Animal (2013), **7:7**, pp 1182–1190 © The Animal Consortium 2013
doi:10.1017/S1751731113000487

animal

Inter-observer agreement, diagnostic sensitivity and specificity of animal-based indicators of young lamb welfare

C. J. Phythian[1†], N. Toft[2], P. J. Cripps[1], E. Michalopoulou[1], A. C. Winter[3], P. H. Jones[1], D. Grove-White[1] and J. S. Duncan[1]

[1]*Department of Epidemiology and Population Health, Institute of Global Health and Infection, University of Liverpool, Leahurst, Neston CH64 7TE, UK;* [2]*Department of Large Animal Sciences, Faculty of Life Sciences, University of Copenhagen, Grønnegaardsvej 8, DK-1870 Frederiksberg C, Denmark;* [3]*School of Veterinary Science, University of Liverpool, Leahurst, Neston CH64 7TE, UK*

(Received 12 August 2012; Accepted 28 January 2013; First published online 8 April 2013)

A scientific literature review and consensus of expert opinion used the welfare definitions provided by the Farm Animal Welfare Council (FAWC) Five Freedoms as the framework for selecting a set of animal-based indicators that were sensitive to the current on-farm welfare issues of young lambs (aged ⩽6 weeks). Ten animal-based indicators assessed by observation – demeanour, response to stimulation, shivering, standing ability, posture, abdominal fill, body condition, lameness, eye condition and salivation were tested as part of the objective of developing valid, reliable and feasible animal-based measures of lamb welfare The indicators were independently tested on 966 young lambs from 17 sheep flocks across Northwest England and Wales during December 2008 to April 2009 by four trained observers. Inter-observer reliability was assessed using Fleiss's kappa (κ), and the pair-wise agreement with an experienced, observer designated as the 'test standard observer' (TSO) was examined using Cohen's κ. Latent class analysis (LCA) estimated the sensitivity (Se) and specificity (Sp) of each observer without assuming a gold standard and predicted the Se and Sp of randomly selected observers who may apply the indicators in the future. Overall, good levels of inter-observer reliability, and high levels of Sp were identified for demeanour (κ = 0.54, Se ⩾ 0.70, Sp ⩾ 0.98), stimulation (κ = 0.57, Se = 0.30 to 0.77, Sp ⩾ 0.98), shivering (κ = 0.55, Se = 0.37 to 0.85, Sp ⩾ 0.99), standing ability (0.54, Se ⩾ 0.80, Sp ⩾ 0.99), posture (κ = 0.45, Se ⩾ 0.56, Sp = 0.99), abdominal fill (κ = 0.44, Se = 0.39 to 0.98, Sp = 0.99), body condition (κ = 0.72, Se ⩾ 0.38 to 0.90, Sp = 0.99), lameness (κ = 0.68, Se > 0.73, Sp = 1.00), and eye condition (κ = 0.72, Se ⩾ 0.86, Sp = 0.99). LCA predicted that randomly selected observers had Se > 0.77 (acceptable), and Sp ⩾ 0.98 (high) for assessments of demeanour, lameness, abdominal fill posture, body condition and eye condition. The diagnostic performance of some indicators was influenced by the composition of the study population, and it would be useful to test the indicators on lambs with a greater level of outcomes associated with poor welfare. The findings presented in this paper could be applied in the selection of valid, reliable and feasible indicators used for the purposes of on-farm assessments of lamb welfare.

Keywords: animal welfare, indicator, sheep, observer agreement, sensitivity, specificity

Implications

A range of animal-based indicators of young lamb welfare have been developed and tested in terms of their reliability, sensitivity, specificity and feasibility for on-farm use. Outcome-based measures of young lamb welfare may be used to inform on-farm management practices or identify areas where further health or welfare investigations are required. Therefore, there is great potential for the application of these indicators by producers, veterinary surgeons, farm assurance and certification assessors, or farm animal welfare inspectors as robust and feasible tools for the on-farm assessment of young lamb welfare.

Introduction

Assessment of the welfare state of farm animal species is required for a number of purposes including demonstration of compliance with national and international legal farm animal welfare standards; compliance with private food sector farm assurance schemes; and for on-farm monitoring of animal welfare by farmers and their veterinary advisors (Veissier *et al.*, 2008).

† Present address: Animal Welfare and Behaviour Group, University of Bristol, School of Veterinary Science, Langford, Bristol BS40 5DU, UK. E-mail: C.J.Phythian@bristol.ac.uk

In the first stages of the development of animal welfare assessment methods, many resource-based measures were used to approximate the welfare state of the animal or animals in a husbandry system (Amon *et al.*, 2001). In an effort to more accurately reflect the animal's own welfare experience, it was considered that animal-based measures of welfare assessment were more appropriate (Main *et al.*, 2003) and were consequently developed for many species (Whay *et al.*, 2003a; Whay *et al.*, 2003b; Anzuino *et al.*, 2010). Such measures included indicators of disease (lameness scoring), nutritional status (body condition score) and environmental conditions (cleanliness scoring). Recently, more holistic behavioural measures of animal welfare state such as Qualitative Behavioural Assessment have been applied to many farm species (Wemelsfelder *et al.*, 2012; Wickham *et al.*, 2012). Consequently, several welfare assessment protocols that include animal- and resource-based assessment measures have been developed for many farm species (Whay *et al.*, 2003a; Knierim and Winckler, 2009; Anzuino *et al.*, 2010). However, currently, there are comparatively few validated welfare assessment measures for the assessment of sheep welfare. Therefore, the aim of this study was to develop and test non-invasive animal-based welfare measures of young lamb welfare for use in on-farm welfare assessments.

