agriculture

MDPI

Article

Assessing the Validity of Animal-Based Indicators of Sheep Health and Welfare: Do Observers Agree?

Clare J. Phythian [1,2], Eleni Michalopoulou [1] and Jennifer S. Duncan [1,*]

1 Department of Epidemiology and Population Health, Institute of Infection and Global Health, University of Liverpool, Leahurst, Neston CH64 7TE, UK; clare.phythian@nmbu.no (C.J.P.); eleni.michalopoulou@liverpool.ac.uk (E.M.)

2 Section for Small Ruminant Research, Faculty of Veterinary Medicine, Institute for Production Animal Clinical Science, Norwegian University of Life Sciences, 4325 Sandnes, Norway

* Correspondence: jsduncan@liverpool.ac.uk; Tel.: +44-151-794-6050

Received: 15 March 2019; Accepted: 24 April 2019; Published: 28 April 2019

Abstract: Sixteen animal-based indicators of sheep welfare, previously selected by a stakeholder panel, and based on the Farm Animal Welfare Council (FAWC) Five Freedoms, were assessed in terms of the level of inter-observer agreement achieved during on-farm testing. Eight observers independently tested the 16 indicators on 1158 sheep from 38 farms in England and Wales. Overall inter-observer agreement was evaluated by Fleiss's kappa (κ), and the pair-wise agreement of each observer was compared to a 'test standard' observer (TSO). Inter-observer assessments of the welfare indicators; dental abnormality, cleanliness score (ventral abdomen), mastitis, tail length, skin lesions, body condition scoring and lameness produced 'fair to good' levels of agreement ($0.40 < \kappa < 0.75$) and joint swellings had 'excellent' levels of agreement ($\kappa \geq 0.75$). The very low apparent prevalence (<0.8%) of sheep with specific outcomes such as pruritis, wool loss, myiasis, thin body condition, diffuse or severe skin lesions limited kappa analysis for these indicators. Overall, findings suggest that observers of differing experience, training and occupation were reliable in assessing key animal-based indicators of sheep health and welfare.

Keywords: animal welfare; sheep; observer agreement; on-farm assessment; welfare indicator

1. Introduction

The assessment of farm animal welfare is undertaken for a number of reasons and by different inspection agencies. These include statutory inspections for legislation enforcement and inspections for private farm assurance schemes, which provide consumer assurance of the welfare standards of products of animal origin. Welfare assessment may also be performed by farmers and their veterinary surgeons, as part of routine stock monitoring and to inform animal health and production plans and intervention strategies [1].

Traditionally, welfare assessment systems have relied primarily on assessment of resource provision to the animals (welfare inputs). Resource-based assessments may include the provision of appropriate food and water, bedding materials, the presence of a health plan and the condition of buildings and equipment. However, it is recognised that direct observations and assessments of the animals, be these physical, physiological and/or behavioural outcomes i.e. animal-based indicators of welfare, provide a more accurate estimation of the welfare experience of the animals themselves over management and resource inputs [1]. Consequently, animal-based welfare indicators have been developed and tested for other animal species, but comparatively few animal-based welfare indicators for sheep have been examined [2,3].

Potential animal-based welfare indicators for sheep were identified by a consenus of stakeholder opinion, using the Five Freedom Framework to identify welfare priorities for sheep managed under British farming systems [4]. Given the potential role of such indicators in statutory and private welfare assessment schemes, it is essential for farmers, consumers and for the individual animal, that indicators used for this purpose are thoroughly tested. Animal-based indicators of sheep welfare need to be evaluated to ensure that they are:

(1) valid measures i.e., genuinely measuring an aspect of an animal's welfare state;
(2) reliable, in that different observers consistently score the same animal or group of animals in a similar way (inter-observer agreement);
(3) feasible to apply under a variety of farm management systems [5].

The welfare indicators are considered using diagnostic tests. The test performance of novel diagnostic tests is usually examined against a reference standard or 'gold standard' [6]. Currently there is no gold standard test for animal welfare assessment. In such circumstances, researchers may use a 'pseudo-gold' standard by comparing whether multiple assessors apply the measures in the same way as their trainer or reference observer [7–9]. Kappa agreement analysis is commonly used to determine the level of inter-observer agreement [10]. However, a limitation of this analytical approach is that it cannot identify whether systematic scoring differences occur between pairs or groups of multiple observers. Therefore, a complimentary approach is to assess for observer disagreement or scoring divergence. The aim of this study was to examine the validity of 16 animal-based welfare indicators of sheep in terms of the level of inter-observer reliability, as assessed by kappa agreement and scoring divergence, and their feasibility for on-farm application by observers of differing experience and training.

2. Materials and Methods

2.1. Sheep Population

Thirty eight sheep farms comprised of 16 lowland, 11 upland and 11 hill farms - representative of the types within the British sheep industry [11], and included commercial, pedigree and smallholding flocks [12], were recruited to the Sheep Welfare Project through contact with veterinary practices, National Sheep Association and independent sheep consultants. All flocks were located in England and Wales and were within a 200 km radius of the University of Liverpool, School of Veterinary Science, Leahurst campus. This convenience sample of sheep farms were selected based on their location and willingness to participate.

All study farms gave informed, written consent to participate and were provided with a written study protocol. Each farm was requested to provide a sample of approximately 100 sheep (aged > 1 year) that they considered were representative of the flock. Out of this group, 30 sheep were then randomly selected for inclusion in this inter-observer reliability study using a pre-determined random number system (based on the order that sheep entered the handling race). A sample size of 30 sheep was based on previous recommendations and research on the number of subjects required to provide sufficient scoring variation for analysis of observer agreement [13,14].

