**National farm assurance scheme demonstrates welfare outcome improvements for sustainable intensification of dairy production**

Siobhan Mullan† \*, Philippa Wiltshire††, Kate Cross††, David CJ Main§, Kate Still§§, Madeleine Crawley§§, Andrew W Dowsey †

† University of Bristol Veterinary School, Langford BS40 5DU, UK

†† Red Tractor Assurance, Rural Innovation Centre, Stoneleigh Park, Kenilworth CV8 2LG, UK

§Royal Agricultural University, Cirencester, Gloucestershire, GL7 6JS, UK

§§Soil Association, Spear House, 51 Victoria Street, Bristol, BS1 6AD, UK

\* Corresponding author: Now at: UCD School of Veterinary Sciences, Dublin. Tel: +44 (0) 7437 962383 e-mail: [siobhan.mullan@ucd.ie](mailto:siobhan.mullan@ucd.ie)

Word count: 5568 words

**Abstract**

Animal welfare is a key pillar of sustainability in livestock farming but can be challenging to deliver improvements in when productivity is high. In October 2013 the Red Tractor UK national dairy assurance scheme introduced welfare outcome monitoring on its 11,500 farms covering 95% of UK milk produced. Here we show that 98 farm assurance assessors achieved high levels of agreement with a gold standard and report data they collected for three years from 248,689 cows, typically 10 cows per farm, during 19,899 audits. Between 2013/14 and 2015/16, the estimated national prevalence fell significantly for lameness (from 10.0% to 7.9%), dirtiness (from 12.4% to 9.2%), ‘hairloss, lesions and swellings’ (from 9.3% to 6.5%), and fat cows (from 2.4% to 1.9%). This occurred at a time when milk yield per cow increased. We have demonstrated an effective implementation strategy suitable for uptake internationally to align with societal sustainability goals.

**Key words**: sustainable intensification, animal welfare, farm assurance, dairy cows, welfare outcomes

**Introduction**

The aim of sustainable intensification of livestock agriculture is to improve yields without commensurate negative impact on the environment or animal welfare. Animal welfare is a key component of sustainability of livestock farming systems, alongside environmental and social impacts (Godfray 2015), although is sometimes overlooked in analyses (e.g. Dillon et al. 2016). Reliable information about animal welfare is required to help guide consumers in animal product purchasing decisions and, when at a large or national-scale, helps to inform policy-makers on the sustainability of their animal farming systems (Whitfield & Marshall 2017; Buller et al. 2018). In some countries, including the UK, privately administered robust farm assurance schemes cover a large proportion of farms in an agricultural sector and therefore have the potential to collect large-scale data relating to animal welfare and other sustainability elements. In countries where farm assurance schemes are not present large corporations, farmer co-operatives or governmental institutions can provide suitable platforms for collation of sustainability data.

Dairy farm assurance in the UK has more than two decades of history. Currently all major buyers of milk require accreditation to the Red Tractor Assurance dairy scheme representing over 95% of all milk produced (Red Tractor 2015), with 11,500 member farms in our study period between 2013 and 2015 (Red Tractor 2013, 2014, 2015). The Red Tractor (RT) scheme has standards at, or slightly above, legislative requirements for a range of areas from traceability to environmental management as well as cattle housing and welfare (Red Tractor 2017). The standards are audited on farms by an independent third party at 18-month intervals.

Cattle welfare assessment was initially focussed on resource-based requirements. Nevertheless, there was increasing political and industry pressure to improve welfare assurance through strengthening the recording of health and welfare in the health plan (FAWC 2009), and the introduction of formal welfare scoring within the audit (CHAWG 2010). In 2011 industry stakeholder discussions were instigated to ensure standardised assessment methods. Subsequently, the Red Tractor, Soil Association and RSPCA Assurance schemes collaborated with us to pilot measures on farms within inspections. During this development phase, national scale welfare outcome monitoring was being implemented with support from us and others for laying hens (Mullan et al. 2011; Main et al. 2012) and pigs (Pandolfi et al. 2017). Considerations were given to the validity, reliability and feasibility of welfare measures that could be undertaken within audit time constraints of approximately 20 minutes for observation followed by 10 minutes for a farmer discussion (Main et al. 2012). Similar theoretical work to that we had already conducted for laying hens (Main et al. 2012) was undertaken to understand the relationship between samples sizes and confidence of prevalence estimates both at the farm level and at the population level for the overall scheme.

In this study we report the success of national scale welfare outcome sustainability metric observations undertaken during RT dairy audits from October 2013 to September 2016. We present the results of training assessors to undertake the assessments, as well as year on year changes in outcome prevalence and their relationship to farm characteristics recorded by the scheme.