Young lambs are managed in a variety of ways in the United Kingdom, from intensive, indoor-lambing flocks that require a high input of labour and resources, to extensive outdoor lambing systems with a consequent risk of exposure to climatic extremes and predation (Dwyer, 2008). Lambs may be raised by their birth ewe or fostered or reared as 'orphan lambs' on milk substitute, often with automated feeding systems. A range of welfare issues for lambs were recently identified by a UK sheep welfare expert panel (Phythian *et al.*, 2011a) and included lamb mortality, starvation, hypothermia and the presence of infectious and inheritable diseases. Interestingly, in a recent report of sheep flock health issues in Scandinavia, a similar range of welfare concerns for lambs were cited (Ulvund, 2012).

The welfare indicators in this study were required to meet a number of criteria. First, they should have validity, in that they should be meaningful measures of an aspect of an animal's welfare state; second, they should be reliable, such that consistent results are identified when applied by different observers; third, as with any diagnostic test, their test accuracy should be determined in terms of their sensitivity (Se) and specificity (Sp). Finally, they should be feasible, meaning they are practical to apply on farms and within any cost and time requirements.

The specific aspect of lamb welfare that was evaluated by each indicator outcome, that is the content validity of each measure, was previously established by a sheep welfare expert panel (Phythian *et al.*, 2011a). The study described here examines their reliability, sensitivity and specificity, and feasibility. Reliability was assessed by examining test agreement between and within different observers; an approach used by other studies in the development of welfare indicators for other species (Mullan *et al.*, 2011). However, the assessment of sensitivity and specificity of novel tests can be problematic in the absence of a true gold standard test (a test, which can be used to determine as far as possible the true welfare status of the animal). A number of approaches have been used to overcome this in the development of welfare indicators, including, the use of a 'Pseudo Gold Standard' (Nielsen *et al.*, 2004; Bertrand *et al.*, 2005) and Bayesian methods, such as latent class analysis (LCA; Hui and Walter, 1980). LCA has frequently been applied in the evaluation of veterinary diagnostic tests (Nielsen *et al.*, 2004; Toft *et al.*, 2007a; Bonde *et al.*, 2010), and in the evaluation of the diagnostic accuracy of clinical observations of animal-based outcomes of pigs (Baadsgaard and Jørgensen, 2003). Accordingly, this approach was applied in the current study. Finally, the feasibility of applying the measures was subjectively assessed by the researchers during the on-farm assessment visits.

Material and methods

Study population

Sheep farms ($n = 50$) located within a 120 mi radius of the University of Liverpool, School of Veterinary Science, Leahurst campus were identified through contact with their local veterinary practice during January 2008 to July 2008. Each farm was contacted by telephone and visited in person to assess suitability for inclusion in the study based on the following criteria: (1) lambing period, (2) farm location and (3) informed consent to participate. Seventeen sheep farms in Northwest England and North Wales, consisting of lowland ($n = 9$), upland ($n = 2$) and hill flock ($n = 3$) types were recruited. This farm population consisted of 14 commercial farms (those that produced lambs for meat consumption – either finished on-farm or sold as 'stores'), two pedigree farms (flocks producing purebred breeding stock with a documented genetic history), and one small-holding in which 21 ewes were managed as pets. The number of breeding ewes present on each farm at lambing time is shown in Table 1. Each study farm was classified as either an indoor ($n = 16$), or outdoor ($n = 1$) lambing flock. Seventy-one percent of the study farms belonged to a farm assurance scheme. Two out of

Table 1 *Flock size according to the number of breeding ewes during lambing visits*

No. of breeding ewes	No. of farms
⩽60	1
61 to 150	1
151 to 250	0
251 to 400	3
401 to 600	4
601 to 800	1
801 to 1000	1
1001 to 1200	3
1201 to 1500	2
⩾1501	1

17 farms were certified as organic and the remaining 15 farms were categorised as conventional flocks.

Animal-based indicators

The animal-based indicators included in the study protocol were selected on the basis of a scientific review and an expert consultation process. This work has been previously described (Phythian *et al.*, 2011a). Briefly, the experts selected animal-based welfare indicators that were considered to be sensitive to current on-farm welfare issues for lambs in Great Britain.

Ten animal-based indicators – demeanour, response to stimulation, posture, standing ability, abdominal fill, body condition, lameness, shivering, eye condition and salivation were scored using binary (presence or absence) and categorical scales to assess each of the welfare-related conditions in young lambs aged $\leqslant$6 weeks (Supplementary Table S1). Lambs were assessed by standing outside the pen or by walking around a field at a distance sufficient to allow examination of demeanour and lameness, but with minimal disturbance to the group. On occasion, following observation of demeanour, response to stimulation and shivering, lambs reared in individual pens would be lifted out of the pen to facilitate the closer observation required for the remaining indicators (posture, standing ability, abdominal fill, body condition, lameness, eye condition and salivation). To prevent mis-mothering and disturbance of ewe–lamb bonding behaviour, young lambs were not gathered at any stage of the assessment, and those aged $\leqslant$6-h old were not examined. In addition to the indicator outcomes, the location of the assessment – indoors or outdoors, the method of rearing – with a non-tethered or tethered ewe (physical restraint of the ewe to minimise aggressive behaviour and rejection of the lamb), or artificial rearing (lambs not reared with a ewe), and the management system – individual or group pens were recorded.

Trained observers

Four trained observers, who were considered to be representative of farm animal welfare assessors, were recruited. Observers 1 and 3 were farm animal veterinary surgeons and classed as experienced assessors, and observers 2 and 4 were classed as inexperienced, non-veterinary assessors (undergraduate animal science students). Observer 1, an experienced, veterinary assessor, who developed the indicator test methods and provided observer training, was designated the 'test standard observer' (TSO). All observers were provided with an on-farm assessment protocol, with which to familiarise themselves. The TSO also provided a 1-day on-farm training session, performed on two lowland, indoor lambing flocks in February 2009. Thereafter, each observer independently applied the set of animal-based indicators on a minimum sample of 30 lambs selected by the TSO. Combinations of two to four observers (Table 2) performed the indicator assessments on each study farm, whereas the TSO performed indicator assessments on all study farms. All observers were blinded to clinical or production records before conducting each farm visit. The study was approved by the University of Liverpool Ethics Committee (reference RETH000287).

