessment depends on assessor training and experience or if this can also be performed reliably by those without training.

Given the ease of comparison-based lameness assessments, we hypothesized that these could be successfully performed by untrained assessors. To test this hypothesis, we presented to crowd workers, recruited via a commercial online platform, videos of 2 cows walking side-by-side. Workers were asked to score which of the 2 cows they considered to be more lame and by how much. To investigate the applicability of our method in terms of time and cost, we evaluated how the number of crowd workers affected the reliability of the lameness assessment. We predicted that crowd-worker assessments would be similar to those of experienced lameness assessors, and only a few workers per assessment would be needed to achieve reliable results.

MATERIALS AND METHODS

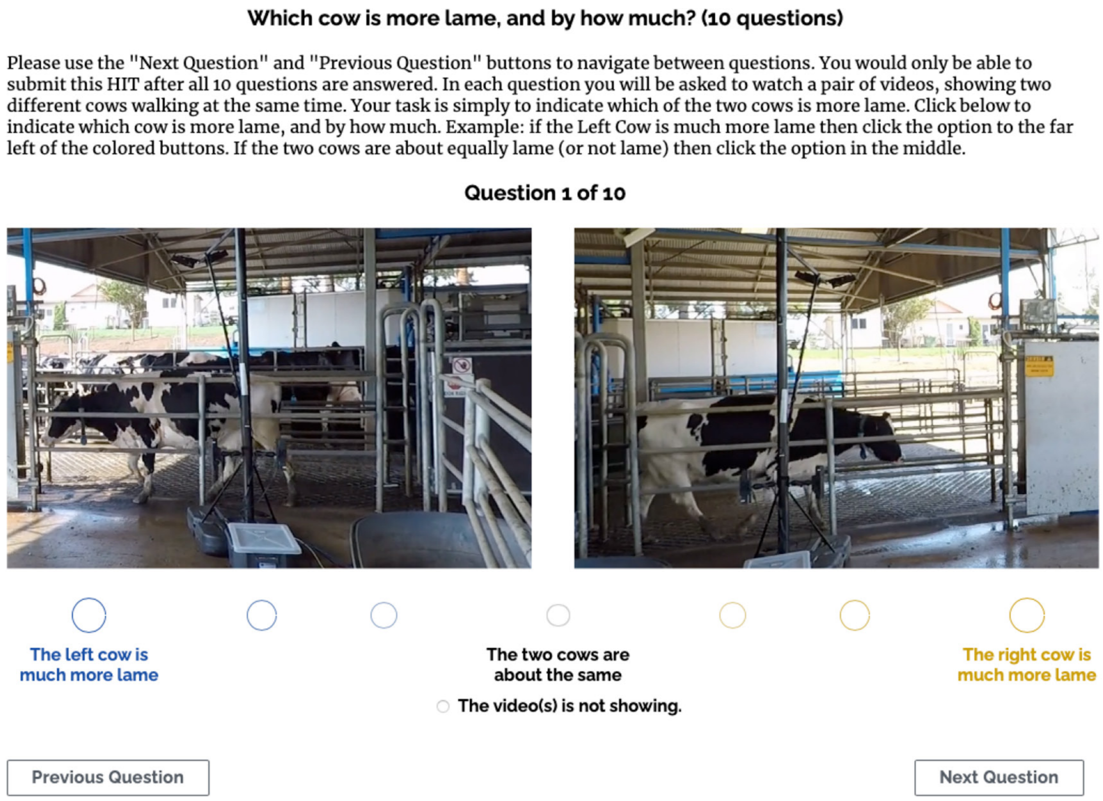
Lameness Assessment

Video Material. No animal ethics permission was required for this study because we used video from an earlier study (Gardenier et al., 2021). For ease of comparison, we used the same videos assessed in an earlier study investigating the reliability of comparison-based lameness scoring (Gardenier et al., 2021). Fifty side-view videos of dairy cows, walking individually into a milking parlor, were used to generate 90 unique video pairs (each showing 2 cows walking side-by-side) for comparative lameness assessment. Of videos used in the current study, 28% showed cows who were clinically lame; this value is reflective of the prevalence of lameness in commercial herds (Denis-Robichaud et al., 2020). Videos were trimmed to exclude frames that did not have cows walking, and were cropped to remove potentially distracting motions in the background and irrelevant objects.