**Materials and Methods**

A requirement for assessors to observe at least 10 cows per farm for four welfare outcomes (mobility (lameness); cleanliness (dirtiness); body condition (thin/fat cows); hairloss, lesions and swellings (HLS); see Table 1) was included for all Red Tractor (RT) dairy farm assurance audits from October 2013. The sample size of 10 cows was decided upon to accommodate time constraints of assessors whilst still providing a reliable whole scheme prevalence estimate. This was a new initiative and whilst some assessors may have previously had informal training in some welfare outcomes this was the first time a national training and standardisation programme had been undertaken.

**Training of assessors**

During September and October 2013, farm assurance assessors from six commercial independent certification companies received a one-day training course in welfare outcome assessment of dairy cattle at one of eight geographical locations around the UK. The aims of the training day were not only to encourage standardisation of assessment of the welfare outcome measures, but also to contextualise these outcome observations within the assurance process. Two trainers, from a total pool of six trainers which included SM, KS and MC, were at each training day. The trainers had all been involved in developing the protocol and all had many years of experience in conducting welfare assessments on dairy farms. Two of the trainers were farm assurance assessors, three worked for an assurance scheme, and one was a welfare scientist. Following the on-farm training day, assessors were required to undertake online testing of mobility, body condition and cleanliness scoring, at approximately annual intervals. Online testing for HLS was not available at this time.

A 2-hour classroom session discussed the use of welfare outcomes and provided detailed training in each measure (see Table 1) with photos/ videos. /. Subsequently, on-farm practice of random sampling strategies and scoring each of the measures was covered in a further 90 minutes. A final 45-minute discussion on the causes and solutions to welfare problems then took place.

**Pre and post training scoring of photos and videos**

On arrival, participants were given the protocols and scored poster photos of 16 cows for body condition (thin/moderate/fat cows) and cleanliness (clean/dirty), poster photos of 20 cows for hair loss, lesions and swellings (each present/absent) and scored videos of 10 cows for mobility (not lame/lame/severely lame). Following the classroom and on-farm training sessions the assessors completed post-training scoring of on-screen photos of 20 cows for body condition, cleanliness, hair loss, lesions and swellings and videos of 15 cows for mobility.

**Online training tool**

Following the on-farm training day, assessors were given access to online training material and required to undertake online tests of mobility (not lame/lame/severely lame), body condition (thin/moderate/fat cows) and cleanliness (clean/dirty) scoring according to the protocol described above at approximately annual intervals for 3 years, although for technical reasons there was no cleanliness test for year 2. The training material included the protocol, practice tests and guidance tips. The tests consisted of a random selection of 10 videos (for mobility) or photos (body condition, cleanliness) from a large bank of material, where each score category was presented at least once within the test. The videos/photos from which test material was drawn all had unanimous agreement between a panel of at least three expert scorers. At the end of each test, feedback was provided to the assessor about the overall numbers of videos/photos scored in each category by them and by the gold standard panel. Feedback on individual videos/photos was not provided to prevent memorising these scores. To ‘pass’ the test, assessors were required to show agreement with the gold standard panel for at least 8 out of 10 videos/photos for two consecutive tests. The decision to set a rolling pass criterium was taken partly for practical reasons as it reduced the number of photos/videos required. The 80% agreement rate was considered by an industry group to be acceptable to farmers.

**Farm Assessments**

Once assessors were trained and ‘passed’ the online tests they were eligible to make observations as part of the standard farm assurance audit process. A sample of at least 10 cows per farm was selected at random, but proportionally according to group size, from the milking herd (*i.e.* not including dry cows and non-milking heifers). For example, if a farm of 100 milking cows kept them in two groups of 50 then 5 cows would be selected for assessment from each group. To achieve random selection a strategy of selecting every nth cow was used, where n=number of milking cows/10. Observations occurred throughout the year in the housing or at pasture. Lying cows were to be encouraged to rise for observation unless it would substantially compromise their welfare. Nevertheless, it emerged following discussions with a group of inspectors that in practice lying cows were rarely included in the sample, inspectors preferring to observe the next nearest standing cow. The same cow did not need to be observed for all measures as understanding the relationships between measures at the cow level was not required. The descriptions for each measure can be found in Table 1 and were agreed by an industry group, based on existing protocols for mobility (lameness) (AHDB Dairy 2020) and body condition (thin/fat cows) (Defra 2001).

