The Old School, Brewhouse Hill, Wheathampstead,
Hertfordshire AL4 8AN, UK
www.ufaw.org.uk

Animal Welfare 2020, 29: 81-87
ISSN 0962-7286
doi: 10.7120/09627286.29.1.081

To dock or not to dock? Faecal soiling measurement in sheep

VS Soriano†, FO Stamm†, CA Taconeli‡ and CFM Molento*†

† Animal Welfare Laboratory, Federal University of Paraná, Rua dos Funcionários, 1540, 80035-050, Curitiba, Paraná, Brazil
‡ Department of Statistics, Federal University of Paraná, Rua Cel Francisco Heráclito dos Santos, 100, 81531-980, Curitiba, Paraná, Brazil
* Contact for correspondence: carlamolento@ufpr.br

Abstract

*Faecal soiling is one of the welfare indicators in the AWIN welfare assessment protocol for sheep (*Ovis aries*) and is measured by dag scores. Studies on dag scoring for ewes with docked and undocked tails have given rise to contradictory results. The aim of this study was to compare faecal soiling between ewes with docked and undocked tails and evaluate inter-rater reliability for faecal soiling of ewes. This study was conducted on a farm in Southern Brazil and included 66 undocked and 94 docked ewes. Dag score was recorded by three assessors on a scale of 1 to 5. There was no significant difference on faecal soiling for docked compared to undocked ewes and the median dag score was 3 (1–5). Repeatability amongst assessors by intra-class correlation coefficient of dag scores on docked and undocked ewes was 0.49 and 0.40, respectively; however, these repeatabilities showed no significant differences. The modest repeatability between three assessors on dag scoring indicates caution as regards the use of faecal soiling as an indicator and suggests a need for further studies. The best field results may be obtained by increasing the assessment sample to at least 160 ewes or by raising the number of assessors to five to promote better inter-observer repeatability. Results suggest that tail-docking did not promote cleanliness on the breech area. Considering the negative impact on welfare, it seems reasonable to reverse the burden of proof and desist from recommending tail-docking in the absence of clear scientific evidence of any benefit.*

Keywords: *animal welfare, dag, faecal soiling, intra-class correlation coefficient, inter-observer reliability, sheep welfare indicator*

Introduction

Dag formation is caused by the adhesion of faecal matter to the breech area of sheep and it has been considered a risk factor for cutaneous myiases or flystrikes (Farm Animal Welfare Council [FAWC] 2008). The consistency of the faeces will affect its ability to accumulate into dags, since stools or diarrhoea are potentially able to adhere to wool and pre-existing faeces thereby developing into dags (Waghorn *et al* 1999). Diarrhoea, in turn, may result from endoparasitic infection or nutritional imbalance (Llonch *et al* 2015).

Regarding the negative impact on animals, faecal soiling is an indicator of good health in the AWIN welfare assessment protocol for sheep, and is measured via dag scores (AWIN 2015). Faecal soiling may be considered an indirect trait for susceptibility to flystrike (Australian Wool Innovation Limited and Meat and Livestock Australia 2007). There are three dag-scoring scales used for sheep: 0–5 (Larsen *et al* 1994), 1–5 (Australian Wool Innovation Limited and Meat and Livestock Australia 2007) and 0–4 (AWIN 2015), in which higher values are an indication of dirtier animals.

Removal of a portion of the tail is a common procedure in sheep. The prevention of flystrike and cleanliness are the two main reasons for tail-docking sheep (Morris 2000; Sutherland & Tucker 2011; Sheep Standards and Guidelines 2013), since it has been suggested that it reduces levels of faecal soiling (FAWC 2008). However, it is unclear whether tail-docking reduces faecal soiling in sheep and the scientific evidence to support the importance of tail-docking in preventing flystrike is sparse (since there have been relatively few controlled studies of flystrike in sheep [FAWC 2009; Sutherland & Tucker 2011]) and somewhat contradictory. While some studies have shown increased faecal soiling with relatively longer tails (Scobie *et al* 1999; Fisher & Gregory 2007), another found little effect of tail length on faecal soiling (French *et al* 1994), while Scobie *et al* (1999) were unable to show any relationship between tail length and cleanliness. In addition, intriguingly, Watts and Marchant (1977) reported more faecal soiling on sheep with very short tails. In a Brazilian study, Madeira *et al* (1998) concluded that tail-docking failed to control myiasis in sheep and, in fact, the resultant lesion facilitated the establishment of flystrike since tail-docked flocks showed an infestation rate twice that of intact flocks. Since docking is considered acutely painful, causing a permanent disfigurement to the animal, it is important to consider the rationale behind its use, in order to evaluate if it is necessary (Sutherland & Tucker 2011).