To assess the quality of assessments by the crowd workers we used positive and negative controls. As positive controls, we selected 2 of the 90 pairs for which the difference in lameness was large (i.e., the 2 cows differed by 3 points in the 4-point scale visual locomotion scoring system assessed by the trained observers in Gardenier et al., 2021). As a negative control we simply mirrored one randomly chosen video, such that there was no difference in lameness between the 2 videos.

We created 11 distinct Human Intelligence Tasks (**HIT**), each consisting of a set of 10 pairs of videos to be assessed for relative lameness. Each set of 10 video pairs included one negative and one positive control; the 8 remaining video pairs were tests, selected randomly from the remaining 88 video pairs (i.e., 90 pairs in the full sample, minus the 2 positive controls). The sequence in which the 10 video pairs was presented to participants was randomized.

Assessment Interface. We developed an interface in HTML5 to enable pairwise comparisons between cow videos for lameness assessment. Video pairs were set to autoplay showing the 2 cows walking in opposite directions. The interface (Figure 1) was designed based on feedback regarding user experience by 21 individuals (friends and family naïve to dairy farming) who volunteered for a pilot test. The interface allowed users to click on 1 of 7 bubble buttons to indicate which cow they considered to be more lame and by how much. For analysis, these responses were represented as whole numbers on a scale of −3 to +3, with −3 indicating



Which cow is more lame, and by how much? (10 questions)

Please use the "Next Question" and "Previous Question" buttons to navigate between questions. You would only be able to submit this HIT after all 10 questions are answered. In each question you will be asked to watch a pair of videos, showing two different cows walking at the same time. Your task is simply to indicate which of the two cows is more lame. Click below to indicate which cow is more lame, and by how much. Example: if the Left Cow is much more lame then click the option to the far left of the colored buttons. If the two cows are about equally lame (or not lame) then click the option in the middle.

Question 1 of 10

○ ○ ○ ○ ○ ○ ○

The left cow is much more lame | The two cows are about the same | The right cow is much more lame

○ The video(s) is not showing.

Figure 1. Lameness assessment interface available to Amazon Mechanical Turk crowd workers. For each question, workers were asked to choose which cow they consider more lame and by how much, by clicking on one of the 7 bubbles below the videos. Videos were autoplayed simultaneously on a loop with the 2 cows walking in opposite directions.

that the left cow was much more lame, 0 indicating that the 2 cows are about the same, and +3 indicating that the right cow was much more lame.

Crowd-Worker Assessment. We recruited 50 crowd workers for each HIT using Amazon Mechanical Turk (MTurk). MTurk is an online platform commonly used for collecting a large number of responses. People from around the world sign up to become crowd workers and are paid for completing each HIT. A script written in Python 3.7 (van Rossum and Drake, 2009) was developed to connect to Amazon MTurk through the application programming interface (API; Amazon Web Services Inc., 2020), assign video pairs to each of the 11 HIT, publish HIT, and extract the responses from the crowd workers once completed. All HIT made available on the Amazon MTurk market were claimed within minutes and completed within 2 h after launch. Crowd workers were paid $1 US per HIT (which required approximately 10 min). Human Intelligence Tasks were programmed to be completed on tablets, desktop computers or laptops; we automatically detected devices used and workers were not able to proceed if mobile devices were detected as these screens were deemed too small to perform an accurate assessment. We automatically approved all submitted HIT and answers were automatically extracted along with the total time spent on each HIT. No personal information was collected.

Experienced Assessors. We recruited 5 assessors through professional connections. Each had at least 10 years of experience in cattle lameness scoring and at least one peer-reviewed publication on lameness assessment in dairy cattle. Each experienced assessor completed all 11 HIT. Experienced assessors used the same interface as crowd workers but did so outside the Amazon MTurk environment and were not paid. Answers were recorded and extracted as described for the crowd workers.

Data Processing

Data processing and analysis was performed using R 3.5.3 (R Core Team, 2022).

Cleaning. We discarded responses from crowd workers who did not answer all 10 questions (e.g., due to some videos not playing, on average 3.4 workers/HIT). We also discarded responses from workers who responded identically to all questions (4 workers across all HIT) as this could be caused by workers going through the task without attending to the questions. The remaining responses were then (1) analyzed without further processing, or (2) were subjected to a combination of filtering and clustering as described below.