Information about the farm was provided to the scheme by the farmer on stock numbers (cows in milk, dry cows, young stock, bulls), as well as through checklists where more than one box could be ticked for: predominant breed (Holstein Friesian, British Friesian, Ayrshire, Shorthorn, Jersey, Guernsey, Montbeliarde, Brown Swiss, Other); milking parlour/ system (herringbone, abreast, tandem, parallel, rotary, cowshed/byre, automatic milking unit/ robot, other); housing system (loose yards, cubicles, other); whether cows had access to a loafing areas (yes/no); and an estimate of the number of weeks the cows spend at grass per year (0-52).

**Data Analysis**

Data was collected on paper by inspectors operating for a range of independent auditing companies. The data was entered by the company into the Red Tractor database via an online portal. Guidance was provided to aid accurate data entry. RT inspections occurred at intervals of 18 months to ensure farms were audited in different seasons. Data for three years from October 2013 to September 2016 were included in the analysis. Unfortunately, technical issues affecting the validity of a full dataset for subsequent years has meant that later data could not be included in the analysis. Comparisons were made over each calendar year. Therefore, each year approximately two-thirds of RT farms were audited, however it was not expected that there was bias between the samples in each year. The data were manipulated in Microsoft Excel and analysed using SPSS Statistics 24 and R v4.0.2.

**Training data**

Each photo was independently assessed by a panel of three trainers. Those with scores that agreed were used for testing the assessors’ score against the ‘gold standard’ score. ‘Gold standard’ mobility scores of cows in the videos were previously determined by an industry mobility scoring panel. The agreement of each assessor with the gold standard score was determined, and the percentage agreement for each measure and test was calculated as this was previously suggested as the most useful way to determine inter-observer reliability for farm assurance welfare outcome assessments (Mullan et al. 2011). To analyse differences between assessors’ scores of percentage agreement with gold standards before and after training, and between their ratings for identical questions in the pre- and post- questionnaires, the Wilcoxon signed rank test was employed. Differences between male and female assessors in their questionnaire ratings and test agreements and between individual trainers and the other trainers were identified using the Mann-Witney U test. Correlations between the ratings of each of the five areas of confidence for each measure in the questionnaire, and the percentage agreement achieved by assessors for corresponding measures, were analysed using Spearman’s rho for (1) the pre-training questionnaire and test and (2) the post-training questionnaire and test. To maximise the data available, changes in the number of attempts to pass the online training tool between years were tested pair-wise using the Mann Witney U test.

**Farm assessment data**

The raw de-duplicated dataset consisted of 19,899 audits assessing 248,689 cows. This represents 6.8%, 6.4% and 6.5% of cows on audited farms in years 1, 2 and 3 respectively. After removal of audits without a recorded herd size and those failing validation checks (data points outside of range), 18,091 audits remained, comprising 182,341 assessed cows and 6,093, 6,472 and 5,526 audits each for years 1, 2, and 3 respectively. Welfare outcome assessment data was missing in 5% of this data, and 2% were missing management information.

The aim of our analysis was to understand the prevalence of each welfare outcome measure, how it changed over time, and the association of each measure with farm characteristics and management factors. For each welfare outcome measure, prevalence within the set of assessed cows was calculated, with the binomial test used to compute confidence intervals. Using survey-weighted logistic regression through the R ‘survey’ package, the analysis was extended to consider prevalence among the full adult herd at audited farms, as well as to compute significance of change of welfare assessment outcomes between years.

Following this, we performed a deeper investigation into welfare outcome change over time and its association to management factors by constructing a Bayesian model with the R package ‘brms’ (Burkner 2018). The model contained: (a) audit-level and farm-level random effects; (b) change over time and herd size modelled as general additive models (Wood 2006) with cubic spline fit; (c) seasonal variation modelled as a general additive model with periodic cubic spline fit; (d) the covariates dominant breed(s), housing type(s), milking type(s), loafing and grazing (as a yes/no dichotomous predictor) each modelled as a multi-membership model (Browne et al. 2001) with equal weight given to each category. Each use of ‘Other’ and each missing data point were assigned as independent categories, so that their influence was effectively imputed from the multi-membership model’s normal distribution estimated from the non-‘Other’ categories. This allowed us to use all records despite missing data and the ambiguous ‘Other’ designator, as multiple imputation was considered computationally infeasible. The model was fitted with Hamiltonian Monte-Carlo simulation on four chains; the chains were deemed to have converged if the Gelman-Rubin Rhat diagnostic was < 1.05 for all derived variables. For cleanliness this required 2,000 warmup and 2,000 sampling iterations, whereas for all other responses a default of 1,000 warmup and 1,000 sampling iterations were sufficient. Each model took two days of compute time to run – as a result, more complex modelling including incorporation of interactions was deemed infeasible.