Figure 1

Dag score scale used to evaluate docked and undocked ewes (Stamm 2015).

The first explanation offered to explain conflicting results on faecal soiling related to docking concerned breed differences (Scobie *et al* 2008). A reduction in wool cover around the perineum — known as breech bareness — was found to reduce faecal soiling (Scobie *et al* 2007) and was negatively correlated with dag scores, both phenotypically and genetically (Scobie *et al* 2008). According to another approach, the dag score is used to assess the extent of faecal soiling, ie as an animal welfare indicator. It is worthwhile considering the following characteristics for animal welfare indicators: (i) validity, ie meaningful with respect to animal welfare; (ii) reliability, ie consistent results when used by different assessors; and (iii) feasibility, ie practical for on-farm assessment (FAWC 2009; AWIN 2015). Due to inconsistencies in the literature, it appears advantageous to further study the reliability of dag scoring.

Reliability depends on the repeatability of results (Chen *et al* 2003). Repeatability is defined as the similarity of repeated measurements on one object or subject and is an important aspect when selecting indicators for an overall assessment system (Dalmau *et al* 2010). The inter-rater repeatability concerns the degree to which two or more assessors evaluating the same animals at the same time report similar scores (Dalmau *et al* 2010). In the case of faecal soiling, apart from wool cover type and quantity related to breed, the tail presence may also influence dag score assessment. This seems especially relevant considering that the scales offer no distinction in terms of docking and the first inter-rater reliability study on dag scoring by Munoz *et al* (2018) only showed moderate agreement amongst assessors by means of Kendall's coefficient of concordance. Thus, an opportunity exists to provide further information on dirtiness assessment to enrich knowledge regarding which indicators offer reliable results. The aim of this study was to compare faecal soiling between ewes with docked and undocked tails and to evaluate inter-rater reliability for faecal soiling measurement in ewes.

Materials and methods

This research was approved by the Animal Use Ethics Committee of the Agricultural Campus of Federal University of Paraná, Brazil with reference number: 2.248.306/2017. The study took place between September 2016 and October 2017, with dag score assessment carried out on a farm located in the town of Quaraí, State of Rio Grande do Sul, Southern Brazil in October 2017. It included 160 unsheared, crossbred Corriedale and Ile de France ewes, aged approximately 13 months of age that randomly allocated into one of two groups. Ninety-four ewes had their tails docked (standard practice, on-farm) and 66 ewes were kept intact.

Tail-docking was administered surgically with a scalpel on anaesthetised lambs at one week old. Local anaesthesia consisted of 1.5 ml of lidocaine (2% lidocaine, Anestex FAGRA, Vétoquinol, Mairiporã, São Paulo, Brazil) injected subcutaneously into three sites in the vicinity of the tail, 5 min prior to the start of the procedure. After docking, the tail was sufficiently long for vulval coverage, as recommended by the *Sheep Standards and Guidelines* (2013) in Australia. After tail-docking, ewes were then given an intramuscular injection of sodium diclofenac (Vetflogin, Vallee, Uberlândia, Minas Gerais, Brazil) (50 mg ml^{-1}, 1 mg kg^{-1} bodyweight) into the breech area for analgesia and oxytetracycline (Oxitec, CALBOS Saúde Animal, São José dos Pinhais, Paraná, Brazil) (200 mg ml^{-1}, 1 ml per 10 kg bodyweight) as preventive antibiotic treatment.

The number of ewes in each group was higher than the minimum sample of 60 from a flock with 650 ewes recommended by AWIN Welfare Assessment Protocol for Sheep in the second level, when there is a requirement for the assessment of individual ewes. The general management of ewes was similar, all ewes were raised as one flock in the same pasture since their birth, and the farm was characterised as extensive (EFSA 2014; AWIN 2015), as all ewes were kept on pasture and received supplementation in some periods.

https://doi.org/10.7120/09627286.29.1.081 Published online by Cambridge University Press

Dag scores were evaluated on ewes in accordance with Stamm (2015), using scores from 1 to 5 (Figure 1) on a five-point scale, as per the Australian Wool Innovation Limited and Meat and Livestock Australia (2007) and AWIN welfare assessment protocol for sheep (2015). The illustrative pictures were collected from ewes in South Brazilian farms and the score descriptors were used according to AWIN welfare assessment protocol for sheep (2015). Score 1 was given when no faecal soiling was present, score 2 for a small quantity of faecal matter in the wool around the anus, score 3 for some soiling around the anus and dags in this area only, score 4 for soiling and dags extending beyond the anus to the tail and onto the upper part of the legs, and a score of 5 designated a wider area of soiling, with dags extending down the legs as far as the hocks.