2.2. Observer Population

Eight observers were recruited from the University of Liverpool, School of Veterinary Science, to include veterinary (n = 1) and animal science (n = 2) students, and veterinary surgeons (n = 5). For the purposes of this study, observers were classified as 'experienced' in sheep welfare assessment if they had the equivalent of three years or more full-time experience of working on sheep farms, handling animals and were classed as 'active' if they had performed sheep health and welfare assessments (including body condition scoring) in at least the year prior to the study. Those that did not meet these criteria were classed as 'inactive' and/or 'inexperienced', which included assessors holding veterinary

qualifications and veterinary specialists of other farmed animals whom had not been actively working with sheep work in the year prior to the study.

Observer 1, a veterinary surgeon classed as an 'experienced' and 'active' assessor who developed the indicator assessment methods and provided on-farm observer training, was designated the 'test standard observer' (TSO). Each observer was provided with an indicator manual that detailed indicator scoring definitions and assessment protocols. Prior to the commencement of the on-farm study, observers were classified as 'trained' after receiving a one-day on-farm training session on the University sheep farm (Table 1). The training led by the TSO involved a short theoretical session introduction to the scoring descriptors with examples of images of specific indicator scores, and a full day handling, turning and examining live sheep. During this practical training session, observers were encouraged to openly discuss and compare lesion scores and compare with the descriptors in manual guide, and practice use of the scoring sheets. Particular emphasis was paid on practical training in body condition scoring, and discussion of assessor scores, although animals in extreme condition scores (e.g., 1 and 5) were not present in the training flock.

The TSO performed welfare indicator assessments on all study farms and was accompanied by 1–2 observers that were different at each visit. Observers independently assessed each indicator on a sample of sheep from each study farm on the same day and were blinded to the clinical history or production performance of the study farms.

The study was approved by the University of Liverpool Ethics Committee (study reference number RETH000287).

Table 1. Observer identity, occupation, experience, activity and training status, and number of sheep assessed by the test standard observer ("TSO") and other study observers.

Observer ID	Occupation	Experience	Status	On-Farm Training	*n* Sheep Assessed
1	Veterinary surgeon	Experienced	Active	TSO	1158
2	Animal science student	Inexperienced	Inactive	Trained	930
3	Veterinary student	Inexperienced	Inactive	Trained	780
4	Animal science student	Experienced	Active	Trained	270
5	Veterinary surgeon	Inexperienced	Inactive	Trained	90
6	Veterinary surgeon	Experienced	Active	Trained	60
7	Veterinary surgeon	Inexperienced	Inactive	Untrained	60
8	Veterinary surgeon	Experienced	Inactive	Untrained	30

2.3. Welfare Indicators

Sheep were examined in an assessment pen and to reduce any effect of behavioural isolation, two animals were placed in the pen at a time. Pen size and shape varied between farms but was always large enough for the sheep to move around freely (approximate minimum of 2 m^2) and to conduct gait assessment. On entry to the assessment pen, the gait of each sheep was examined (lameness was defined as score ≥2 [14]. Individual sheep were then caught, minimally restrained and examined in the standing position to assess: demeanour, eye abnormality, dental abnormality, breech cleanliness, tail length, wool loss, pruritis, skin lesions, injuries and wounds, myiasis, body condition score (BCS), and fit-fat-thin score (Table 2). The sheep was then turned over to assess abdominal cleanliness, joint swellings, mastitis in ewes and to complete examination of the skin and fleece. Ewes were not turned over during the month following the mating or late pregnancy (4 to 5 months gestation). In this instance, all indicators were examined as fully as possible in the standing animal. Inspection and recording took between 5 to 10 min per sheep. All observations took place on the same day. Each observer attending the farm visit performed independent assessments on the same sheep, consecutively. Observers were not permitted to discuss or disclose their assessments during the study.

Table 2. Description of indicators and scoring scales.

Indicator	Method of Assessment	Scoring Scale
Lameness	As individual sheep entered the assessment area the gait was evaluated for signs of lameness defined as score ≥2 of Kaler et al., scale [14].	0 = Sound 1 = Lame
Demeanour	The demeanour of a sheep was assessed as the sheep approached the assessment area. Signs of dull demeanour included behavioural separation, lowered head carriage, and reduced responsiveness to the approach of the assessor were signs that were used to assess the presence of dull demeanour.	0 = Bright, alert, responsive 1 = Dull, depressive, reduced responsiveness
Eye Abnormality	An eye condition was deemed to be present if any one of the following signs was observed—blepharospasm, corneal opacity, abnormal ocular discharge, lacrimation with tear-staining of skin, conjunctivitis, or entropion.	0 = Absence of eye abnormality 1 = Presence of eye abnormality
Dental Abnormality	Missing permanent incisors were scored as incisor loss. The molar teeth were assessed by palpation along the maxilla and mandible. The thickness, sharpness, length and position of the molar teeth were assessed as well as outward displacement of the molars, palpable mandibular swellings and bony growths.	0 = No evidence of incisor loss or molar disease 1 = Loss of incisors (broken mouth) 2 = Presence of molar tooth abnormality 3 = Incisor loss and molar tooth abnormality
Abdomen Cleanliness	Soiling of the ventral abdomen with mud, faeces or dirt was scored along a 3-point scale.	0 = Clean: may be minor splashing 1 = Discreet to solid plaques 2 = Very heavily contaminated
Breech Cleanliness	Faecal soiling of the rear area was scored along a 3-point scale. The breech area was defined to include the perineum, the superficial and medial aspects of the gluteal region to the top of the hind limbs.	0 = Clean: may be minor splashing with faeces 1 = Discreet to solid plaques of faeces 2 = Very heavily contaminated with faeces
Mastitis	Mammary glands were palpated for areas of focal or diffuse thickening, swelling, heat, pain or discomfort.	0 = No evidence of mastitis in any gland 1 = One gland affected by mastitis 2 = Both glands affected by mastitis
Tail Length	The length of the tail was assessed according to the Welfare of Farmed Animals (England) regulations 2007 [15]	0 = Appropriate tail length: tail covers anus in males or vulva in females 1 = Inappropriate tail length: tail does not cover anus in males or vulva in females
Wool Loss	Assessment of fleece cover was made in the shorn and unshorn sheep and an area of absence of wool was recorded.	0 = No wool loss observed 1 = Area of wool loss observed