Young lamb welfare assessments

During December 2008 to April 2009, a range of two to four observers independently assessed 10 animal-based welfare indicators on 966 young lambs from the 17 study farms. For the purposes of assessing observer reliability, a sample of 30 to 90 (median 59) lambs per farm was selected (Walter *et al.*, 1998). Given the nature and timing of the study, the farmer was not always available for interview, and it was not feasible to record the location of all lambs or the total number of animals present at the time of assessment. Therefore, the TSO estimated the total number of lambs available for assessment at the time of the visit and selected sample animals to include the variety of rearing and management systems on each farm.

Analysis of observer reliability

Reliability data were analysed using Stata version 10 (StataCorp LP, College Station, Texas, USA). The overall level of inter-observer reliability for multiple observer assessments was determined by Fleiss's kappa (κ; Fleiss, 1981). Cohen's κ (Cohen, 1960) was used to examine the paired agreement between the TSO and observers 2 to 4. All κ values were interpreted according to Fleiss (1981), whereby values $\geqslant$0.75 suggested 'excellent' agreement, $\kappa = 0.40$ to 0.75 indicated 'fair to good' agreement, and $\kappa \leqslant 0.40$ suggested 'poor' agreement. Graphical representation of scoring differences between the paired assessments of study observers and the TSO were used to examine for evidence of scoring bias.

Latent class models

Estimates of Se and Sp were produced for each observer of the inter-observer study using LCA (Hui and Walter, 1980). A Bayesian approach was selected in view of the relatively small population, and low proportion of some indicator outcomes (Bonde *et al.*, 2010). The model was essentially that proposed by Baadsgaard and Jørgensen (2003), where observers were assumed to be similar, that is, from the same random-effects model. LCA was performed in OpenBUGS software (Lunn *et al.*, 2009) using Markov Chain Monte Carlo (MCMC) sampling to obtain the joint posterior distribution of the model, and observer identity was included as a random effect. The first 10 000 samples were discarded as burn-in and the subsequent 10 000 iterations were used for posterior inference. MCMC chain convergence was visually assessed using time-series plots and Gelman–Rubin diagnostic plots using three sample chains with different initial values (Toft *et al.*, 2007b). The Se and Sp of each observer were provided with 95% posterior credibility intervals – the Bayesian analogue of confidence intervals. In addition, LCA predicted the Se and Sp of 'random' observers – randomly selected observers who may be expected to apply the welfare indicators in the future.

Table 2 *Total number and percentage (%) of young lambs observed with a positive score for each welfare indicator (assessed by the test standard observer)*

Indicator outcome	Total *n* observed	Percentage (%) observed (95% CI)
Dull demeanour	35	3.62 (0.77 to 6.48)
Unresponsive to stimulation	20	2.07 (0.60 to 3.54)
Signs of shivering	5	0.55 (0.01 to 1.12)
Recumbent, weak on standing	29	2.90 (1.56 to 4.24)
Hunched, tucked-up posture	27	2.81 (0.43 to 5.19)
Abnormal abdominal fill	19	1.97 (0.21 to 3.73)
Inappropriate body condition	48	5.09 (2.83 to 7.35)
Lame	20	2.09 (0.46 to 3.73)
Eye abnormality	63	6.66 (3.74 to 9.39)
Excess salivation	1	0.12 (0.00 to 0.38)

Results

Of the 966 lambs observed, 652 were from lowland (67.5%), 150 from upland (15.5%) and 164 lambs from hill study flocks (17.0%). Fifty percent of study lambs were managed in individual pens ($n = 487$), 23.5% managed indoors in groups ($n = 227$) and 26.1% managed outdoors ($n = 252$). Eighty-nine percent were observed as being reared with a non-tethered ewe ($n = 859$), 3.6% were observed to be reared with a tethered ewe ($n = 35$) and 5.6% were artificially reared (orphan) lambs ($n = 54$). The breed of each individual lamb was not recorded, but the sample population comprised a variety of purebred lambs including Welsh Mountain, Lleyn, Derbyshire Gritstone, Charollais, Grey-faced Dartmoor, Suffolk, Texel, Jacob, Welsh Mountain, Welsh Hill Speckled Face, Hampshire Down, and a variety of cross-bred lambs and terminal-sire crosses. The age distribution of sample lambs, estimated at the time of assessment from farmer reports, observed management practices and available lambing records, was categorised. Forty-one percent of lambs were aged 6 h to $<$3 days ($n = 399$), 38.7% were 3 to 7 days ($n = 372$) and 19.4% were $>$1 to $<$6 weeks ($n = 187$).

The assessment and manual recording of all 10 indicators took between 3 and 5 min per lamb. However, as close observation was required for the assessment of body condition, abdominal fill, eye condition and salivation, these indicators could not be assessed in 96% of lambs managed outdoors.

The TSO results for the proportion of each welfare indicator observed in the sample population are shown in Table 2. Overall, levels of inter-observer reliability (Table 3) for all welfare indicators were interpreted as 'fair–good' agreement and the assessment of eye condition was highly reliable ($\kappa > 0.72$) with very few scoring disagreements. Paired assessments with the TSO provided a range of Cohen's κ results (Table 3), but few scoring differences were found with the exception of the body condition assessments of observer 3, which resulted in a lower Se (0.38) for this indicator. LCA estimated that mean values for all observers were; Se $\geqslant 0.86$ and Sp $= 0.99$. With the exception of the indicators of posture and shivering, which were observed at a very low prevalence ($<$0.6%), latent class models predicted that randomly selected assessors would have had Se $\geqslant 0.74$ and Sp $\geqslant 0.98$ for all indicator assessments (Table 4). Few lambs with external salivation were observed ($n = 2$), which provided insufficient observations for latent class modelling (Table 4).