Filtering and Clustering. To remove low quality responses (i.e., answers from workers who did not pay attention to the task), we used 2 different filtering approaches based on worker responses to the positive (one cow clearly more lame) and negative (no lameness difference) control questions. The weak filter only kept answers from workers who correctly identified the cow that was more lame in the positive control (irrespective of how much more lame). The strong filter only kept responses from workers who correctly picked the cow that was more lame and clicked the button indicating much more lame in the positive control question, and selected no difference in the negative control question.

To remove outlier responses (i.e., responses very different from the majority answers), we applied clustering on the cleaned data and on the data retained after applying the different filtering approaches. For each HIT, we used Euclidean distance-based hierarchical clustering of crowd workers and retained responses only from the largest cluster of workers with similar responses (distance among workers in the retained cluster was less than the mean distance among all workers in the specific HIT).

Statistical Analysis

First, we calculated the median, minimum, and maximum number of workers retained across 11 HIT after applying each filtering and clustering combination.

Second, we assessed the agreement among and between experienced assessors and crowd workers. We used the irr (Gamer et al., 2019) package to calculate intraclass correlation coefficient (**ICC**) with a 2-way random effect model focusing on agreement. As all 5 experienced assessors completed all of the 11 HIT, we used ICC based on the full data set to describe the interobserver agreement among them. For crowd workers, we calculated the ICC separately for each HIT to assess the interobserver agreement among workers who completed that particular set of questions. We also calculated the ICC to assess the agreement between average responses from experienced assessors and crowd workers in each HIT, again considering all filtering and clustering combinations. To assess the overall performance of our method in correctly determining which cow in each pair was more lame, we calculated the ICC between the average responses of crowd workers and experienced assessors using all 88 test video pairs.

Third, using responses after data cleaning but without filtering or clustering, we investigated if we could use fewer workers per HIT while maintaining the high agreement between the average responses of workers and experienced assessors. We evaluated the changes in agreement between workers and assessors as we decreased the number of workers used per HIT from 43 (1 less than the minimum number of workers retained after data cleaning in any HIT) to 2, resulting in 42 levels in total. The workers retained at each level were generated through random sampling, extracting 100 random samples at each level. We then averaged the responses from these random samples and assessed the correlation between the number of workers per HIT and the agreement between experienced assessors and crowd workers across the complete sample of 88 test questions.

RESULTS

Workers assessed all video pairs in a HIT in 10.6 ± 7.5 min (mean ± SD). Data cleaning, filtering and clustering reduced the number of workers from 50 to between 17 and 42 workers per HIT, depending upon the method (Table 1). Responses from the 5 experienced

Table 1. Effect of filtering and clustering on the number of crowd workers retained, among-worker agreement, and the agreement between experienced assessors and workers[1]

Processing method	No. of workers per HIT (minimum; median; maximum)	Agreement among workers in a HIT (ICC; mean ± SD)	Agreement between worker and experienced assessor average in a HIT (ICC; mean ± SD)
No filter	44; 46; 49	0.46 ± 0.15	0.84 ± 0.13
No filter + clustering[2]	37; 40; 42	0.65 ± 0.17	0.85 ± 0.13
Weak filter[3]	39; 42; 45	0.59 ± 0.14	0.84 ± 0.15
Weak filter[3] + clustering[2]	33; 37; 40	0.70 ± 0.12	0.84 ± 0.14
Strong filter[4]	11; 22; 30	0.64 ± 0.15	0.85 ± 0.14
Strong filter[4] + clustering[2]	5; 18; 27	0.77 ± 0.10	0.86 ± 0.12

[1] Agreement was assessed using the intraclass correlation coefficient (ICC) across each of 11 Human Intelligence Tasks (HIT), each with 10 comparison-based lameness assessment questions answered by 50 Amazon Mechanical Turk crowd workers.

[2] Retaining only the largest cluster of workers with similar responses based on hierarchical clustering.

[3] Only retaining workers who correctly identified the cow that was more lame in the positive control question.

[4] Only retaining workers who correctly selected the cow that was more lame, and clicked the button indicating much more lame in the positive control question, and no difference in the negative control question.

assessors showed very high interobserver agreement (ICC = 0.87, $P < 0.001$). Agreement among workers completing the same HIT was moderate to good; the average ICC across HIT ranged from 0.46 to 0.77 considering all data processing methods. ICC of agreement among workers was lowest when considering all workers and highest when we applied the strong filter combined with clustering (Table 1). The average of crowd-worker and experienced assessor responses showed high agreement with little variation among HIT or among data processing methods (Table 1).