Table 2. *Cont.*

Indicator	Method of Assessment	Scoring Scale
Pruritis	Signs of skin irritation were assessed using the nibble test [16]. This was performed by rubbing the fingertips on the skin of the sheep along the lumbar, flank and shoulder regions. A positive response and presence of skin irritation was interpreted as positive if the animal showed head and neck extension, and nibbling and chewing behaviours associated with head and tongue movements after manual stimulation.	0 = No response to nibble test 1 = Positive response to nibble test: extension of head and neck, mouthing, or nibbling
Skin Lesion	The skin was assessed in both the standing and turned sheep for lesions including abscesses, seborrhoea, moisture, erythema, discolouration, abnormal odour. In the fully fleeced sheep, the observer ran their hands through the wool and areas of the wool were parted in order to examine the integrity of skin.	0 = No skin lesions observed 1 = Presence of a single lesion ≤3 cm diameter 2 = Presence of multiple lesions >3 cm diameter 3 = Presence of a single lesion (≤10 × 5 cm) 4 = Presence of multiple lesions (≥10 × 5 cm) 5 = Diffuse, generalised skin lesion(s)
Injuries and Wounds	Injuries and wounds were simultaneously assessed along with skin lesion and wool loss indicators. The skin of the entire body and head was examined for signs of injury such as wounds, bruises, cut and scratches.	0 = No injuries, wounds or scratches observed 1 = Superficial scratches or superficial cuts ≤5 cm 2 = Superficial scratches or superficial cuts ≥5 cm long 3 = Healing or healed wounds 4 = Single open wound on the body or head 5 = Multiple open wounds on body or head
Myiasis	The presence of maggots anywhere on the sheep was recorded.	0 = No maggots observed on the sheep 1 = Presence of any live maggot(s)
Body Condition Score (BCS)	Body condition was assessed using the Russel [17] six-point scoring scale. Briefly, using both hands the lumbar vertebrae and transverse processes were manually palpated. An assessment of the sharpness and prominence of the spinal process together with coverage over the *longissimus dorsi* (loin) and degree of fat cover was made by pressing the fingertips underneath the ends of the lumbar processes to assess the amount of muscle.	0 to 5 scale of Russel [17]
Fit – Fat – Thin	The assessment of body condition using the fit – fat – thin system was based on the Russel [17] scoring descriptions.	Fit = Body condition score ≥2–4 Fat = Body condition score >4 Thin = Body condition score <2 (as above)
Joint Swelling	The presence of swelling, heat and pain assessed visually and palpation of the elbow, stifle, carpal, tarsal, metacarpal and metatarsal joints.	0 = No visible or palpable joint swelling 1 = Visible joint swelling

2.4. Evaluation of Diagnostic Performance

Data were analysed in Stata version 13 (StataCorp LP, College Station, TX, USA). A total of 1158 sheep were assessed, including 1027 ewes and 131 rams. There were five cases in which the udder was not fully palpated and recorded as a missing observation. No other missing data was recorded. The number of sheep observed by the TSO with each animal-based indicator was divided by the total number of sheep assessed to determine the apparent prevalence (percentage and 95% confidence interval, 95% CI) for each indicator outcome. The overall level of agreement for multiple observer assessments (every instance a sheep was examined by ≥2 assessors) was evaluated by Fleiss's kappa (κ), which provides a coefficient value based on the average observed proportion of agreement for cases where three or more observers assessed the same sheep [10]. With the exception of BCS, paired agreement between the TSO and each observer for each indicator was examined using unweighted Cohen's κ [10]. The categorical scoring indicator of mastitis was analysed as a binary indicator (0 = absence, 1 = presence of mastitis and/or palpable lesions in any gland). There were few observations of extreme BCS (Table 2) and the scores of "fit" and "fat" from the fit-fat-thin condition scoring scale were examined as binary indicators. Since BCS is scored along an ordinal scale and scoring categories are assumed to be equi-distant [14], a quadratic weighted κ [10] was used to assess observer agreement using the kap STATA command. All κ results were interpreted whereby values ≥0.75 suggested 'excellent', κ 0.40–0.74 'fair to good', and κ ≤ 0.39 indicated 'poor' agreement [10]. Graphical distributions of the scoring differences between the paired assessments of the TSO and each observer (TSO score–observer score) were visually examined for evidence of systematic scoring differences i.e. divergence with the TSO.

3. Results

During July to December 2008, 16 animal-based indicators were tested on 1158 sheep from 38 farms across England and Wales by a varying group of 2–3 observers. The proportion of the study population observed by the TSO with each indicator is shown in Table 3. For indicators where the apparent prevalence was less than 1%; namely dull demeanour, eye abnormality, pruritis, myiasis, wool loss, diffuse skin lesions and wounds, and thin category of the fit-fat-thin scale (BCS < 2) (Table 3), kappa values were not determined. The remaining ten indicators were examined in terms of Fleiss's κ agreement and paired κ agreement (Table 4).

3.1. Lameness

Just under 7% of sheep in this study were recorded as lame (Table 3). Overall, fair to good inter-observer agreement was achieved (Table 4).

3.2. Dental Abnormalities

The apparent prevalence of dental abnormalities, as determined by the TSO, was 3.97% (Table 3). Overall inter-observer κ agreement of 0.50 (95% CI 0.48–0.52) was classed as fair to good agreement (Table 4).