Discussion

The objective of this study was to develop and test the validity, reliability and feasibility of animal-based welfare indicators, which were developed following consultation of the scientific literature and expert opinion, as proxy measures of the welfare status of a young lamb.

Evaluation of diagnostic performance

The reliability of each welfare indicator was evaluated in accordance with the quality of reporting of reliability studies (QAREL; Lucas *et al.*, 2010). Kappa (κ) was selected as the method of analysing between-observer agreement. The interpretation of κ values requires consideration of the prevalence of the condition of interest, as a population with few affected animals will provide artificially low estimates of reliability (Hoehler, 2000). Lameness and eye condition were both found to have excellent levels of observer agreement. By contrast, other indicators including salivation and shivering that were observed in $<$1% of the population produced lower levels of test performance. This is likely a consequence of the low prevalence of these conditions in the study population. It is possible that higher κ values would have been estimated if the prevalence of affected animals was higher in the test population. Therefore, it may be argued that the prevalence of some of the conditions of interest in this study may have been too low to assess observer reliability effectively. However, similar issues for the evaluation of κ reliability may occur in populations with a high prevalence of outcomes associated with poor animal welfare (Burn *et al.*, 2009). Although a balanced prevalence of indicators, that is, a 50% prevalence of affected and unaffected animals is ideal for reliability studies (Hoehler, 2000), achieving a representative sample with a mixed prevalence of welfare indicators is difficult in

Table 3 *Overall level of inter-observer agreement (Fleiss κ, 95% confidence interval), and paired agreement between the test standard observer and observers 2 to 4 (Cohen's κ, 95% CI)*

		Paired agreement	
Indicator	Overall agreement Fleiss κ (95% CI)	Observer identity	Cohen's κ (95% CI)
Demeanour	0.54 (0.45 to 0.59)	2	0.55 (0.39 to 0.71)
		3	0.52 (0.21 to 0.83)
		4	0.44 (0.27 to 0.62)
Stimulation	0.57 (0.52 to 0.63)	2	0.72 (0.42 to 1.00)
		3	0.56 (0.34 to 0.79)
		4	0.56 (0.34 to 0.79)
Shivering	0.55 (0.35 to 0.66)	2	0.75 (0.41 to 1.00)
		3	a
		4	0.40 (−0.14 to 0.94)
Standing ability	0.54 (0.45 to 0.55)	2	0.70 (0.57 to 0.86)
		3	0.66 (0.30 to 1.00)
		4	0.58 (0.43 to 0.74)
Posture	0.45 (0.34 to 0.48)	2	0.50 (0.32 to 0.69)
		3	0.76 (0.50 to 1.00)
		4	0.50 (0.32 to 0.67)
Abdominal fill	0.44 (0.42 to 0.47)	2	0.57 (0.38 to 0.76)
		3	1.00 (1.00 to 1.00)
		4	0.39 (0.12 to 0.66)
Body condition	0.72 (0.60 to 0.74)	2	0.71 (0.60 to 0.82)
		3	0.49 (0.07 to 0.92)
		4	0.76 (0.66 to 0.87)
Lameness	0.68 (0.53 to 0.69)	2	0.70 (0.53 to 0.87)
		3	0.72 (0.42 to 1.00)
		4	0.81 (0.67 to 0.95)
Eye condition	0.72 (0.63 to 0.77)	2	0.76 (0.66 to 0.86)
		3	0.84 (0.69 to 0.99)
		4	0.66 (0.54 to 0.78)
Salivation	0.71 (0.54 to 1.00)	2	0.75 (0.41 to 1.00)
		3	a
		4	0.39 (−0.14 to 0.94)

[a]Insufficient observations to produce estimate.

field studies (Burn *et al*., 2009) such as this. In the present study, the evaluation of diagnostic test performance was conducted under field conditions (Lucas *et al*., 2010), and in spite of some of the issues outlined, all indicators were intepreted with 'fair–good' levels of observer agreement.

LCA

Reliability is only the first step in evaluation of diagnostic tests, the second step being evaluation of sensitivity (Se) and specificity (Sp). In the absence of a gold standard, observer 1 was designated as 'TSO' on the basis of their role in the development of the indicators. Latent class models do not require a comparative reference standard, and also offer a means of predicting the Se and Sp of randomly selected observers (Hui and Walter, 1980). LCA assumes the prevalence of indicator outcomes differs across different populations (Hui and Walter, 1980); a feature which was observed across different farms in this study. Therefore, LCA can offer an amenable method of evaluating tests conducted on farms where *a priori* information may not be known. To our knowledge, no prior estimates of conditions in young lambs, such as lameness, thin body condition or ocular abnormalities, have been previously published, and this is the first time that LCA has been applied to assess the test validity of young lamb welfare indicators.