To determine the significance of the change in the number of weeks of grazing year on year the Mann-Witney U test was used and the Chi2 test was employed to determine any change in the proportion of farms that did not provide grazing to their cows at any time during the year.

**Results**

**Training of assessors**

Ninety-eight assessors (73 male, 25 female) were trained in total with between 10 and 17 assessors attending each day. Two trainers completed four training days, one trainer did three days, two trainers did two days each and one trainer did one training day.

Prior to training the mean percentage agreement with the gold standard was between 59% (range 20-80%) for mobility (3-point scale) to 83% (range 45-100%) for presence/absence of swellings. Post training the percentage agreement was significantly higher for body condition (thin/moderate/fat) (p=0.001), cleanliness (p=0.003), swellings (p=0.003) and mobility using both a 2-point scale (p<0.001) and 3-point scale (p<0.001), however it was lower for hairless patches (p<0.001) and not significantly different for lesions (p=0.429) (see Figure 1). The proportion of assessors achieving at least 80% agreement with the gold standard, as would be required for the online testing tool, significantly increased following training for all measures except hairloss and lesions (see Figure 1).

Three approximately annual online webtool tests were conducted following the initial on-farm training for mobility, cleanliness and body condition. The number of assessors taking the tests reduced in the third year for administrative reasons. In most cases the assessors taking the online tests had completed the on-farm training, however some new assessors were trained subsequently, and others left their role and did not complete the test(s). The number of attempts required by assessors to pass the mobility tests reduced year on year and was significantly fewer in year 3 (mean 3.5) compared to year 1 (mean 4.9) (p=0.031). The number of attempts to pass the body condition tests increased from year 1 (mean 3.0) in the second year (mean 3.8) and reduced to the year 1 level in the third (mean 3.0) (although neither significantly). There were significantly fewer attempts to pass required from year 1 (mean 2.9) to year 3 (mean 2.2) for cleanliness (p=0.013) (see Figure 2). The proportion of assessors requiring more than 6 attempts to pass (at which point they were asked to seek additional training from their employer) was 19%, 18% and 11% for mobility and 5%, 11% and 6% for body condition for years 1,2 and 3 respectively and 6% and 0% for cleanliness for years 1 and 3 respectively. There was no significant effect of gender, or whether they attended the original (reported here) or subsequent welfare outcome training on the number of attempts it took to pass the online test for any measure in any year. In addition, there were no significant correlations between the final percentage agreement achieved following on-farm training and the number of attempts to pass the webtool tests for any measure in any year.

**Farm assessments**

Welfare outcome assessments were conducted during 93.2%, 93.2% and 93.9% of the 6738, 7083 and 6068 audits conducted in Years 1, 2 and 3 respectively. The prevalence of the welfare outcome measures within assessed cattle, and the estimated adult national herd prevalence in audited farms, is illustrated in Table 4 and Figure 3. There were significant improvements in all measures from year 1 to 2 (p=0.0143 for thin cows, all others p<0.0001). A further significant improvement was seen for HLS from year 2 to 3 (p=0.0001) but the improvement in thin cows reversed (p<0.001). Over the three-year study period, estimated adult herd lameness dropped from 10.0% to 7.9%, dirty cows from 12.4% to 9.2%, and HLS from 9.3% to 6.5%. Fat cows dropped modestly from 2.4% to 1.9%, while thin cows increased marginally from 3.3% to 3.8%.

Additional data collected by RT on farm characteristics is presented in Table 2 (stock numbers) and Table 3. The mean number of milking cows on each farm was 128, 135, 131 in years 1, 2 and 3 respectively. The most popular response in each farm category was a cubicle housing system (90% in year 3), parlour feeding (84% in year 3), Holstein Friesian breed (74% in year 3) and herringbone parlours (79% in year 3). Most farms provided outdoor loafing (60% in year 3). Figure 5 shows the farmer-estimated number of weeks cows spent at grass over the preceding 12 months. There were significantly fewer weeks at grass in year 1 compared to year 2 (p=0.020) and significantly more farms with 0 weeks at grass in year 2 compared to year 1 (p=0.002) (4.2%, 5.6% and 5.6% of farms not providing grazing access at all in years 1, 2 and 3 respectively (see Figure 5)- these farms had 7.1%, 10.3% and 9.6% of cows on them respectively).