3.3. Cleanliness Scoring (Abdomen and Breech)

Overall inter-observer agreement for the cleanliness scoring of the abdomen was fair to good (Table 3). In contrast, breech cleanliness produced poor agreement (Table 4).

3.4. Mastitis

Under 6% of ewes were identified with palpable intra-mammary masses (Table 4). Fair-good agreement (Table 4) was found for all observers.

3.5. Tail Length

Excellent levels of inter-observer agreement (Table 4) on tail length were found.

Table 3. Apparent prevalence of sheep health and welfare indicator outcomes (percentage, 95% confidence interval) in descending order as determined by the test standard observer (TSO).

Indicator	*n* Observed	Percentage (%) Observed (95% CI)
'Fit' body condition (scores 2–4)	1088	93.96 (92.58–95.32)
Body condition score 3	601	51.90 (49.02–54.78)
Body condition score 2	335	28.93 (26.31–31.54)
Breech cleanliness scores 1	261	22.54 (20.13–24.94)
Abdomen cleanliness score 1	203	17.53 (15.33–19.72)
Body condition score 4	151	13.04 (11.10–14.89)
Molar abnormality	132	11.40 (9.57–13.23)
Lameness	80	6.91 (5.44–8.37)
'Fat' body condition (>4)	64	5.52 (4.21–6.84)
Mastitis [#] (one or both glands)	57	5.58 (4.17–6.99)
Short tail length	49	4.23 (3.07–5.39)
Incisor tooth loss	46	3.97 (2.84–5.10)
Dull demeanour	9	0.78 (0.27–1.28)
Eye abnormality	7	0.60 (0.16–1.05)
Skin lesions score 1 and 2	32	2.76 (1.81–3.71)
Breech cleanliness score 3	24	2.07 (1.25–2.89)
Body condition score 5	64	5.53 (4.21–6.84)
Injuries/wounds score 3	17	1.47 (0.77–2.16)
Joint swelling	12	1.03 (0.45–1.62)
Pruritis	9	0.78 (0.27–1.28)
Skin lesions score 3 and 4	7	0.60 (0.16–1.05)
Injuries/ wounds score 4 and 5	7	0.60 (0.16–1.05)
Body condition score 1	7	0.60 (0.16–1.05)
Skin lesion score 5	6	0.52 (0.10–0.93)
'Thin' body condition (>2)	6	0.52 (0.10–0.93)
Myiasis	3	0.26 (0–0.55)
Abdomen cleanliness score 2	2	0.17 (0–0.41)
Injuries/wounds scores 1 and 2	1	0.09 (0–0.26)

[#] Indicator assessed in ewes (n = 1027).

3.6. Skin Lesions

Overall between-observer agreement was fair to good (Table 4). Less than 3% of sheep were identified with single or multiple lesions <3 cm in diameter (Table 3).

3.7. Injuries and Wounds

For injuries and wounds the overall agreement was categorised as 'poor'. The apparent prevalence of injuries and wounds was very low and, only the 'healing wound' category (score 3) observed at just over 1% by the TSO (Table 3).

3.8. Body Condition Scoring and Fit-Fat-Thin

There was little scoring variation—most sheep were within scores 2–3 and very few thin sheep were found (BCS <2) (Table 3). Inter-observer agreement of the Russel condition scoring system [17] was fair to good (Table 4). When BCS categories were grouped into Fit-Fat-Thin scores, the level of observer agreement improved (Table 4). Graphical evaluation suggested that only slight scoring differences arose between the TSO and most observers, although observer 6 consistently scored sheep one BCS lower than the TSO. Most disagreements occurred between the mid-range of condition scores i.e., between BCS 2 and 3.

3.9. Joint Swellings

Overall, excellent levels of agreement were identified for joint swelling (Table 4).

Table 4. Overall inter-observer kappa (κ) agreement for multiple observer assessments ($n > 2$) and paired κ agreement between the test standard observer ('TSO': observer 1) and observers 2–8.

Indicator	Overall Agreement (κ)	Interpretation of κ	Paired Agreement (κ) by Observer Identity						
			2	3	4	5	6	7	8
Joint swellings	0.77	Excellent	0.45	0.73	0.85	[a]	1	[a]	[a]
Tail length	0.71	Fair-good	0.72	0.80	1	0.79	[a]	[a]	[a]
Lameness	0.67	Fair-good	0.66	0.69	0.77	0.41	1	[a]	[a]
'Fit-fat-thin' body condition	0.67	Fair-good	0.80	0.69	[a]	0.80	[a]	0.39	[a]
Abdomen cleanliness	0.62	Fair-good	0.60	0.53	0.73	0.14	0.54	0.97	[a]
Dental abnormalities	0.50	Fair-good	0.50	0.51	0.64	0.31	0.50	0.44	0.65
Body condition score [W]	0.46	Fair-good	0.55	0.63	0.61	0.62	0.35	0.42	0.06
Mastitis	0.44	Fair-good	0.30	0.61	0.83	0.48	[a]	0.41	[a]
Skin lesions	0.42	Fair-good	0.37	0.59	0.50	0.13	0.88	[a]	[a]
Injuries and wounds	0.38	Poor	0.21	0.39	0.66	0.33	1	[a]	[a]

[a] Insufficient scoring variation to calculate kappa coefficients (all indicator scores were 0). [W] Weighted kappa was used to evaluate the paired inter-observer agreement for body condition scores. Kappa interpretation: $\kappa \geq 0.75$ 'excellent', κ 0.40–0.74 'fair to good', and $\kappa \leq 0.39$ 'poor' agreement [10].

4. Discussion

For the purposes of an on-farm welfare assessment, any animal-based indicator of sheep health and welfare needs to be consistently and accurately applied. In this study 16 animal-based indicators, used as proxy measures of sheep welfare, were analysed using the principles of diagnostic test evaluation to assess the level of overall and paired between-observer agreement.