Trained observers

In accordance with QAREL standards, the indicators were tested by observers from veterinary and animal science backgrounds as these were considered to be a fair representation of the assessors, who may be expected to apply the measures in the future (Lucas *et al*., 2010). For example, statutory, on-farm welfare inspections, in the United Kingdom, are currently undertaken by veterinary surgeons, whereas private welfare assurance schemes typically employ inspectors with a background in agriculture or other animal-related experience. Certain observer characteristics, such as level of training, experience and occupation can influence the level of observer agreement (Kristensen *et al*., 2006), thus observers with differing levels of experience in assessing the health and

Table 4 *Bayesian posterior estimates (median, 95% PCI) of the sensitivity (Se) and specificity (Sp) of assessments of animal-based indicators of young lamb welfare performed by study observers (1 to 4) and randomly selected observers (random)*

Indicator	Observer identity	Se (95% PCI)	Sp (95% PCI)
Demeanour	1	0.75 (0.58 to 0.89)	0.98 (0.97 to 0.99)
	2	0.85 (0.69 to 0.99)	1.00 (0.99 to 1.00)
	3	0.77 (0.50 to 0.96)	0.98 (0.96 to 1.00)
	4	0.70 (0.47 to 0.86)	1.00 (1.00 to 1.00)
	Random	0.78 (0.52 to 0.96)	0.98 (0.88 to 1.00)
Stimulation	1	0.55 (0.40 to 0.70)	1.00 (1.00 to 1.00)
	2	0.74 (0.57 to 0.89)	1.00 (1.00 to 1.00)
	3	0.72 (0.34 to 0.95)	1.00 (1.00 to 1.00)
	4	0.30 (0.18 to 0.45)	1.00 (1.00 to 1.00)
	Random	0.77 (0.48 to 0.99)	0.98 (0.86 to 1.00)
Shivering	1	0.85 (0.41 to 1.00)	1.00 (1.00 to 1.00)
	2	0.56 (0.23 to 0.90)	1.00 (1.00 to 1.00)
	3	0.58 (0.01 to 1.00)	1.00 (1.00 to 1.00)
	4	0.37 (0.23 to 0.81)	1.00 (1.00 to 1.00)
	Random	0.64 (0.00 to 1.00)	0.99 (0.99 to 1.00)
Standing ability	1	0.82 (0.62 to 0.94)	1.00 (0.99 to 1.00)
	2	0.80 (0.61 to 0.91)	1.00 (0.99 to 1.00)
	3	0.81 (0.59 to 0.96)	0.99 (0.98 to 1.00)
	4	0.80 (0.60 to 0.90)	0.99 (0.97 to 0.99)
	Random	0.80 (0.60 to 0.82)	0.99 (0.98 to 1.00)
Posture	1	0.75 (0.42 to 1.00)	0.99 (0.99 to 1.00)
	2	0.56 (0.30 to 0.82)	0.99 (0.98 to 1.00)
	3	0.70 (0.38 to 0.99)	0.99 (0.98 to 1.00)
	4	0.62 (0.36 to 0.87)	0.99 (0.98 to 1.00)
	Random	0.67 (0.32 to 1.00)	0.99 (0.99 to 1.00)
Abdominal fill	1	0.96 (0.56 to 1.00)	0.99 (0.99 to 1.00)
	2	0.98 (0.75 to 1.00)	0.99 (0.98 to 1.00)
	3	0.98 (0.78 to 1.00)	0.99 (0.99 to 1.00)
	4	0.39 (0.12 to 0.71)	0.99 (0.99 to 1.00)
	Random	0.91 (0.00 to 1.00)	0.99 (0.98 to 1.00)
Body condition	1	0.84 (0.70 to 0.95)	0.99 (0.99 to 1.00)
	2	0.80 (0.66 to 0.91)	0.99 (0.99 to 1.00)
	3	0.38 (0.07 to 0.80)	0.99 (0.99 to 1.00)
	4	0.90 (0.76 to 0.99)	0.99 (0.99 to 1.00)
	Random	0.74 (0.21 to 0.97)	0.99 (0.99 to 1.00)
Lameness	1	0.80 (0.60 to 1.00)	1.00 (0.99 to 1.00)
	2	0.73 (0.54 to 0.87)	1.00 (0.99 to 1.00)
	3	0.73 (0.47 to 0.91)	1.00 (0.99 to 1.00)
	4	0.76 (0.59 to 0.91)	1.00 (0.99 to 1.00)
	Random	0.76 (0.56 to 0.96)	1.00 (0.99 to 1.00)
Eye condition	1	0.89 (0.80 to 0.89)	0.99 (0.98 to 0.99)
	2	0.87 (0.75 to 0.88)	0.99 (0.99 to 1.00)
	3	0.89 (0.79 to 0.89)	0.99 (0.98 to 1.00)
	4	0.86 (0.73 to 0.87)	0.99 (0.98 to 0.99)
	Random	0.88 (0.77 to 0.97)	0.99 (0.99 to 1.00)

PCI = posterior credibility interval.

welfare of individual lambs were included. All observers performed independent assessments and were blinded to historical and clinical information (Lucas *et al.*, 2010). However, because of the study setting, it was not possible to blind observers to cues such as farm cleanliness and hygiene, and presence of rearing equipment or medicines, which may have alerted observers to the presence of certain health or welfare issues (Petersen *et al.*, 2004).

The size of the observer pool ($n = 4$) was determined by feasibility and farmer compliance, as it was found to be impractical to have more than four people simultaneously observing and recording the same animals in a lambing shed, while the farmers were trying to continue with the work of supervising the lambing flock. Although this may be considered a small number of observers in comparison to some studies (Mullan *et al.*, 2011), it is similar to the numbers

used in other published welfare indicator reliability studies (Channon *et al.*, 2009; Kaler *et al.*, 2009; Foddai *et al.*, 2012), and is above the minimum number of observers recommended for reliability studies (Walter *et al.*, 1998).

Comparison of the LCA results of the different observers suggested that observer 1 was a suitable choice for the role of the 'test standard' and provider of observer training as this observer generally had higher levels of Se compared with other observers. As TSO developed the scoring systems, this higher level of test performance may have been due to a greater understanding of the indicator case definitions. It also may reflect differences in observational qualities of different assessors, which were not explored in this study.