The influence of certain farm characteristics on welfare outcomes is described in Table 5. The odds ratios are presented, and it can be seen, for example, that breed is frequently influential on the likelihood of being recorded with any of the outcome measures, with Holstein Friesians typically more likely to exhibit mobility, thinness, HLS and cleanliness issues, but less likely to be overweight. HLS was most influenced by housing characteristics, with provision of a loafing area, loose yards and grazing all being protective against being recorded with HLS. Loose yards and grazing were also protective against lameness and excess weight, but thin cows were more prevalent in loose yards.

As illustrated in Figure 4, the likelihood of observing each measure was found to vary during the year. Lameness was 1.5 times more likely to be observed in February compared to September, the rate of observation of thin cows in April was 1.3 times that of December, while fat cows were 1.5 times more incident in March than September. The greatest seasonal differences were found for dirty cows, which were five times as likely to be observed in June compared to December, and HLS with the rate in March being 3 times that in September.

The influence of herd size on the likelihood of observing each measure is also shown in Figure 4. Generally, the likelihood of observing thin or fat cows reduced as herd size increased. Much smaller and larger herds than average were less likely to have cows observed as lame or with HLS – this may also be the case for dirtiness but the results are inconclusive. It is also important to note that model uncertainty increases significantly at herd size extremes due to the small number of farms with these characteristics and their heterogeneity.

**Discussion**

This is the first national-scale reporting of improvement of welfare outcomes sustainability metrics following inclusion of welfare assessments in a private dairy farm assurance scheme, and occurred during a period of increased productivity. Average milk yield was 7,543 litres per cow in 2013 (mostly prior to implementation) and rose to 7,897 litres/cow in 2014 and 2015 in the UK (AHDB 2019). Although this dropped in 2016 to 7,559 litres/cow, likely due to low milk price leading to reduced concentrate feed use (FarmingUK 2016), this value was still higher than prior to implementation and a short-lived reduction as the increase returned in 2017 (7,893 litres/cow) and further in 2018 (7,959 litres/cow). It has long been suggested that when productivity is low improvements in animal welfare and productivity often go together, but when productivity is already high increasing it further is likely to be at the detriment of animal welfare (McInerney 1991). Here we show that sustainable intensification goals relating to animal welfare can be achieved in a country that ranked 12th in the world for productivity per cow in 2012 (Compassion in World Farming 2012). The data was collected on Red Tractor farms producing over 95% of all UK milk (Red Tractor 2015) and can be considered to represent the national herd. The necessary processes are now embedded to allow monitoring of trends in both resource provision and welfare outcomes over time. This study follows the publication of similar UK initiatives for laying hens (Mullan et al. 2016) and pigs (Pandolfi et al. 2017), both of which showed significant animal welfare improvements.

**Training of assessors**

To date this is the largest scale study reporting training and standardisation results of commercial farm assessors, and importantly, includes long-term follow-up for key measures. Following initial training, assessors improved their agreement with a gold standard and showed high levels of standardisation of lameness, body condition, swellings and cleanliness. A range of measures of standardisation have previously been used and after recently reviewing the impact of applying 10 different indices of inter-observer reliability on a goat outcome measure Giammarino et al considered Bandigwala’s B and Gwet’s Gamma (AC/1) to be the best approach (Giammarino et al. 2021). It should be noted that either 2 or 3 point scales were utilised to maximise repeatability as had been found previously (Thomsen & Baadsgaard 2006; Brenninkmeyer et al. 2007) and we also found that inter-observer reliability was greater for a 2-point lame/not lame than a 3-point mobility scale. It is not clear why assessment of hairloss and lesions did not improve similarly, although the initial agreement was greater than 75% for both measures. Mobility showed the poorest inter-observer reliability prior to training but, perhaps because assessors could be considered experienced rather than inexperienced observers, a relatively small amount of training improved their standardisation substantially (Main et al. 2000; March et al. 2007). Identifying suitable farms with a variation in scores for each measure for practice can be difficult, therefore complementary videos and photos are useful, and chosen for test purposes in this multi-site study, despite these having been previously shown to result in higher inter-observer reliability than observations of live animals (Mullan et al. 2011).

The significant reduction in the number of attempts required to pass the online webtool for mobility and cleanliness from year 1 to 3 likely demonstrates an increased familiarity with the scoring system through frequent use. However, it is not clear why the body condition scores should worsen and then improve to the original level again across the 3 years. Setting a threshold for the number of attempts before assistance should be sought (six) was instigated during the first test when it was found some assessors were taking a long time repeatedly trying to pass. Although it is not known how many assessors sought assistance, or what this help consisted of, it was welcomed by assessors as a practical way to limit their time taking the online tests. This type of on ongoing assessment is now integrated into the assessor monitoring process for the scheme. However, there is a substantial administrative burden to managing the online access and results of the webtool, and to generating new material and gold standard scores for additional tests. This is currently undertaken by a third party and fee charged for users.