4.1. Impact of Study Population on Kappa Analysis

The indicators were examined on 1158 sheep of different breeds, from 38 farms in a range of geographical locations and managed under a variety of systems. Background details on farm management type were reported following guidelines on the reporting of the design and conduct of reliability and agreement studies [18]. The purpose of this study was to examine indicator validity in terms of the level of observer agreement or 'reliability'. Clearly, these indicators also need to be feasible for use under a range of farming conditions. Experiences from this study and the wider Sheep Welfare Project will inform decisions on the selection of valid, reliable and feasible indicators for use in on-farm welfare assessment protocols. As well as being reliable [5], the internal and external validity of these indicators also needs to be examined in terms of their ability to detect differences in welfare outcomes across different farms and management types, and across the different seasons and periods of the production cycle. This will be the next objective following the identification of reliable indicators.

Sampling bias was reduced by using a random number identifier to pre-select 30 sample sheep. However, with the non-random selection of farms, and the unavoidable voluntary participation of farmers, it is likely that the study population may be biased towards farms with higher welfare status. Statutory animal welfare agencies and actors were approached for support in active recruitment of sheep farms undergoing compulsory welfare inspection processes or follow-up in order to include farms with more extreme indicator scores. Despite these efforts, no flocks of this type were recruited by this approach. The inclusion and on-farm application of these indicators by government and or farm assurance bodies could provide a more representative and varied sheep population and may help further validation and reliability testing of these indicators.

Indeed, the very low apparent prevalence of conditions such as dull demeanour, eye abnormality, pruritis, myiasis, and wool loss, meant that it was not possible to fully examine kappa agreement for all indicators. Observer agreement studies should ideally include all scoring levels of each indicator [19]. Since κ takes account of the fact that chance agreement is high when conditions occur at a very low or very high prevalence [8]. Where specific scores are not identified, the observed indicator prevalence should be taken into account as this affects the 2×2 table used in kappa analyses. A low prevalence of disease (conversely a high level of score 0's) can result in artificially low levels of κ agreement [20]. A minimum sample size of thirty subjects has been previously suggested for observer agreement studies [13]. However, one of the issues for this study was that this sample size was insufficient for assessing the performance of several indicators. Our experiences suggest that much larger sample sizes than 30 sheep and the inclusion of farms with greater variation in indicator outcomes may be required to obtain sufficient scoring variation and 'extreme' scores required to facilitate agreement analyses. However, experiences from this study reveal that gaining access to farms with a higher prevalence of welfare issues appears to remain a particular challenge for animal health and welfare researchers.

4.2. Welfare Indicator Test Performance

Overall, results identified excellent kappa levels ($\kappa > 0.75$) for joint swelling and fair to good agreement ($0.40 < \kappa < 0.75$) for the assessment of tooth loss/abnormalities, cleanliness scores (abdomen), mastitis, tail length, skin lesions, body condition scoring and lameness. Detection of disease by observation and palpation, akin to a clinical examination, is not perfect and the TSO was not considered a 'gold standard' but a reference for evaluating observer agreement [8]. In some cases, divergence from a TSO could represent a closer approximation to the true positive or true negative health and

welfare state. Multiple assessments by different observers may establish a consensus that may or may not be close to true diagnostic status. For example, all observers could misdiagnose a condition and agree about this, which would result in high kappa agreement. This remains a familiar and, as yet, unresolved issue for studies based on detecting specific physical signs. However, significant deviation from the trainer or wider observer group remains a cause for concern for observational studies [19]. Previous studies have identified that training can be important in attaining high levels of observer agreement [7,9]. Whilst training is considered key to reducing any observer effects, slight individual differences can also occur in spite of training due to the effects of personal bias and prior experience [9].

Due to limitations in resources and time available for multiple farm visits, different groupings of 2 to 3 observers performed the 38 study farm assessments. Only two observers (7 and 8) were classed as 'untrained' in that they received the indicator scoring definitions but did not attend an on-farm practical workshop and the full day "hands-on" and practical on-farm training session led by the TSO. Observers 7 and 8 examined 60 and 30 individual sheep respectively, but we found there was insufficient scoring variation for analysis of observer agreement for several indicators, including joint swellings. For other indicators it was considered that these observers were responsible for some outlier results, for example the lower kappa for condition scoring by observer 8 and fit-fat-thin scoring and tail length assessments by observer 7. For example, observer 8, classed as an 'experienced' veterinary assessor but was 'inactive' and 'untrained' in indicator assessments, consistently scored animals on a single farm one condition score unit higher than the TSO, which resulted in poor paired agreement. As this observer only performed a single farm visit, it was difficult to fully elucidate the effect of observer training and the influence of prior experience here.

Evidence of systematic scoring bias, such as disagreement with other observers and the TSO, might suggest that prior assessor experience or lack of standardisation or calibration with other 'trained' observers influenced the condition scoring assessments by this particular observer. The prior observer experience or 'bias' of this assessor with sheep managed in a lower BCS may also explain scoring divergence that occurred between the TSO and other observers present at the single farm visit. This finding highlights that being 'experienced' alone cannot predict high levels of agreement or indeed compliance to a particular indicator scoring system. Whilst this study does not provide strong evidence of the impact of training, this particular example supports our suggestion that practical training of assessors, particularly training with body condition scoring of sheep across different farms and breeds, is an important feature for becoming calibrated to the scoring system and providing consistency of on-farm animal welfare assessments across different assessors. A process of on-farm training and reliability testing could be used to select the most suitable observers and those with high reliability in order to maintain the validity, confidence and transparency in indicators assessments. An earlier study on the body condition scoring of dairy cows found that training resulted in improved kappa agreement [9]. Pilot work in sheep suggested that a short assessor recalibration exercise might improve observer BCS agreement [21] but this was only examined over a limited time period. More attention on observer training and alternative training methods for sheep BCS may also facilitate higher levels of observer agreement [22].