Assessments were performed by one animal scientist (VSS) and two sheep farmers. Assessors scored the animals simultaneously but independently. As assessors were inexperienced in dag score assessment, a picture with dag scores scale with each score description was presented to illustrate dag soiling differences between scores and training was conducted on five random ewes prior to data collection. Assessor 2 scored 89 docked ewes from a total of 94 and 65 undocked ewes from a total of 66; the other two assessors evaluated all the animals in the experiment. In total, 277 assessments were carried out on docked ewes and 197 on undocked ewes.

Statistical analysis

A linear mixed effects model was used to compare docked and undocked ewes regarding dag scores on a 1–5 scale with tail as a fixed effect and animal and assessor as random effects. Based on the first level welfare assessment in AWIN Welfare Assessment Protocol for Sheep (2015), ie flock level, where the first three ratings are considered acceptable cleanliness on the breech area, we tested joint scores as a binary parameter. For this, scores 1, 2 and 3 were deemed as 'acceptable for cleanliness' and scores 4 and 5 'dirty ewes' in order to compare docked and undocked ewes by generalised linear mixed models (Verbeke & Molenberghs 2009).

To study repeatability amongst raters on dag scores, intra-class correlation coefficient (ICC) was used. Repeatability was classified as poor, if $0.00 < ICC < 0.39$; fair, if $0.40 < ICC < 0.59$; good, if $0.60 < ICC < 0.74$; and excellent, if $0.75 < ICC < 1.00$ (Cicchetti & Sparrow 1981). A 95% confidence interval (CI) was obtained by parametric bootstrap and comparison between docked and undocked ewes was made through permutation test, both based on 5,000 simulations.

An additional simulation study was performed to evaluate the impact of sample size and number of assessors on repeatability precision. In this study we have considered ranges from three to 12 assessors and from 20 to 500 ewes were considered as sample size. For each combination of sample size and number of assessors, a total of 1,000 samples were simulated based on linear mixed effects model previously adjusted and the repeatability between assessors was calculated by ICC. Standard deviation was used to evaluate the impact of sample size and number of assessors on repeatability precision. Analyses were performed using R Statistical Computing Environment version 3.4.2 (2017).

Results

There was no significant faecal soiling differences comparing docked and undocked ewes on the 1–5 dag scores scale ($P = 0.733$; Figure 2). Even when scores were combined as a binary parameter, there was no significant difference between docked and undocked ewes ($P = 0.232$).

The median and average score in both situations was 3; out of 277 assessments on docked ewes from three raters, 46% were scored 3 and from 197 assessments on undocked ewes from three raters, 42% were scored 3 (Figure 3). Two was the second most frequently observed dag score on docked ewes (28% of assessments) while 4 was the second most frequent score for undocked ewes (27% of assessments) (Figure 2). From a total of 474 assessments by three raters, 75% of the animals were scored as 1–3 and 25% as 4–5; from all assessments, 44% were scored as 3, 26% were scored as 2 and 24% were scored as 4 (Figure 3). Most ewes were given a score of 3 according to two assessors; however, one assessor scored the majority of the animals as 4. The greatest variation occurred within score 4, based on the difference in numerical range of ewes assessed between assessors in each score (Figure 3).

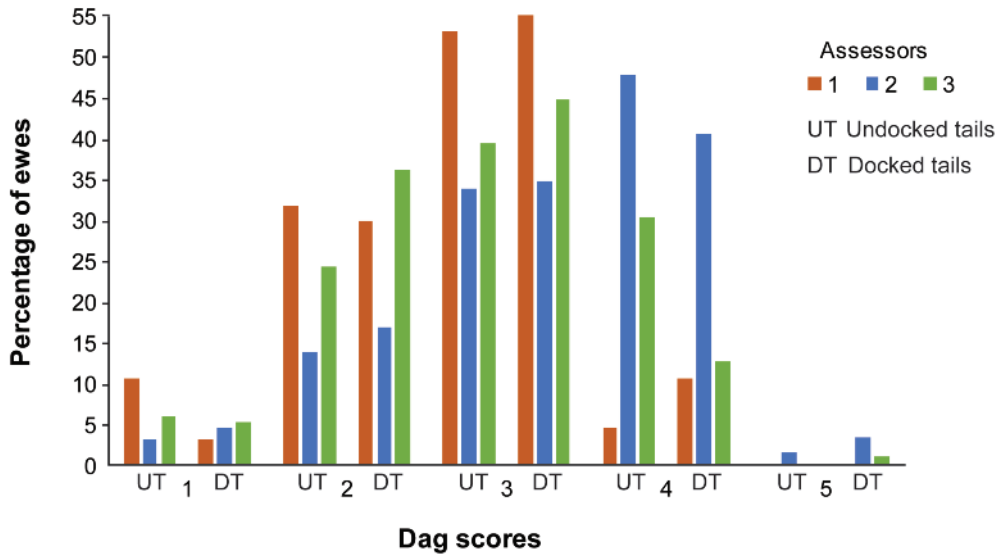
The ICC for dag scores of docked and undocked ewes was 0.49 (95%; CI: 0.33–0.62) and 0.40 (95%; CI: 0.23–0.57), respectively; repeatability did not differ significantly ($P = 0.340$) and amongst raters it was fair. Our simulation revealed that the accuracy of dag score assessment was improved by increases in sample size and number of assessors; however, more than five assessors implied only a slightly more accurate assessment (Figure 4). In addition, the precision had low boosting beyond 160 ewes. Results suggest that when it is not possible to increase the number of animals to be assessed on dirtiness, it is recommended to increase the number of assessors.