The training was successful in improving the confidence of assessors in the various aspects of their work for almost all the welfare outcome measures involved, and importantly, eliminated a comparative lack of confidence reported by female assessors prior to the session. It could be expected that higher levels of confidence by assessors encourages reporting of poor welfare, both formally to the scheme and informally in discussions with a farmer, both of which are part of potential mechanisms for improving welfare.

**Farm assessments**

Welfare outcome assessments were reported for 93.4% of Red Tractor farm audits. Where outcome data was not available it was not known whether the assessment was not conducted, or the data not entered. Data from almost 182,341cows over 18,091 audits were included in the analysis providing a high level of confidence in the national farm characteristics and outcome prevalences reported.

From our most common responses a typical UK farm could be characterised at that time as having approximately 130 Holstein or British Friesian milking cows in cubicle housing in the winter, probably with a loafing area, and at pasture in summer. They would have a herringbone parlour in which feed is given and feed either a Total Mixed Ration (TMR) or silage when fresh grass is not grazed. However, there are difficulties in interpreting some of the data, and in particular, the predominant breed category, as more than one breed was frequently recorded without any further information about proportions within the herd.

There has been increasing interest in the welfare of dairy cows that do not have access to pasture in the UK, with new labelling initiatives such as ‘Pasture for life’ and ‘Free Range Dairy’ introduced to reflect perceived consumer concern (Free Range Dairy 2018; Pasture for Life 2018). Our results showed there was a significant increase in continuously housed farms over the three years reported to approximately 6% of farms, which are larger than average, having nearly 10% of all UK cows on them. A wide-ranging review of pasture access for cattle concluded that not only did cows show a preference for pasture access but there were a variety of other welfare benefits, including being protective against some negative welfare outcomes, as we also showed in this study (Arnott et al. 2017). In 2018 in the Netherlands, a collaborative approach between a wide range of actors started to reverse the 15-year trend for increasing use of continuous housing of dairy cattle (Runhaar 2021).

Welfare outcome measure prevalences for the UK have previously been reported in smaller research studies. In most instances the prevalences reported here are lower than those in research studies. For example, an industry body review of studies of UK lameness published between 1996 and 2014 reported a range of mean prevalences from 18% to 37% (CHAWG 2018) and a more recent study, where data was collected at the same time as this study reported a mean prevalence of 32% (Griffiths et al. 2018), whereas in this study assessors recorded 10% or fewer lame cows. There could be several reasons for this, an important one being that many of these studies were conducted during the winter housing period when we have shown lameness is highest, whereas data in this study was collected all year round. In addition, in this study, cattle were observed in the housing/pasture, rather than the more commonly used research method of during exiting the milking parlour, and only standing cattle were assessed. A small study by the authors was undertaken on 14 farms to try to ascertain whether location (by scoring the same cows exiting milking and in the housing/pasture) and only scoring standing cows (by noting whether cows were standing or lying at the time of assessment and comparing prevalences of these groups when scored exiting milking) could account for lower than expected reported lameness prevalence. In line with other studies (Ito et al. 2010; Solano et al. 2016; Hut et al. 2021; Ji et al. 2021) the lameness prevalence was lower for standing than lying cows (22% vs 29%) and only 3/14 farms had a lower lameness prevalence (and 6/14 had higher prevalence) when the same animals were assessed in housing/pasture compared with exiting milking (unpublished data). Therefore, it was concluded that other factors were also important, likely including assessor under-reporting. Despite the on-going training and assessment programme assessors may still under-score compared to a gold standard or researcher and assessors may also be unconsciously or consciously more likely to err on the side of caution when scoring an animal with a welfare issue. Nevertheless, the data reported here have been recorded consistently for three years and therefore year-on-year comparisons are valid.

The improvement in mobility, hairloss, lesions and swellings (HLS), and cleanliness in particular from year 1 to year 2 was significant and by year three there were approximately 36,000 fewer lame cows, 52,000 fewer dirty cows and 55,000 fewer cows with HLS in the UK than there would have been had the prevalences stayed as they were in year 1. This is a welcome achievement and our analysis indicates that this effect started to occur even before some farms were visited for their first outcome assessment, suggesting that the process of implementation of the outcome monitoring within the farm assurance scheme was responsible for some of the improvement. A strong communication strategy was employed over a period of approximately two years to signal to farmers that changes to welfare assessment would occur. However, the challenge remains to ensure that initial improvements are at least maintained and ideally continued. It is expected that additional strategies will need to be employed to achieve further improvements in welfare outcome measures.