As well as issues with between-study variation in disease prevalence [8] arbitrary interpretation tools can limit cross-study comparisons of kappa results [18]. There are well-known issues and limitations with kappa analysis and interpretation [20]. Therefore, for this study, a more stringent interpretation of kappa [10] was selected over that of Landis and Koch [23]. However, use of the latter would have resulted in more indicators being interpreted with substantially higher levels of agreement. With the exception of breech cleanliness and injuries and wounds, all indicators were interpreted as "clinically useful" (i.e., $\kappa \geq 0.4$) [24].

Lameness is a key welfare issue for sheep [25] and whilst high levels of inter- and intra-observer agreement have been reported for a six-point categorical lameness scoring scale assessing video footage of individual sheep [14], similar levels of agreement had not been reported under field conditions [26]. Following field pilot testing (C.J Phythian, unpublished observations), a simple binary gait scoring

system, based on Kaler and co-authors scale [14] was used in which sheep were classed as either 'lame' (scores ≥ 2) or 'sound'.

Gait assessment relied on the ability to quietly walk individual sheep around the assessment pen, which could be challenging when very responsive or 'flighty' sheep were examined. In spite of differences in assessment conditions between farms, such as the quality of flooring, and availability of lighting, results suggested that gait assessment of individual sheep produced 'fair to good' agreement. A comparison of group-based and individual animal gait assessment approaches for sheep are reported elsewhere by the authors' [27]. We found that on the majority of farms assessed, the most accurate approach for assessing lameness was by group assessment in the field. However, depending on the circumstances, field conditions and sample size may preclude this and individual gait assessment may provide a more feasible alternative.

Whilst standardisation of the conditions for gait assessment of individual sheep was attempted, in practice this was found to be challenging on some farms [27]. A portable lighting source, in addition to taking along gates to set-up an assessment pen of standard shape and size, and a portable, level flooring surface could be useful equipment for improving the standardisation of gait assessments of individual sheep. For the present study, a binary gait scoring system was deemed most appropriate for identifying a potential or actual risk for sheep welfare. Certainly, we have not resolved the ethical dilemma and practical challenges of weighing up and interpreting lameness severity, duration or prevalence as important criteria for sheep welfare [28]. Even ovine foot lesions associated with severe pathology and pain are not always associated with an observed high prevalence of lameness [27]. Furthermore, a single visit cannot capture data on lameness duration. Our simplistic approach to classifying sheep simply as sound or lame (≥score 2 [14]), requires further prompt investigation and application of a second-tier of indicators, including foot examination [29], in order to diagnose the specific reasons for lameness in individuals and/or groups of sheep. More recent video clip studies have reported high inter- and intra-observer agreement when farmers and veterinary surgeons applied a four-point sheep lameness scale [30]. With further assessor training and calibration, more detailed sheep gait scoring (i.e., sound, mild, moderate and severe) systems could be applied where facilities and farm management systems allow.

Due to the low prevalence of skin lesion scores, kappa was only obtained for single or multiple skin lesions <3 cm in diameter. Other researchers found poor observer agreement for the assessment of skin lesions in sheep and highlighted issues due to the masking of skin lesions in fleeced animals. This led to the suggestion to only score shorn sheep and skin lesions over 2 cm in size [2]. Whilst this approach may improve the level of observer agreement it risks missing skin lesions of great welfare importance, such as sheep scab. Similarly, whilst pruritis might be detected by observation of areas of wool loss and signs of intense rubbing against fences and objects, it is possible that animals with pruritic skin conditions may not always be identified during a relatively brief period of undisturbed observation. Therefore, it is recommended that a combination of group-assessment and individual, physical examination of a selection of sheep, particularly those with signs of wool loss, to assess skin, and fleece condition, and the 'nibble test' [16] to avoid missing small skin lesions that may be of high welfare importance. This concurs with more recent findings which identified that fleece condition and skin lesions were reliable indicators for Australian sheep [22] and possibly reflects the higher prevalence of these lesions, which facilitated agreement analysis.

In common with earlier work [3,21], most of the animals in this study were found to be 'fit' with a BCS of 2–3. The majority of scoring disagreements occurred over this mid-range of condition (BCS 2–3). Whilst half and quarter-unit scoring precision is recommended for management purposes [17], full unit scores were selected based on earlier findings [21] and the aim of identifying animals in extreme scores. It is possible that higher agreement could be achieved if the scale were applied to a population with more varied condition scores, and more recent research using quarter-point scores found evidence of moderate inter-observer and poor intra-observer reliability of some assessors [22], again highlighting the importance for calibration and practical training. Whilst the fit-fat-thin scale may lack sufficient

precision for clinical and management applications [31], it may be sufficiently sensitive for welfare assessment protocols focused on identifying sheep below the minimum standards.

4.3. Feasibility

All the indicators tested here were found to be feasible for use across the study farms. The physical examination, scoring and recording per observer using all 16 indicators took approximately five minutes per sheep, and required no input from the farmer during indicator assessments. Use of these indicators does require sheep to be gathered and individually handled, and at certain times of the sheep production cycle, such as late pregnancy, it may not be appropriate to turn ewes. Therefore, some indicators may take longer to assess and others, for example: mastitis, ventral abdominal skin lesions and cleanliness, were examined in the standing animal. As well as removing those indicators with poorer levels of test performance, the time required for assessments could be reduced through targeted selection of key welfare indicators depending on the focus and priorities of the specific assessment procedure. Recording time may also be reduced by use of electronic recording systems.