When considering the complete data set of 88 test questions posted for assessment, the average crowd-worker response showed high agreement with the average of experienced assessors (Figure 2), regardless of filtering or clustering, with the ICC ranging between 0.89 and 0.91 across data processing methods (no filter: 0.89, no filter + clustering: 0.90, weak filter: 0.90, weak filter + clustering: 0.91, strong filter: 0.89, strong filter + clustering: 0.90). Agreement between the crowd workers and experienced assessors was highest when the difference between the 2 cows in the pair was large. Conversely, the agreement diminished when the difference between the 2 cows in the pair was minimal, a pattern discernible from the central bulge depicted in Figure 2.

The number of workers considered for a HIT influenced the agreement between worker and experienced assessor responses; however, even a small sample of randomly selected workers showed good agreement with experienced assessors (ICC >0.80) when more than 10 worker responses were averaged (Figure 3).

DISCUSSION

Crowd workers were able to reliably perform comparison-based lameness assessments. In our experiment, 5 lameness assessors with years of experience achieved excellent interobserver reliability when determining which cow was more lame and by how much. Traditional lameness assessments are associated with variable inter-rater reliability (Flower and Weary, 2009), but the current results corroborate the findings of Gardenier et al. (2021) who found that trained observers showed good agreement in comparison-based lameness assessments. In the current study crowd workers only showed moderate interobserver agreement, but when answers were averaged our lameness assessment method yielded high agreement between experienced assessors and crowd workers. Agreement among crowd workers increased after filtering and clustering but this did not further increase the agreement between experienced assessor and crowd-worker averages. Previous work has shown that averaged responses from multiple independent respondents can provide surprisingly accurate estimates (i.e., the wisdom of crowds, Surowiecki, 2005; Navajas et al., 2018).