The reasons for there being little change in the prevalence of fat cows and an increase in thin cows in year 3 are not clear. It has been suggested in industry discussions that the prevalence of thin cows is influenced by milk and feed prices, however further analysis would be needed to confirm this. These data provide an important guide for industry to ensure the most effective use of their resources to improve welfare and interventions to reduce the proportion of thin cows in the national herd should be prioritised.

In line with findings of a review of 23 studies (Robbins et al. 2016) we found a curvilinear relationship between lameness and herd size, with lower levels being recorded on smaller and larger farms, and similar relationships with hairloss, lesions and swellings and dirtiness. The reasons for this are not clear, but may be related to individual care being more possible in smaller farms and very large farms being able to efficiently put in place robust monitoring and treatment systems. Larger farms in our study also appeared better able to manage body condition, with fewer thin or fat cows, something not found by Adams et al (2017) in their study of 181 dairy farms across the USA.

The outcome measures showed substantial seasonal variation, most notably HLS and cleanliness, with the worst prevalences recorded around the end of the winter housing period. This has implications for discussions with farmers and the implementation of thresholds for actions to be undertaken if outcome results are high. If thresholds were based upon fractions of the data, for example, the worst 25% of farms, it may be that two or more thresholds would be needed throughout the year to reflect the seasonal variation found. In addition, the seasonal variation, as well as the small farm sample size, would make it difficult to compare an individual farm’s audit result with the previous one if they were conducted at 18-month intervals. However, since we model this variation in our Bayesian model, it could be employed to make farm and time-specific audit predictions that the subsequent audit results could be compared against, essentially creating a personalised threshold for each audit. As demonstrated in Table 5 and Figure 4, since the Bayesian model is fully explainable, this would go some way to demonstrate a fair and transparent approach to determining which farms require remedial action.

**Conclusion**

This national monitoring of UK dairy cow welfare outcomes through a private farm assurance scheme demonstrates sustainable intensification occurred during 2013-16. Our data has shown significant improvements in lameness, cleanliness and hairloss, lesions and swellings (HLS) over three years, during a time when productivity per cow increased. The training programme for 98 assessors improved standardisation in outcome assessments and follow-up using online webtool training and testing over three years demonstrated further improvements in standardisation. Our models indicated that the improvements in welfare outcomes started to occur even before the first outcome audit on many farms, suggesting that the activities prior to, and the process of, implementation were effective at improving welfare. In order to maximise further improvements it is recommended that an evidence-based industry wide welfare improvement plan is instigated.

**Acknowledgements**

We thank the farm assessors who collected the data for this study, in all weather.

**Funding details**

This work was funded by the Tubney Charitable Trust through the AssureWel project under which this work was conducted.

**Disclosure statement**

No authors have received any financial interest or benefit that has arisen from direct result of this research.

**Data availability**

The welfare outcome data of dairy cattle presented in this paper is owned by the Red Tractor farm assurance scheme as part of Assured Food Standards and is commercially sensitive and therefore not publicly available. Requests for access should be directed in the first instance to the corresponding author.

**Author contributions**

SM helped design the implementation programme, conducted training of assessors, undertook analysis and wrote the manuscript. PW helped design the implementation programme and facilitated data extraction. KC facilitated data extraction and contributed to the manuscript. DCJM led the implementation programme design and contributed to the analysis. KS helped design the implementation programme and conducted training of assessors. MC helped design the implementation programme, led the training of assessors and undertook some analysis. AWD undertook the modelling and contributed to the writing of the manuscript.

**References**

Adams AE, Lombard JE, Fossler CP, Roman-Muniz IN, Kopral CA. 2017. Associations between housing and management practices and the prevalence of lameness, hock lesions, and thin cows on US dairy operations. Journal of Dairy Science. 100(3):2119-2136.

AHDB. 2019. UK milk yield. AHDB Dairy; [accessed 2020 14 August]. https://ahdb.org.uk/dairy/uk-milk-yield.

AHDB Dairy. 2020. Mobility scoring: how to score your cows. AHDB Dairy; [accessed 2020 17 January]. https://ahdb.org.uk/knowledge-library/mobility-scoring-how-to-score-your-cows.

Arnott G, Ferris CP, O'Connell NE. 2017. Review: welfare of dairy cows in continuously housed and pasture-based production systems. Animal. 11(2):261-273.