4.4. Future Implementation of Sheep Welfare Indicator Assessments

The indicators and scoring systems tested in this study were considered to be a first step in the development of an overall on-farm welfare assessment protocol for sheep, for use in farm assurance schemes and animal welfare certification frameworks, or for verifying compliance with minimal statutory regulations. They do not offer a complete on-farm welfare assessment tool and other approaches including assessment of positive welfare states of sheep using qualitative behaviour assessment (QBA) [32] or stockperson-animal handling and behaviour [33] offer complementary tools for assessing positive states and elements of animal welfare beyond measures of physical health and welfare. Since this study was completed other conceptual frameworks, such as Welfare Quality®, have been examined for identifying welfare concerns for sheep [5]. Regardless of the framework or approach taken, a common theme of on-farm assessment protocols is the application of indicators of physical health and welfare of sheep, which have been tested across a range of countries and farm management systems including Italy [2], Norway [3], England and Wales [12] Scotland [34], and Australia [22]. Many of the animal-based indicators tested in these earlier studies appear in the final AWIN sheep welfare assessment protocol [35]. Our study provides evidence of the validity of many of the included animal-based indicators in terms of the reliability of these indicators. In the future, these indicators might be selected and applied as part of farm assurance schemes, assessment with legal compliance, or benchmarking of individual farms to inform improvements in management or inputs [1,29], as well as providing positive feedback to producers on the physical health and welfare of individual and groups of sheep.

5. Conclusions

Animal-based indicators of sheep welfare with good inter-observer agreement and related to key welfare concerns for sheep included lameness, joint swellings, body condition and tail length. These indicators may be applied by producers, veterinary surgeons, farm assurance and certification assessors, or farm animal welfare inspectors as robust and feasible tools in on-farm assessments. With further training and wider testing on a sheep population with greater scoring variation, it is likely that higher levels of observer reliability could be achieved. As well as being feasible on-farm measures, these indicators also appear feasible to assess sheep welfare at transport and markets [36]. Overall, they may inform flock management practices or identify areas where further investigation and/or additional interventions or inputs are required. Further work examining the effect of farm management type and seasonal influences on physical outcomes of sheep health and welfare could provide further evidence of the validity of these animal-based welfare indicators.

Author Contributions: Conceptualization, J.S.D., E.M.; Methodology, C.J.P, E.M., J.S.D.; Software, C.J.P.; Validation, C.J.P., E.M. and J.S.D.; Formal Analysis, C.J.P.; Investigation, C.J.P.; Resources, J.S.D., E.M.; Data Curation, C.J.P.; Writing-Original Draft Preparation, C.J.P.; Writing-Review & Editing, E.M., J.S.D.; Supervision, E.D., J.S.D.; Project Administration, J.S.D.; Funding Acquisition, J.S.D.

Funding: The study was funded by the Department of the Environment and Rural Affairs (Defra) as part of the AW1025 grant—'Development of indicators for the on-farm assessment of sheep welfare'.

Acknowledgments: The authors also gratefully acknowledge the support and participation of the project expert panel, study farms and the on-farm assistance of Daniel Hughes, Rachel Wyllie, Stephen Beer, Dai Grove-White and Phil Jones. We are particularly grateful to Peter Cripps for statistical support.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Main, D.C.J.; Kent, J.P.; Wemelsfelder, F.; Ofner, E.; Tuyttens, F.A.M. Applications for methods of on-farm welfare assessment. *Anim. Welf.* **2003**, *12*, 523–528.
2. Napolitano, F.; De Rosa, G.; Ferrante, V.; Grasso, F.; Braghieri, A. Monitoring the welfare of sheep in organic and conventional farms using an ANI 35 L derived method. *Small Rumin. Res.* **2009**, *83*, 49–57. [CrossRef]
3. Stubsjøen, S.M.; Hektoen, L.; Valle, P.S.; Janczak, A.M.; Zanella, A.J. Assessment of sheep welfare using on-farm registrations and performance data. *Anim. Welf.* **2011**, *20*, 239–251.
4. Phythian, C.J.; Michalopoulou, E.; Jones, P.H.; Winter, A.C.; Clarkson, M.J.; Stubbings, L.A.; Grove-White, D.; Cripps, P.J.; Duncan, J.S. Validating indicators of sheep welfare through a consensus of expert opinion. *Animal* **2011**, *5*, 943–952. [CrossRef]
5. Richmond, S.; Wemelsfelder, F.; De Heredia, I.B.; Ruiz, R.; Canali, E.; Dwyer, C.M. Evaluation of animal-based indicators to be used in a welfare assessment protocol for sheep. *Front. Vet. Sci.* **2017**, *4*, 201. [CrossRef] [PubMed]
6. Greiner, M.; Gardner, I.A. Epidemiologic issues in the validation of veterinary diagnostic tests. *Prev. Vet. Med.* **2000**, *45*, 3–22. [CrossRef]
7. Kristensen, E.; Dueholm, L.; Vink, D.; Andersen, J.E.; Jakobsen, E.B.; Illum-Nielsen, S.; Petersen, F.A.; Enevoldsen, C. Within and across-person uniformity of body condition scoring in Danish holstein cattle. *J. Dairy Sci.* **2006**, *89*, 3721–3728. [CrossRef]
8. Burn, C.C.; Pritchard, J.C.; Whay, H.R. Observer reliability for working equine welfare assessment: Problems with high prevalences of certain results. *Anim. Welf.* **2009**, *18*, 177–187.
9. Vasseur, E.; Gibbons, J.; Rushen, J.; De Passillé, A.M. Development and implementation of a training program to ensure high repeatability of body condition scoring of dairy cows. *J. Dairy Sci.* **2013**, *96*, 4725–4737. [CrossRef] [PubMed]
10. Fleiss, J.L.; Levin, B.; Paik, M.C. The Measurement of Interrater Agreement. In *Statistical Methods for Rates and Proportions*, 3rd ed.; John Wiley & Sons Inc.: Hoboken, NJ, USA, 2003.
11. Agriculture and Horticulture Development Board. *The Breeding Structure of the British Sheep Industry*; EBLEX Publications: Warwickshire, UK, 2014.
12. Phythian, C.J.; Cripps, P.J.; Michalopoulou, E.; Jones, P.H.; Grove-White, D.; Clarkson, M.J.; Winter, A.C.; Stubbings, L.A.; Duncan, J.S. Reliability of indicators of sheep welfare assessed by a group observation method. *Vet. J.* **2012**, *193*, 257–263. [CrossRef]
13. Walter, S.D.; Eliasziw, M.; Donner, A. Sample size and optimal designs for reliability studies. *Stat. Med.* **1998**, *17*, 101–110.
14. Kaler, J.; Wassink, G.J.; Green, L.E. The inter and intra-observer reliability of a locomotion scoring scale for sheep. *Vet. J.* **2009**, *180*, 189–194. [CrossRef]
15. Welfare of Farmed Animals (England) Regulations (SI 2007 No. 2018). 2007. Available online: https://www.legislation.gov.uk/uksi/2007/2078/contents/made (accessed on 19 February 2019).
16. D'Angelo, A.; Maurella, C.; Bona, C.; Borrelli, A.; Caramelli, M.; Careddu, M.E.; Jaggy, A.; Ru, G. Assessment of clinical criteria to diagnose scrapie in Italy. *Vet. J.* **2007**, *174*, 106–112. [CrossRef]
17. Russel, A. Body condition scoring of sheep. *Practice* **1984**, *6*, 91–93. [CrossRef] [PubMed]
18. Kottner, J.; Audigé, L.; Brorson, S.; Donner, A.; Gajeweski, B.J.; Hróbjartsson, A.; Robersts, C.; Shoukri, M.; Streiner, D.L. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *J. Clin. Epidemiol.* **2011**, *64*, 96–106. [CrossRef] [PubMed]