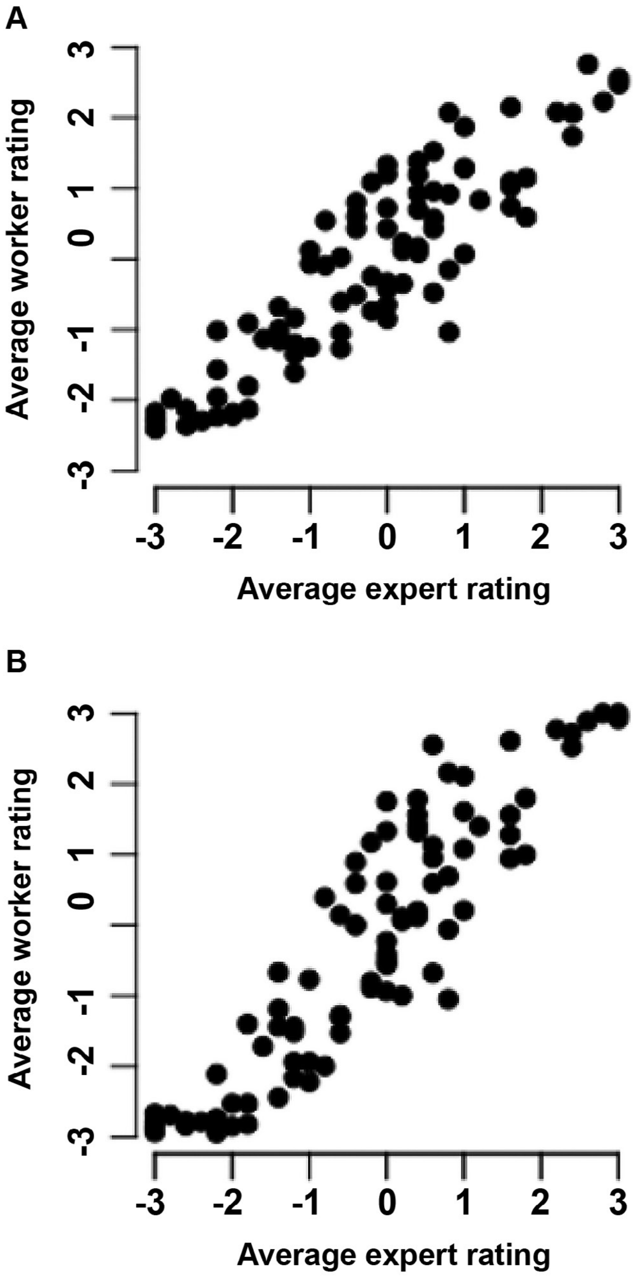


Figure 2. Agreement between the average of the 5 experienced lameness assessors (as shown on the x-axis) and the average of crowd workers (as shown on the y-axis) for each of the 88 test video pairs presented. Values <0 indicate the cow in the left video was more lame and values >0 indicate the cow on the right video was more lame. (A) Average of all worker responses with no filter, intraclass correlation coefficient (ICC) = 0.89. (B) Average of worker responses after strong filtering and clustering, ICC = 0.90.

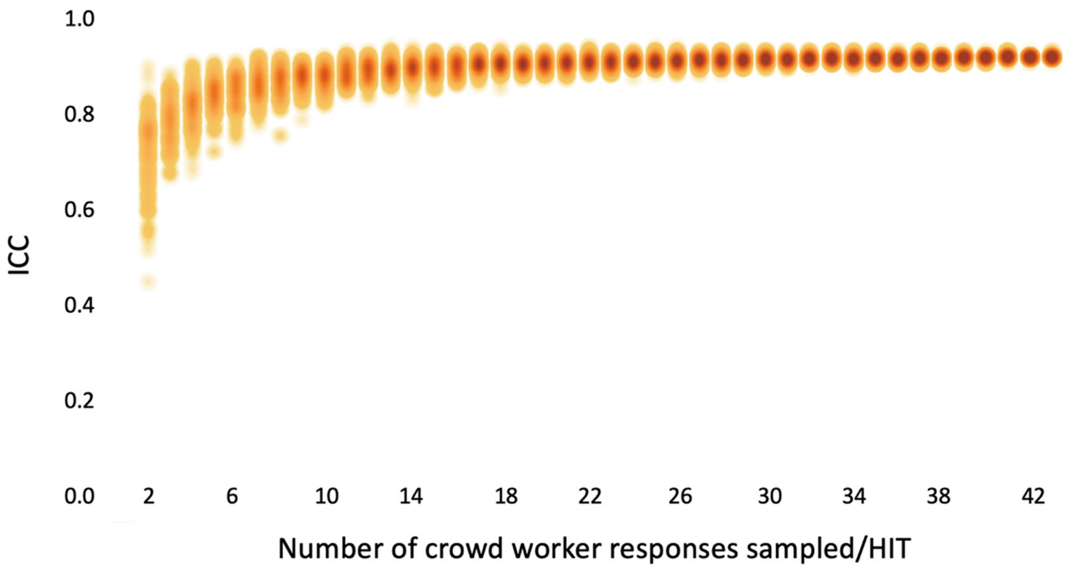


Figure 3. A heatmap showing the relationship between the number of Amazon Mechanical Turk crowd workers sampled per Human Intelligence Tasks (HIT; 10 video pairs per HIT for comparison-based lameness assessment) and the agreement (intraclass correlation coefficient; ICC) between the average of crowd workers and the average of 5 experienced lameness assessor responses based on comparisons from 11 HIT. We evaluated the changes in agreement between worker responses and experienced assessor responses as the number of workers used per HIT decreased from 43 (1 less than the minimum number of workers retained after data cleaning in any HIT) to 2. The workers retained at each level were identified through random sampling, with 100 samples calculated for each level; each sample is illustrated as a point on the graph, with darker colors representing more densely distributed data clouds.

The results of our study demonstrate that reliable relative lameness assessments can be achieved using crowd workers recruited with the MTurk platform. Workers were able to complete the 10 relative comparisons within one task in about 10 min. All HIT were claimed within minutes after release into the market, indicative of a large number of crowd workers motivated to participate in this task. Moreover, our results show that only 10 click workers are needed per HIT to produce responses that are in high agreement with experienced assessors, illustrating that comparison-based lameness detection can be conducted in a rapid and inexpensive manner. Comparison-based scoring using videos have recently been used to assess motor dysfunction in people with multiple sclerosis (Burggraaff et al., 2020). These studies confirmed finer granularity compared with traditional assessments, indicating the potential of this approach for a range of clinical assessments in different species and settings.