Observers 1 and 3 were considered to be the most experienced assessors, both achieved excellent levels of reliability and Se $>$ 0.98 for many indicator assessments. Comparing observer performance against a test standard or within groups of observers can be one way of identifying specific issues with welfare indicator assessments. For example, a lower observer Se might suggest that further training of observers would be advantageous. This might be of particular relevance for measures included as part of on-farm welfare inspections and farm assurance schemes.

Welfare indicator outcomes

Demeanour, standing ability, lameness and eye condition produced good levels of Se and Sp when applied by all study observers. In particular, eye condition stood out as an indicator with high Se and Sp suggesting that ocular abnormalities in young lambs were clearly recognised. Since before their participation in this study not all observers were familiar with the assessment of eye conditions, the high level of diagnostic performance may reflect both the clear scoring scales used, and the ease of identifying lambs with ocular discharges and lesions. The presence of entropion was considered to be a particular on-farm welfare issue for young lambs by an expert panel in Great Britain (Phythian *et al.*, 2011a) and Scandinavia (Ulvund, 2012). This was supported by the study results that identified over 6% of the study population had an ocular abnormality; most frequently diagnosed as entropion. This finding might inform farmer knowledge-exchange events and routine management actions aimed at improving the health and welfare of young lambs. Therefore, eye condition appears to be a highly relevant indicator to include in future lamb health and welfare inspection tools.

A simple binary lameness scoring system was developed to distinguish between 'sound' and 'lame' lambs and the clarity of this scale might explain the high level of observer reliability achieved. Lameness in this study population was most frequently diagnosed to be the result of septic arthritis (joint ill), which produces severe pathological changes in synovial joints, resulting in stiffness, joint swellings and severe gait abnormalities, which are easily recognised (Angus *et al.*, 1991).

With the exception of observer 4, the assessment of abdominal fill produced good inter-observer agreement and Sp $\geqslant$ 0.98. For demeanour assessments, it was important to determine the difference between a healthy, sleeping lamb and a dull, depressed lamb of poor welfare status and the responsiveness of young lambs to stimuli, such as movement or palpation by the assessor. It may be considered useful to combine demeanour and stimulation tests into a single indicator in which lambs could be scored as either bright, alert and responsive to stimulation or dull, depressed and unresponsive to stimulation.

Latent class models also predicted that a randomly selected observer would have Sp $>$ 0.97 for all indicators. This predicted diagnostic ability is derived from the results obtained by observers 1 to 4, which may reflect the quality of training and indicator definitions provided. Additional body condition training may be required, but overall the results suggest that other trained assessors may achieve good level of Se and Sp when applying these measures in the future.

The high level of Sp ($>$0.98) found for all indicators may suggest that the measures are better at identifying lambs with good welfare, whereas the lower Se might imply that observers may miss some animals with poor welfare. However, the results do need to be interpreted in light of the proportion of affected and unaffected lambs in the study population, as well as acknowledging the trade-off that occurs between the level of diagnostic Se and Sp (Greiner and Gardner, 2000).

In common with Baadsgaard and Jørgensen (2003), because of the feasibility of conducting on-farm studies, study farms were not randomly selected. Instead, farm types commonly found in England and Wales were selected so that the tests were applied under the range of management styles. Given the non-random selection of farms, it is recognised that this population may be biased towards farms of higher welfare status or those with regular contact with a veterinary surgeon. Similarly within the flock, for reasons of feasibility the lambs were not randomly selected for indicator assessments. It is acknowledged that this approach can introduce bias at flock level, but as our purpose was to evaluate the indicators and not the welfare status of the flock or the national sheep population, it was considered appropriate. Further work is needed to determine the best approach for selecting the sample of lambs that are assessed for the purposes of an on-farm welfare inspection.

A sample size of 17 farms (three different farm types), 966 lambs and four observers were used to assess the performance of the welfare indicators in this study. Examination of the literature on reliability studies of welfare indicators shows that there is no consistent pattern to the number and type of farms sampled, number of animals examined, or the number of observers used. For example, in a recent reliability study of pig welfare indicators, 53 observers were used to assess $\sim$400 pigs (Mullan *et al.*, 2011), a study of lameness scoring in cattle (Channon *et al.*, 2009) used five observers and 83 cows on one farm. Studies investigating the reliability of sheep lameness scoring used a sample size of three observers and 65 video clips of individual sheep (Kaler *et al.*, 2009b); and three observers and 80 photographs of sheep

(Foddai *et al.*, 2012). As no prevalence estimates were available for the lamb conditions under examination in this study, optimal sample size estimates were not performed. This is not an unfamiliar issue and is a factor that the WelfareQuality® project recognises as an outstanding issue for other farm animal welfare assessment protocols (Knierim and Winckler, 2009). However, the number and range of farm types that the indicators were tested on in the study presented here, together with the larger sample size of lambs, can be considered to add to the robustness of the results. Furthermore, the fact the reliability, Se and Sp of these indicators was tested on farms rather than using video clips and photographs could also be considered to enhance their external validity and applicability in on-farm assessments.

The timing of assessment visits is also worth further consideration. A one-off on-farm welfare assessment outside of the lambing season may not take into account the wide variety of welfare issues that specifically arise during lambing time and are of concern for young lamb welfare (Dwyer, 2008; Phythian *et al.*, 2011a). However, on-farm welfare assessments conducted during lambing time may previously have been discouraged because of concerns over disturbing the flock and the producer during this critical period. Lambing performance is the ultimate outcome of short- and long-term management decisions and inputs, which can have a significant impact on the health and welfare of both the ewe and lamb during the pre- and peri-parturient period. Welfare issues, such as dystocia, hypothermia and starvation, may not be identified using resource and management-based measures alone, which frequently rely on the maintenance of accurate lambing records. Similarly, welfare issues specific to the lambing period might not be transparent if the outcomes are applied later in the production calendar. Unlike previous lamb welfare assessment tools that were developed to be used by producers (Matheson *et al.*, 2011), a clear advantage of the animal-based measures tested in this study was that they did not rely heavily on farmer involvement, did not impinge on farm management practices, and were designed specifically for ease of use. Overall, the measures were also found to be feasible, requiring 3 to 5 min to assess and record per individual lamb. Therefore, the indicators could be readily applied during the lambing season by both external assessors and producers alike, and therefore may be considered to have a high degree of social acceptability by these different sectors.