19. Ruddat, I.; Scholz, B.; Bergmann, S.; Buehring, A.-L.; Fischer, S.; Manton, A.; Prengel, D.; Rauch, E.; Steiner, S.; Wiedmann, S.; et al. Statistical tools to improve assessing agreement between several observers. *Animal* **2014**, *8*, 643–649. [CrossRef] [PubMed]
20. Feinstein, A.R.; Cicchetti, D.V. High agreement but low kappa. 1. The problems of 2 paradoxes. *J. Clin. Epidemiol.* **1990**, *43*, 543–549. [CrossRef]
21. Phythian, C.J.; Hughes, D.; Michalopoulou, E.; Duncan, J.S. Reliability of body condition scoring of sheep for cross-farm assessments. *Small Rumin. Res.* **2012**, *104*, 156–162. [CrossRef]
22. Munoz, C.; Campbell, A.; Hemsworth, A.; Doyle, R. Animal-Based Measures to Assess the Welfare of Extensively Managed Ewes. *Animals* **2018**, *8*, 2. [CrossRef] [PubMed]
23. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. [CrossRef]
24. Sim, J.; Wright, C.C. The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physi. Ther.* **2005**, *85*, 257–268.
25. Farm Animal Welfare Committee. *Opinion on Lameness in Sheep*; Farm Animal Welfare Comittee Publications: London, UK, 2011.
26. Welsh, E.M.; Gettinby, G.; Nolan, A.M. Comparison of a visual analogue scale and a numerical rating scale for assessment of lameness, using sheep as a model. *Am. J. Vet. Res.* **1993**, *54*, 976–983.
27. Phythian, C.J.; Cripps, P.J.; Michalopoulou, E.; Jones, P.H.; Grove-White, D.; Duncan, J.S. Observing lame sheep: Evaluating test agreement between group-level and individual animal methods of gait assessment. *Anim. Welf.* **2013**, *22*, 417–422. [CrossRef]
28. Goddard, P. Welfare assessment in sheep. *Practice* **2011**, *33*, 508–516. [CrossRef]
29. Phythian, C.J.; Cripps, P.J.; Grove-White, D.; Michalopoulou, E.; Duncan, J.S. Inter-observer agreement for clinical examinations of foot lesions of sheep. *Vet. J.* **2016**, *216*, 189–195. [CrossRef]
30. Angell, J.W.; Cripps, P.J.; Grove-White, D.H.; Duncan, J.S. A practical tool for locomotion scoring in sheep: Reliability when used by veterinary surgeons and sheep farmers. *Vet. Rec.* **2015**, *176*, 521–523. [CrossRef]
31. Lovatt, F.M. Clinical examination of sheep. *Small Rumin. Res.* **2010**, *92*, 72–77. [CrossRef]
32. Phythian, C.J.; Michalopoulou, E.; Cripps, P.J.; Duncan, J.S.; Wemelsfelder, F. On-farm qualitative behaviour assessment in sheep: Repeated measurements across time, and association with physical indicators of flock health and welfare. *Appl. Anim. Behav. Sci.* **2016**, *175*, 23–31. [CrossRef]
33. Coleman, G.J.; Rice, M.; Hemsworth, P.H. Human-animal relationships at sheep and cattle abattoirs. *Anim. Welf.* **2012**, *21*, 15–21. [CrossRef]
34. Morgan-Davies, C.; Waterhouse, A.; Pollock, M.L.; Milner, J.M. Body condition score as an indicator of ewe survival under extensive conditions. *Anim. Welf.* **2008**, *17*, 71–77.
35. AWIN. AWIN Welfare Assessment Protocol for Sheep. 2015. Available online: https://air.unimi.it/retrieve/handle/2434/269114/384851/AWINProtocolSheep.pdf (accessed on 19 February 2019). [CrossRef]
36. Llonch, P.; King, E.M.; Clarke, K.A.; Downes, J.M.; Green, L.E. 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. *Vet. J.* **2015**, *206*, 289–297. [CrossRef] [PubMed]