The agreement between crowd workers and experienced assessors was greatest with the difference between the 2 cows within a pair was large. Out of 90 comparisons, 26 showed cows with ≥2 points lameness difference (as defined by absolute scores in Gardenier et al., 2021). Detecting small differences in gait is expected to be associated with higher uncertainty; further work is required to determine how the magnitude of lameness difference between cows assessed in a HIT influences the reliability of our method. Assigning more workers to comparisons where initial assessment indicates no or small differences (i.e., comparison score averages between −1 and 1), or to comparisons with high variance in responses, may improve the robustness of the proposed method.

One limitation of our proposed method is that it requires high-quality video recordings of individual cows walking by one at a time; acquiring such videos may not be feasible on some dairy farms. Even when these videos are available, an automatic and accurate method of identifying individual cows would be needed for assessing lameness on a larger scale. Computer vision approaches to automatically detect which cows are lame require large data sets of reliable lameness assessments to train models; we suggest that the approach used here could provide an efficient method of generating this training data.

In this study we performed between-cow comparisons. Given a sufficient number of video comparisons between different cows, a continuous lameness hierarchy could be derived for the entire herd. Numerous methods exist for creating a ranking system based on pairwise comparison results (e.g., insertion sorting algorithm: Biernacki and Jacques, 2013; TrueSkill algorithm: Minka et al., 2018; Elo-rating algorithm: Neumann et al., 2011). A continuous and herd-specific lameness hierarchy could provide more granular information than conventional locomotion scores. Due to the finer granularity, the comparison-based methodology is also well suited to track changes in cow gait over time by comparing the videos of the same cow at different times. The combination of regular between- and within-cow comparisons may allow for earlier and more accurate detection of lameness.

CONCLUSIONS

We investigated if comparison-based cattle lameness assessment is scalable by outsourcing video-based assessments to crowd workers recruited via an online platform. Building upon previous work that found comparison-based lameness assessment more reliable than absolute scoring, we showed that 10–15 naïve crowd workers can perform comparison-based lameness assessments that, on average, agree well with experienced assessors.

ACKNOWLEDGMENTS

Data and the code used for our analysis are available at https://doi.org/10.5683/SP3/3EEQEQ. We thank the 5 experienced assessors for performing the pairwise lameness comparisons, and the staff of Corstorphine farm (Cobbitty, NSW, Australia) for helping acquire the video footage (with the University of Sydney team supported by Dairy Australia and the Dairy Science and Livestock Production and Welfare groups). The Animal Welfare Program at The University of British Columbia is funded by Canada's Natural Sciences and Engineering Research Council Industrial Research Chair Program with contributions from Alberta Milk (Edmonton, AB, Canada), British Columbia Dairy Association (Burnaby, BC, Canada), Boehringer Ingelheim (Burlington, ON, Canada), Lactanet (Sainte-Anne-de-Bellevue, QC, Canada), BC Cattle Industry Development Fund (Kamloops, BC, Canada), Dairy Farmers of Canada (Ottawa, ON, Canada), Dairy Farmers of Manitoba (Winnipeg, MB, Canada), Intervet Canada Corporation (Kirkland, QC, Canada), Saputo Inc. (Montreal, QC, Canada), SaskMilk (Regina, SK, Canada), and Semex Alliance (Guelph, ON, Canada). The authors have not stated any conflicts of interest.

REFERENCES

Afonso, J. S., M. Bruce, P. Keating, D. Raboisson, H. Clough, G. Oikonomou, and J. Rushton. 2020. Profiling detection and classification of lameness methods in British dairy cattle research: A systematic review and meta-analysis. Front. Vet. Sci. 7:542 https://doi.org/10.3389/fvets.2020.00542.

Alsaaod, M., I. Locher, J. Jores, P. Grimm, I. Brodard, A. Steiner, and P. Kuhnert. 2019. Detection of specific treponema species and *Dichelobacter nodosus* from digital dermatitis (Mortellaro's disease) lesions in Swiss cattle. Schweiz. Arch. Tierheilkd. 161:207–215. https://doi.org/10.17236/sat00201.

Amazon Web Services Inc. 2020. Amazon MTurk Application Programming Interface.

Bennett, R. M., K. Christiansen, and R. S. Clifton-Hadley. 1999. Estimating the costs associated with endemic diseases of dairy cattle. J. Dairy Res. 66:455–459. https://doi.org/10.1017/S0022029999003684.