Discussion

There was no difference in faecal soiling between docked and undocked ewes in this study, adding further complication to this already contentious issue. Scobie *et al* (1999) found an increase in dag accumulation as the tail stump length got larger, although level of significance depended on the time of the year and management system. Other studies have shown that undocked lambs, ranging in age from 2 to 8 weeks of age, presented a significantly higher mean dag score than docked lambs, on a 6-point scale (Ware *et al* 2000) and increased faecal soiling with relatively longer tails (Fisher & Gregory 2007); however, this last study was based on a questionable sample of ten lambs. Sutherland and Tucker (2011) noted that sheep with very short tails were observed as having more faecal soiling, with diarrhoea more prevalent.

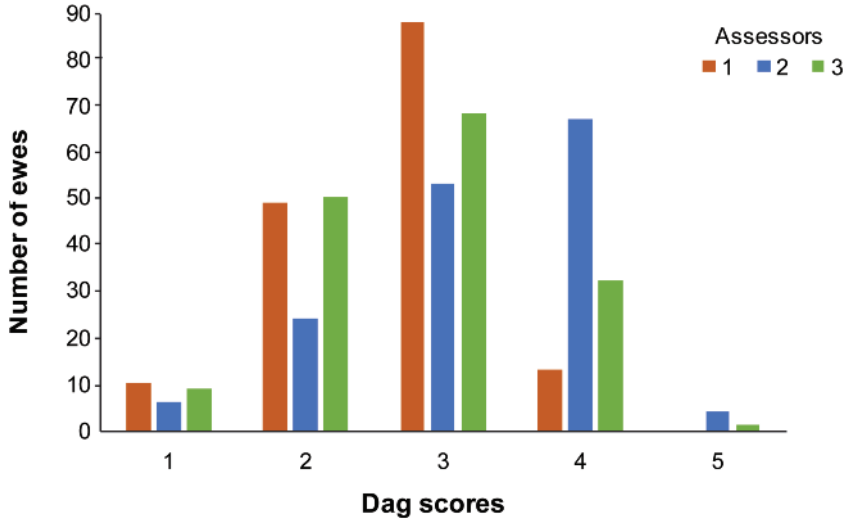
The development of dags begins with faecal adherence to wool, its accumulation promoting faeces-to-faeces adhesion (Waghorn *et al* 1999). In addition to specific breed traits, the factors contributing to dirtiness may be increased rainfall in winter (Karlssom *et al* 2004), gastrointestinal parasites (Waghorn *et al* 1999) and low levels of neutral detergent

Figure 2



Percentage of unsheared, crossbred Corriedale and Ile de France ewes with docked (DT) and undocked tails (UT), assessed on a 1–5 dag score scale by three raters on a farm in Southern Brazil, October 2017.