Development of a young lamb welfare assessment system

We have studied individual indicators in terms of their diagnostic performance and have not attempted to produce any overall system of assessment. The indicators tested in the present study should now be evaluated for inclusion as part of the development of a welfare assessment system for use across different conditions and management systems. All of the tested indicators could be applied to indoor-housed lambs, but it was not always feasible to assess measures requiring close inspection – body condition, eye condition, salivation and abdominal fill, in 96% of lambs managed in extensive, outdoor environments. Indicators that could be consistently applied to lambs in all the rearing and management systems tested included demeanour, stimulation, standing ability, posture and lameness.

A simple approach would include only the common indicators to an assessment system. Alternatively, the selection of young lamb indicators could be tailored to the on-farm management system. However, one implication of this approach is that it might fail to identify neonatal and young lambs affected by conditions such as entropion, or poor body condition (Dwyer, 2008, Phythian *et al.*, 2011a). Further research into the development of lamb welfare assessment systems, combining different indicators according to the management system, should aim to provide sufficiently robust results to provide a reliable overall picture of the welfare status of the animals and to prevent some welfare issues from remaining undetected.

One solution could be to develop a young lamb welfare score, similar to the lamb vigour scoring system of Matheson *et al.* (2011). A welfare index made up of composite indicators could be weighted using expert opinion to ensure that lambs managed under different farming systems are capable of attaining the same welfare scores. This is of particular relevance, if different indicators are used as part of farm assurance or statutory welfare inspection protocols.

A different approach would be to use the indicators on all management systems as initial or 'iceberg' measures in the first stages of a young lamb welfare assessment protocols and at a subsequent stage complement them by the inclusion of resource-, and management-based measures, such as information on perinatal mortality of ewes and lambs, tail-docking and castration policies, on-farm records, housing hygiene and nutritional provision.

Indicators tested in this study were based on the Five Freedoms and as such concentrate on the inputs and resources required to provide good animal welfare. This approach, though not unique, may have been the reason why many indicators were focused on physical measures of health and welfare. It may also reflect the awareness or bias of the expert panel towards measures of health, injury and production (Phythian *et al.*, 2011a). There is an increasing move in the field of animal welfare science towards assessments, which focus on the quality of an animal's life and the authors recognise the value of including positive measures of animal welfare and suggest that a qualitative behaviour assessment (QBA) approach (Wemelsfelder and Lawrence, 2001) could be tested as a means of exploring positive welfare states in young lambs. QBA has previously been examined in terms of its reliability and feasibility as an on-farm indicator of adult sheep and growing/fat lamb welfare (Phythian *et al.*, 2011b), and may be particularly useful for extensive conditions, which present particular difficulties for on-farm welfare assessments (Dwyer, 2009).

Conclusion

Observers achieved good levels of test reliability, Se and Sp for animal-based indicators of young lamb welfare. Evaluation

of diagnostic performance may be affected by the low level of lambs in this study population observed with specific welfare conditions, such as hypothermia and starvation. Given that the tests were highly capable of detecting welfare conditions, such as lameness, and thin body condition, on a sample population with relatively few affected animals, it is recommended that the measures are tested on a sample of young lambs with a higher level of these conditions.

Acknowledgements

The study was funded by Defra as part of the AW1025 grant. 'Development of indicators for the on-farm assessment of sheep welfare'. The authors also gratefully acknowledge the support of the expert panel, study farms and observers.

Supplementary materials

For supplementary material referred to in this article, please visit http://dx.doi.org/10.1017/S1751731113000487

References

Amon T, Amon B, Ofner E and Boxberger J 2001. Precision of assessment of animal welfare by the 'TGI 35 L' Austrian needs index. Acta Agriculturae Scandinavica Section A – Animal Science 51, 114–117.

Angus K 1991. Arthritis in lambs and sheep. In Practice 13, 204–207.

Anzuino K, Bell NJ, Bazeley KJ and Nicol CJ 2010. Assessment of welfare on 24 commercial UK dairy goat farms based on direct observations. Veterinary Record 167, 774–780.

Baadsgaard NP and Jørgensen E 2003. A Bayesian approach to the accuracy of clinical observations. Preventive Veterinary Medicine 59, 189–206.

Bertrand P, Benichou J, Grenier P and Chastang C 2005. Hui and Walter's latent-class reference-free approach may be more useful in assessing agreement than diagnostic performance. Journal of Clinical Epidemiology 58, 688–700.

Bonde M, Toft N, Thomsen PT and Sorensen J 2010. Evaluation of sensitivity and specificity of routine meat inspection of Danish slaughter pigs using latent class analysis. Preventive Veterinary Medicine 94, 165–169.

Burn CC, Pritchard JC and Whay HR 2009. Observer reliability for working equine welfare assessment: problems with high prevalences of certain results. Animal Welfare 18, 177–187.

Channon AJ, Walker AM, Pfau T, Sheldon IM and Wilson AM 2009. Variability of manson and leaver locomotion scores assigned to dairy cows by different observers. Veterinary Record 164, 388–392.

Cohen J 1960. A coefficient of agreement for nominal scales. Education and Psychological Measurement 20, 37–46.

Dwyer CM 2008. The welfare of the neonatal lamb. Small Ruminant Research 76, 31–41.