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. https://doi.org/10.3168/jds.2007-0076.

Biernacki, C., and J. Jacques. 2013. A generative model for rank data based on insertion sort algorithm. Comput. Stat. Data Anal. 58:162–176. https://doi.org/10.1016/j.csda.2012.08.008.

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. https://doi.org/10.3168/jds.2009-2721.

Burggraaff, J., J. Dorn, M. D'Souza, C. Morrison, C. P. Kamm, P. Kontschieder, P. Tewarie, S. Steinheimer, A. Sellen, F. Dahlke, L. Kappos, and B. Uitdehaag. 2020. Video-based pairwise comparison: Enabling the development of automated rating of motor dysfunction in multiple sclerosis. Arch. Phys. Med. Rehabil. 101:234–241. https://doi.org/10.1016/j.apmr.2019.07.016.

Cook, N. B. 2020. Symposium review: The impact of management and facilities on cow culling rates. J. Dairy Sci. 103:3846–3855. https://doi.org/10.3168/jds.2019-17140.

Denis-Robichaud, J., D. Kelton, V. Fauteux, M. Villettaz-Robichaud, and J. Dubuc. 2020. Short communication: Accuracy of estimation of lameness, injury, and cleanliness prevalence by dairy farmers and veterinarians. J. Dairy Sci. 103:10696–10702. https://doi.org/10.3168/jds.2020-18651.

Dominiak, K. N., and A. R. Kristensen. 2017. Prioritizing alarms from sensor-based detection models in livestock production - A review on model performance and alarm reducing methods. Comput. Electron. Agr. 133:46–67. https://doi.org/10.1016/j.compag.2016.12.008.

Eriksson, H. K., R. R. Daros, M. A. G. von Keyserlingk, and D. M. Weary. 2020. Effects of case definition and assessment frequency on lameness incidence estimates. J. Dairy Sci. 103:638–648. https://doi.org/10.3168/jds.2019-16426.

Flower, F. C., and D. M. Weary. 2009. Gait assessment in dairy cattle. Animal 3:87–95. https://doi.org/10.1017/S1751731108003194.

Gamer, M., J. Lemon, I. Fellows, and P. Singh. 2019. irr: Various coefficients of interrater reliability and agreement. R package version 0.84.1. https://CRAN.R-project.org/package=irr.

Gardenier, J., J. Underwood, D. M. Weary, and C. E. F. Clark. 2021. Pairwise comparison locomotion scoring for dairy cattle. J. Dairy Sci. 104:6185–6193. https://doi.org/10.3168/jds.2020-19356.

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. https://doi.org/10.3168/jds.S0022-0302(02)74304-X.

Gulliksen, H. 1946. Paired comparisons and the logic of measurement. Psychol. Rev. 53:199–213. https://doi.org/10.1037/h0061673.

Kang, X., X. D. Zhang, and G. Liu. 2021. A review: Development of computer vision-based lameness detection for dairy cows and discussion of the practical applications. Sensors (Switzerland) 21:753. https://doi.org/10.3390/s21030753.

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. https://doi.org/10.1016/j.rvsc.2010.02.014.

McDermott, J. H., A. J. Lehr, and A. J. Oxenham. 2008. Is relative pitch specific to pitch? Psychol. Sci. 19:1263–1271. https://doi.org/10.1111/j.1467-9280.2008.02235.x.

Minka, T., M. Research, R. Cleven, and Y. Zaykov. 2018. TrueSkill 2: An improved Bayesian skill rating system. Microsoft (MSR-TR-2018-8).

Miyazaki, K. 1995. Perception of relative pitch with different references: Some absolute-pitch listeners can't tell musical interval names. Percept. Psychophys. 57:962–970. https://doi.org/10.3758/BF03205455.

Navajas, J., T. Niella, G. Garbulsky, B. Bahrami, and M. Sigman. 2018. Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. https://www.nature.com/articles/s41562-017-0273-4.

Neumann, C., J. Duboscq, C. Dubuc, A. Ginting, A. M. Irwan, M. Agil, A. Widdig, and A. Engelhardt. 2011. Assessing dominance hierarchies: Validation and advantages of progressive evaluation with Elo-rating. Anim. Behav. 82:911–921. https://doi.org/10.1016/j.anbehav.2011.07.016.