Figure 3



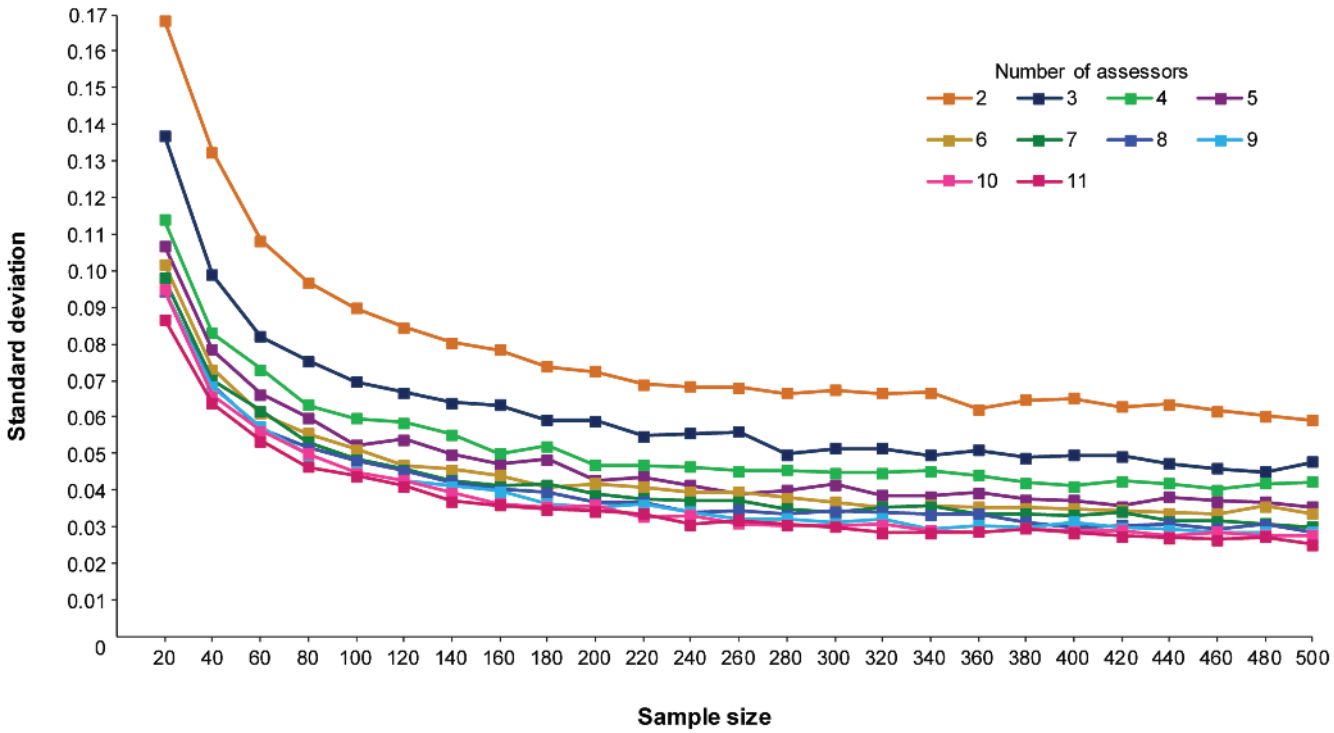
Number of unsheared, crossbred Corriedale and Ile de France ewes assessed on a 1–5 dag score scale from three assessors on a farm in Southern Brazil, October 2017.

fibre, due to the consistency of faeces (Davidson *et al* 2006). Since all ewes were the same mixed-breed, our results indicated that tail-docking had no influence on faecal soiling controlling for the breed effect. Similarly, the presence of a tail did not affect faecal soiling on lambs, according to a 6-point scale in New Zealand, when they were compared at weaning and at transport (Rogers *et al* 2011). Rogers and co-workers also observed that 2.2 was highest average dag score at weaning with 1.2 being the highest at transport; however, breeds differed within docked and undocked groups. Again in New Zealand, Scobie *et al* (2008) scored lambs on a 6-point scale and observed a maximum average dag score of 1.4. It is important to view faecal soiling as a time-dependent measure when carrying out comparisons across studies. Thus, the literature is mostly useful in providing context in terms of an overall scenario regarding faecal soiling in sheep, appearing to indicate relatively low dag scores in lambs.

The majority of studies were performed on lambs. In this regard, the two 5-point scales for dag scores were recommended for animals at least four months old (Australian Wool Innovation Limited and Meat and Livestock Australia

https://doi.org/10.7120/09627286.29.1.081 Published online by Cambridge University Press

Figure 4



Simulation of sample size and number of assessors on the precision of ewe dag scoring on a 1–5 scale, according to the standard deviation of intra-class correlation between assessors.

2007) and for greater than one year of age (AWIN 2015), probably because wool in lambs increases with age and the studies demonstrated low faecal soiling in this age category. The effect of age on scouring and dag in Merino ewes was described by Larsen *et al* (1999). The authors observed a higher prevalence of severe dag (score > 4 on a 1–5 scale), and an increased mean dag score for the two year old ewes, which may be related, in part, to an increased susceptibility to gastrointestinal nematodes at their first lambing. Larsen *et al* (1999) also noted that after two years of age, there was a tendency for the prevalence of severe dag and the mean dag score to decrease. In our study the median dag score was 3 and ewes were not susceptible to immune responses as a result of pregnancy (Barger 1993) or under 12 months of age (Colditz *et al* 1996).

According to Fisher and Gregory (2007), animals with dag scores of 1–3 was considered acceptable but sheep farmers had an aversion to animals scoring 4 and 5. Therefore, if we consider ewes that scored 1–3 as acceptable breech clean ewes, then, according to AWIN Welfare Assessment Protocol for Sheep (2015) at flock level, 75% of ewes assessed in this study presented acceptable dag scores.