Dwyer CM 2009. Welfare of sheep: providing for welfare in an extensive environment. Small Ruminant Research 86, 14–21.

Farm Animal Welfare Council (FAWC) 1994. Report on the welfare of sheep. PB 1755. Farm Animal Welfare Council (FAWC) publication, London, UK.

Fleiss LL 1981. Statistical methods for rates and proportions. John Wiley and Sons, New York, USA.

Foddai A, Green LE, Mason SA and Kaler J 2012. Evaluating observer agreement of scoring systems for foot integrity and footrot lesions in sheep. BMC Veterinary Research 8, 65.

Greiner M and Gardner IA 2000. Epidemiologic issues in the validation of veterinary diagnostic tests. Preventive Veterinary Medicine 45, 3–22.

Hoehler FK 2000. Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. Journal of Clinical Epidemiology 53, 499–503.

Hui SL and Walter SD 1980. Estimating the error rates of diagnostic tests. Biometrics 36, 167–171.

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.

Knierim U and Winckler C 2009. On-farm welfare assessment in cattle: validity, reliability and feasibility issues and future perspectives with special regard to the Welfare Quality® approach. Animal Welfare 18, 451–458.

Kristensen E, Dueholm L, Vink D, Andersen JE, Jakobsen EB, Illum-Nielsen S, Petersen FA and Enevoldsen C 2006. Within- and across-person uniformity of body condition scoring in Danish Holstein cattle. Journal of Dairy Science 89, 3721–3728.

Lucas NP, Macaskill P, Irwig L and Bogduk N 2010. The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). Journal of Clinical Epidemiology 63, 854–861.

Lunn D, Spiegelhalter D, Thomas A and Best N 2009. The Bugs project: evolution, critique and future directions. Statistics in Medicine 28, 3049–3067.

Main DCJ, Kent JP, Wemelsfelder F, Ofner E and Tuyttens FAM 2003. Applications for methods of on-farm welfare assessment. Animal Welfare 12, 523–528.

Matheson SM, Rooke JA, McIlvaney K, Jack M, Ison S, Bünger L and Dwyer CM 2011. Development and validation of on-farm behavioural scoring systems to assess birth assistance and lamb vigour. Animal 5, 776–783.

Mullan S, Edwards SA, Butterworth A, Whay HRA and Main DCJ 2011. Inter-observer reliability testing of pig welfare outcome measures proposed for inclusion within farm assurance schemes. Veterinary Journal 190, 100–109.

Nérette P, Stryhn H, Dohoo I and Hammell L 2008. Using pseudogold standards and latent-class analysis in combination to evaluate the accuracy of three diagnostic tests. Preventive Veterinary Medicine 85, 207–225.

Nielsen LR, Toft N and Ersboll AK 2004. Evaluation of an indirect serum ELISA and a bacteriological faecal culture test for diagnosis of Salmonella serotype Dublin in cattle using latent class models. Journal of Applied Microbiology 96, 311–319.

Petersen HH, Enoe C and Nielsen EO 2004. Observer agreement on pen level prevalence of clinical signs in finishing pigs. Preventive Veterinary Medicine 64, 147–156.

Phythian CJ, Michalopoulou E, Jones PH, Winter AC, Clarkson MJ, Stubbings LA, Grove-White D, Cripps PJ and Duncan JS 2011a. Validating indicators of sheep welfare through a consensus of expert opinion. Animal 5, 943–952.

Phythian CJ, Wemelsfelder F, Michalopoulou E and Duncan JS 2011b. Qualitative behaviour assessment in sheep: consistency across time and association with health indicators. In Proceedings of the 5th international workshop on the assessment of animal welfare at farm and group level (ed. T Widowski, P Lawlis and K Sheppard), pp. 14. Wageningen Academic Publishers, Wageningen, The Netherlands.

Toft N, Akerstedt J, Tharaldsen J and Hopp P 2007a. Evaluation of three serological tests for diagnosis of Maedi-Visna virus infection using latent class analysis. Veterinary Microbiology 120, 77–86.

Toft N, Innocent GT, Gettinby G and Reid SWJ 2007b. Assessing the convergence of Markov Chain Monte Carlo methods: an example from evaluation of diagnostic tests in absence of a gold standard. Preventive Veterinary Medicine 79, 244–256.

Ulvund MJ 2012. Important sheep flock health issues in Scandinavia/northern Europe. Small Ruminant Research 106, 6–10.

Veissier I, Butterworth A, Bock B and Roe E 2008. European approaches to ensure good animal welfare. Applied Animal Behaviour Science 113, 279–297.

Walter SD, Eliasziw M and Donner A 1998. Sample size and optimal designs for reliability studies. Statistics in Medicine 17, 101–110.

Wemelsfelder F and Lawrence AB 2001. Qualitative assessment of animal behaviour as an on-farm welfare-monitoring tool. Acta Agriculturae Scandinavica Section A – Animal Science 30, 21–25.

Wemelsfelder F, Hunter AE, Paul ES and Lawrence AB 2012. Assessing pig body language: agreement and consistency between pig farmers, veterinarians, and animal activists. Journal of Animal Science 90, 3652–3665.

Whay HR, Main DCJ, Green LE and Webster AJF 2003a. Animal based measures for the assessment of welfare state of dairy cattle, pigs, and laying hens: consensus of expert opinion. Animal Welfare 12, 205–217.

Whay HR, Main DCJ, Green LE and Webster AJF 2003b. An animal-based welfare assessment of group-housed calves on UK dairy farms. Animal Welfare 12, 611–617.

Wickham SL, Collins T, Barnes AL, Miller DW, Beatty DT, Stockman C, Blache D, Wemelsfelder F and Fleming PA 2012. Qualitative behavioral assessment of transport-naïve and transport-habituated sheep. Journal of Animal Science 90, 4523–4535.