Nutter, F. W. Jr., and P. D. Esker. 2006. The role of psychophysics in phytopathology: The Weber-Fechner law revisited. Eur. J. Plant Pathol. 114:199–213. https://doi.org/10.1007/s10658-005-4732-9.

O'Leary, N. W., D. T. Byrne, A. H. O'Connor, and L. Shalloo. 2020. Invited review: Cattle lameness detection with accelerometers. J. Dairy Sci. 103:3895–3911. https://doi.org/10.3168/jds.2019-17123.

Ózsvári, L. 2017. Economic cost of lameness in dairy cattle herds. J. Dairy Vet. Anim. Res. 6. https://doi.org/10.15406/jdvar.2017.06.00176.

Qiao, Y., H. Kong, C. Clark, S. Lomax, D. Su, S. Eiffert, and S. Sukkarieh. 2021. Intelligent perception-based cattle lameness detection and behaviour recognition: A review. Animals (Basel) 11:3033 https://doi.org/10.3390/ani11113033.

R Core Team. 2022. R: A language and environment for statistical computing (3.5.3). R Foundation for Statistical Computing. https://www.R-project.org/.

Sadiq, M. B., S. Z. Ramanoon, W. M. S. Mossadeq, R. Mansor, and S. S. S. Hussain. 2019. Dairy farmers' perceptions of and actions in relation to lameness management. Animals (Basel) 9:270. https://doi.org/10.3390/ani9050270.

Sahar, M. W., A. Beaver, R. R. Daros, M. A. G. von Keyserlingk, and D. M. Weary. 2022. Measuring lameness prevalence: Effects of case definition and assessment frequency. J. Dairy Sci. 105:7728–7737. https://doi.org/10.3168/jds.2021-21536.

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. 2014. Effect of merging levels of locomotion scores for dairy cows on intra- and interrater reliability and agreement. J. Dairy Sci. 97:5533–5542. https://doi.org/10.3168/jds.2014-8129.

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. 2015. Comparison of locomotion scoring for dairy cows by experienced and inexperienced raters using live or video observation methods. Anim. Welf. 24:69–79. https://doi.org/10.7120/09627286.24.1.069.

Silva, S. R., J. P. Araujo, C. Guedes, F. Silva, M. Almeida, and J. L. Cerqueira. 2021. Precision technologies to address dairy cattle welfare: Focus on lameness, mastitis, and body condition. Animals (Basel) 11:2253 https://doi.org/10.3390/ani11082253.

Surowiecki, J. 2005. The Wisdom of Crowds. Knopf Doubleday Publishing Group.

Van De Gucht, T., W. Saeys, J. Van Meensel, A. Van Nuffel, J. Vangeyte, and L. Lauwers. 2018. Farm-specific economic value of automatic lameness detection systems in dairy cattle. J. Dairy Sci. 101:637–648. https://doi.org/10.3168/jds.2017-12867.

van Nuffel, A., I. Zwertvaegher, L. Pluym, S. van Weyenberg, V. M. Thorup, M. Pastell, B. Sonck, and W. Saeys. 2015a. Lameness detection in dairy cows: Part 1. How to distinguish between non-lame and lame cows based on differences in locomotion or behavior. Animals (Basel) 5:838–860.

van Nuffel, A., I. Zwertvaegher, S. van Weyenberg, M. Pastell, V. M. Thorup, C. Bahr, B. Sonck, and W. Saeys. 2015b. Lameness detection in dairy cows: Part 2. Use of sensors to automatically register changes in locomotion or behavior. Animals (Basel) 5:861–885.

van Rossum, G., and F. L. Drake. 2009. Python 3 Reference Manual. CreateSpace.

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. https://doi.org/10.1136/vr.153.7.197.

Willshire, J. A., and N. J. Bell. 2009. An economic review of cattle lameness. Cattle Pract. 17:136–141.

ORCIDS

Kehan Sheng https://orcid.org/0000-0001-6442-5284
Borbala Foris https://orcid.org/0000-0002-0901-3057
Marina A. G. von Keyserlingk https://orcid.org/0000-0002-1427-3152
John Gardenier https://orcid.org/0000-0003-4926-5950
Cameron Clark https://orcid.org/0000-0002-7644-2046
Daniel M. Weary https://orcid.org/0000-0002-0917-3982