Here, a crucial flaw in comparing faecal soiling in both situations was the only fair level of repeatability between assessors. This is not surprising, considering the reasons given by Tuyttens *et al* (2009): (i) short training sessions; and (ii) lack of previous experience in creating dag scores on sheep for the assessors. In a review of animal-based indicators of sheep welfare, Llonch *et al* (2015) cited two studies (Napolitano *et al* 2011; Stubsjøen *et al* 2011) with high validity and reliability for a diarrhoea indicator used in lambs; no information was provided for adult animals. Napolitano *et al* (2011) showed a 0.46 correlation on a test-retest reliability for the parameter 'other diseases', with diarrhoea or cough, or ocular, nasal, vaginal discharges, or a combination of them — it was considered a fair correlation. Also, Stubsjøen *et al* (2011) assessed body cleanliness as an indicator of diarrhoea, a different measure to our study. The only study that measured inter-rater reliability was that of Munoz *et al* (2018) who found moderate agreement amongst three assessors by means of Kendall's coefficient of concordance (from 0.68 to 0.70), which was an improvement on this study, and was performed after an on-farm training session using 75 ewes with experienced assessors. Kendall's coefficient of concordance may be compared with the present study because it determines inter-rater reliability when there are more than two assessors and it is equivalent to ICC when the assumption of normality is violated (Kraemer 1976). While providing pictures and details about the dag scale to sheep managers is important, Munoz *et al* (2018) highlight the need for training sessions before the assessment.

The inter-rater repeatability in this study was lower than that reported for other welfare indicators, such as Qualitative Behavioural Assessments of sheep (Phythian *et al* 2013); locomotion scoring for sheep (Kaler *et al* 2009) and even dirtiness at hindquarters and udder on ewes (Napolitano *et al* 2009). Our results validate the relevance of improving inter-rater reliability since it is fundamental to the develop-

ment of valid welfare assessment protocols (Wemelsfelder & Mullan 2014) and also because it is an important issue for decisions with considerable consequences for animal welfare, such as the abolition of tail-docking.

Detailed descriptors and other refinements of dag scales may constitute tools to improve dag-scoring reliability. One possibility is the inclusion of number and size of dags, percentage of area covered by soiling, taking into account the tail, and the combination of both traits for each score in the description of the different scores in the scale. This appears to warrant further work, since assessors tend to more readily agree when indicators are clearly defined (Wemelsfelder & Mullan 2014). Alternatively, a continuous modified visual analogue scale could be tested, since the use of ordinal scales for assessment was not the best instrument for all parameters. Tuyttens *et al* (2009) noted that inter-rater reliability was significantly improved for a continuous modified visual analogue scale with multiple anchors, than for a 3-point ordinal scale for scoring lameness in dairy cattle from videos. Moreover, as cited by Tuyttens *et al* (2009), it is important to test the effect of longer training sessions on reliability amongst assessors. March *et al* (2007) concluded that intensive training procedures are required to obtain high inter-rater repeatability with the 5-point gait-scoring system to assess dairy cows. The authors noted that inter-rater reliability by Spearman rank correlation coefficient between two assessors ranged from 0.55 to 0.87 after experience gained in the course of data collection in 46 dairy herds.

Other ways to improve inter-rater repeatability on dag score assessment include raising the sample size and increasing the number of assessors. Improvement in inter-rater repeatability was expected as a result of increases in the number of ewes assessed, because estimates become more precise as the sample size increases (Kufs 2011); this was also observed in this study. Considering labour and time restrictions usually present on field, it seems advisable to assess 160 ewes for best precision since further increases in the number of ewes assessed bring small gains in precision (Figure 4). If the assessment of 160 ewes is not feasible, our results suggest increasing the number of raters. Additionally, considering on-farm feasibility and optimum repeatability, our data showed that more than five assessors implied only a slight increase in accuracy, characterising worthless labour.

Animal welfare implications and conclusion

There is no evidence for differences in faecal soiling for docked and undocked ewes, which suggests no benefit of tail-docking as regards ewe cleanliness. The limitations related to repeatability between assessors on dag scoring suggest caution when using it to justify management decisions. Additionally, improving inter-rater reliability by either raising sample sizes or increasing the number of raters seems essential if faecal soiling is to be used as an argument for defending the practice of tail-docking. Considering the negative impact of the procedure on the welfare of the animals, it seems reasonable to reverse the burden of proof and desist from recommending the docking of ewes in the absence of clear scientific evidence of any benefit.

Acknowledgements

This project was funded by Capes Forensic Sciences, as a grant to the first author. The authors are grateful to the sheep farmers that agreed to participate as well as the contribution from Frank AM Tuyttens.

References

Animal Welfare Indicators (AWIN) 2015 *AWIN welfare assessment protocol for sheep*. https://doi.org/10.13130/AWIN_SHEEP_2015

Australian Wool Innovation Limited and Meat and Livestock Australia 2007 *Visual sheep scores*. Australian Wool Innovation Ltd: Sydney, NSW, Australia

Barger IA 1993 Influence of sex and reproductive status on susceptibility of ruminants to nematode parasitism. *International Journal for Parasitology 23*: 463-469. https://doi.org/10.1016/0020-7519(93)90034-V

Chen WJ, Bonillo C and Lecointre G 2003 Repeatability of clades as a criterion of reliability: a case study for molecular phylogeny of Acanthomorpha (Teleostei) with larger number of taxa. *Molecular Phylogenetics and Evolution 26*: 262-288. https://doi.org/10.1016/S1055-7903(02)00371-8

Cicchetti DV and Sparrow SS 1981 Developing criteria for establishing interrater reliability of specific items: application to assessment of adaptive behavior. *American Journal of Mental Deficiency 86*: 127-137

Colditz IG, Watson DL, Gray GD and Eady SJ 1996 Some relationships between age, immune responsiveness and resistance to parasites in ruminants. *International Journal for Parasitology 26*: 869-877. https://doi.org/10.1016/S0020-7519(96)80058-0

Dalmau A, Geverink NA, Van Nuffel A, Van Steenbergen L, Van Reenen K, Hautekiet V, Vermeulen K, Velarde A and Tuyttens FA 2010 Repeatability of lameness, fear and slipping scores to assess animal welfare upon arrival in pig slaughterhouses. *Animal 4*: 804-809. https://doi.org/10.1017/S1751731110000066

Davidson BS, Chaplin SJ and Laird C 2006 Effect of fibre supplementation on dag formation and flystrike in sheep grazing spring pastures. *Australian Journal of Experimental Agriculture 46*: 783-786. https://doi.org/10.1071/EA05353

EFSA Panel on Animal Health and Welfare (AHAW) 2014 Scientific opinion on the welfare risks related to the farming of sheep for wool, meat and milk production. *EFSA Journal 12*: 1-128. https://doi.org/10.2903/sp.efsa.2015.EN-738

Farm Animal Welfare Council (FAWC) 2008 *FAWC report on the implications of castration and tail docking for the welfare of lambs*. FAWC: London, UK

Farm Animal Welfare Council (FAWC) 2009 *Report on farm animal welfare in Great Britain: past, present and future*. FAWC: London, UK

Fisher MW and Gregory NG 2007 Reconciling the differences between the length at which lambs' tails are commonly docked and animal welfare recommendations. *Proceedings of the New Zealand Society of Animal Production 67*: 32-38

French NP, Wall R and Morgan KL 1994 Lamb tail docking: a controlled field study of the effects of tail amputation on health and productivity. *Veterinary Record 134*: 463-467. https://doi.org/10.1136/vr.134.18.463

https://doi.org/10.7120/09627286.29.1.081 Published online by Cambridge University Press

Kaler J, Wassink GJ and Green LE 2009 The inter- and intra-observer reliability of a locomotion scoring scale for sheep. *The Veterinary Journal 180*: 189-194. https://doi.org/10.1016/j.tvjl.2007.12.028

Karlsson LJE, Pollott GE, Eady SJ, Bell A and Greeff JC 2004 Relationship between faecal worm egg counts and scouring in Australian Merino sheep. *Animal Production in Australia 25*: 100-103

Kraemer HC 1976 The small sample nonnull properties of Kendall's Coefficient of Concordance for normal populations. *Journal of the American Statistical Association, 71*: 608-613. https://doi.org/10.1080/01621459.1976.10481536

Kufs C 2011 *Stats with cats: The domesticated guide to statistics, models, graphs, and other breeds of data analysis* p 376. Wheatmark: Tucson, AZ, USA

Larsen JWA, Anderson N, Vizard, AL, Anderson GA and Hoste H 1994 Diarrhoea in Merino ewes during winter: association with trichostrongylid larvae. *Australian Veterinary Journal 7*: 365-372. https://doi.org/10.1111/j.1751-0813.1994.tb00930.x

Llonch P, King EM, Clarke KA, Downes JM and Green LE 2015 A systematic review of animal based indicators of sheep welfare on farm, at market and during transport, and qualitative appraisal of their validity and feasibility for use in UK abattoirs. *The Veterinary Journal 206*: 289-297. https://doi.org/10.1016/j.tvjl.2015.10.019

Madeira NG, Amarante AFT and Padovani CR 1998 Effect of management practices on screw-worm among sheep in São Paulo State, Brazil. *Tropical Animal Health and Production 30*: 149-157. https://doi.org/10.1023/A:1005055518916

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

Morris MC 2000 Ethical issues associated with sheep fly strike research, prevention, and control. *Journal of Agricultural and Environmental Ethics 13*: 205-217. https://doi.org/10.1023/A:1009541810740

Munoz C, Campbell A, Hemsworth P and Doyle R 2018 Animal-based measures to assess the welfare of extensively managed ewes. *Animals 8*: 1-16. https://doi.org/10.3390/ani8010002

Napolitano F, De Rosa G, Ferrante V, Grasso F and Braghieri A 2009 Monitoring the welfare of sheep in organic and conventional farms using an ANI 35 L derived method. *Small Ruminant Research 83*: 49-57. https://doi.org/10.1016/j.smallrumres.2009.04.001

Napolitano N, Giuseppe DR, Girolamia A, Scavonea M and Braghieri A 2011 Avoidance distance in sheep: Test-retest reliability and relationship with stockmen attitude. *Small Ruminant Research 99*: 81-86. https://doi.org/10.1016/j.smallrumres.2011.03.044

Phythian C, Michalopoulou E, Duncan J and Wemelsfelder F 2013 Inter-observer reliability of Qualitative Behavioural Assessments of sheep. *Applied Animal Behaviour Science 144*: 73-79. https://doi.org/10.1016/j.applanim.2012.11.011

Rogers AR, Dowling SK and Webster JR 2011 Raising lambs with intact tails to meet retailer welfare requirements: on farm feasibility and farmer perspectives. *Proceedings of the New Zealand Society of Animal Production 71*: 270-274

Scobie DR, Bray AR and O'Connell D 1999 A breeding goal to improve the welfare of sheep. *Animal Welfare 8*: 391-406

Scobie DR, O'Connell D, Morris CA and Hickey SM 2007 A preliminary genetic analysis of breech and tail traits with the aim of improving the welfare of sheep. *Australian Journal of Agricultural Research 58*: 161-167. https://doi.org/10.1071/AR05444

Scobie DR, O'Connell D, Morris CA and Hickey SM 2008 Dag score is negatively correlated with breech bareness score of sheep. *Australian Journal of Experimental Agriculture 48*: 999-1003. https://doi.org/10.1071/EA07397

Sheep Standards and Guidelines 2013 Sheep standards and guidelines: tail docking. https://www.animalwelfare-standards.net.au/files/2011/05/Sheep-Tail-docking-discussion-paper-5.3.13.pdf

Stamm FO 2015 *First glimpse on sheep welfare in the State of Parana and the case of tail docking*. Dissertation, Universidade Federal do Paraná, Curitiba, Brazil

Stubsjøen SM, Hektoen LA, Valle PS, Janczak AM and Zanella AJ 2011 Assessment of sheep welfare using on-farm registrations and performance data. *Animal Welfare 20*: 239-251

Sutherland MA and Tucker CB 2011 The long and short of it: A review of tail docking in farm animals. *Applied Animal Behaviour Science 135*: 179-191. https://doi.org/10.1016/j.applanim.2011.10.015

Tuyttens FAM, Sprenger M, Van Nuffel A, Maertens W and Van Dongen S 2009 Reliability of categorical versus continuous scoring of welfare indicators: lameness in cows as a case study. *Animal Welfare 18*: 399-405

Verbeke G and Molenberghs G 2009 *Linear Mixed Models for Longitudinal Data*. Springer Science & Business Media: New York, NY, USA

Waghorn GC, Gregory NG, Todd SE and Wesselink R 1999 Dags in sheep: a look at faeces and reasons for dag formation. *Proceedings of the New Zealand Grassland Association 61*: 43-49

Ware JW, Vizard AL and Lean GR 2000 Effects of tail amputation and treatment with an albendazole controlled-release capsule on the health and productivity of prime lambs. *Australian Veterinary Journal 78*: 838-842. https://doi.org/10.1111/j.1751-0813.2000.tb10504.x

Watts JE and Marchant RS 1977 The effects of diarrhoea, tail length, and sex on the incidence of breech strike in modified mulesed Merino sheep. *Australian Veterinary Journal 53*: 118-123. https://doi.org/10.1111/j.1751-0813.1977.tb00132.x

Wemelsfelder F and Mullan S 2014 Applying ethological and health indicators to practical animal welfare assessment. Scientific and technical review. *Office International des Epizooties 33*: 11-20. https://doi.org/10.20506/rst.33.1.2259