Griffiths B, Grove White D, Oikonomou G. 2018. A Cross-Sectional Study Into the Prevalence of Dairy Cattle Lameness and Associated Herd-Level Risk Factors in England and Wales. Frontiers in Veterinary Science.

Brenninkmeyer C, Dippel S, March S, Brinkmann J, Winckler C, Knierim U. 2007. Reliability of a subjective lameness scoring system for dairy cows. Animal Welfare. 16(2):127-129.

Browne WJ, Goldstein H, Rasbash J. 2001. Multiple membership multiple classification (MMMC) models. Statistical Modelling. 1(2):103-124.

Buller H, Blokhuis H, Jensen P, Keeling L. 2018. Towards Farm Animal Welfare and Sustainability. Animals. 8(6).

Burkner PC. 2018. Advanced Bayesian Multilevel Modeling with the R Package brms. R Journal. 10(1):395-411.

CHAWG. 2010. Dairy Cow Welfare Strategy. UK.

CHAWG. 2018. Fourth Report. UK.

Compassion in World Farming. 2012. Statistics: Dairy Cows. UK.

Defra. 2001. Condition Scoring of Dairy Cows. London, UK: Defra Publications.

Dillon EJ, Hennessy T, Buckley C, Donnellan T, Hanrahan K, Moran B, Ryan M. 2016. Measuring progress in agricultural sustainability to support policy-making. International Journal of Agricultural Sustainability. 14(1):31-44.

FarmingUK. 2016. Smaller herd and reduced milk yield reasons for decrease in British milk production. [accessed 2020 11th September]. https://www.farminguk.com/news/smaller-herd-and-reduced-milk-yield-reasons-for-decrease-in-british-milk-production\_44284.html.

FAWC. 2009. Opinion on the Welfare of the Dairy Cows. London, UK.

Free Range Dairy. 2018. Who we are. Free Range Dairy; [accessed 2018 15 June]. http://www.freerangedairy.org/who-we-are/.

Giammarino M, Mattiello S, Battini M, Quatto P, Battaglini LM, Vieira ACL, Stilwell G, Renna M. 2021. Evaluation of Inter-Observer Reliability of Animal Welfare Indicators: Which Is the Best Index to Use? Animals. 11(5).

Godfray HCJ. 2015. The debate over sustainable intensification. Food Security. 7(2):199-208.

Hut PR, Hostens MM, Beijaard MJ, van Eerdenburg F, Hulsen J, Hooijer GA, Stassen EN, Nielen M. 2021. Associations between body condition score, locomotion score, and sensor-based time budgets of dairy cattle during the dry period and early lactation. Journal of Dairy Science. 104(4):4746-4763.

Ito K, von Keyserlingk MAG, LeBlanc SJ, Weary DM. 2010. Lying behavior as an indicator of lameness in dairy cows. Journal of Dairy Science. 93(8):3553-3560.

Ji KJ, Booth RE, Blackie N. 2021. A Retrospective Case Study into the Effect of Hoof Lesions on the Lying Behaviour of Holstein-Friesian in a Loose-Housed System. Animals. 11(4).

Main DCJ, Clegg J, Spatz A, Green LE. 2000. Repeatability of a lameness scoring system for finishing pigs. Veterinary Record. 147(20):574-576.

Main DCJ, Mullan S, Atkinson C, Bond A, Cooper M, Fraser A, Browne WJ. 2012. Welfare outcomes assessment in laying hen farm assurance schemes. Animal Welfare. 21(3):389-396.

Main DCJ, Rogerson I, Crawley MC, Avizenius J, Fraser A, Mullan S. 2012. Welfare outcomes assessment in dairy farm assurance schemes. Cattle Practice. 20:142-145.

March S, Brinkmann J, Winkler C. 2007. Effect of training on the inter-observer reliability of lameness scoring in dairy cattle. Animal Welfare. 16(2):131-133.

McInerney J. 1991. A SOCIOECONOMIC PERSPECTIVE ON ANIMAL-WELFARE. Outlook on Agriculture. 20(1):51-56.

Mullan S, Edwards SA, Butterworth A, Ward M, Whay HR, Main DCJ. 2011. Welfare science into practice: a successful case example of working with industry. Animal Welfare. 20(4):597-601.

Mullan S, Edwards SA, Butterworth A, Whay HR, Main DCJ. 2011. Inter-observer reliability testing of pig welfare outcome measures proposed for inclusion within farm assurance schemes. Veterinary journal (London, England : 1997). 190(2):e100-109.

Mullan S, Szmaragd C, Cooper MD, Wrathall JHM, Jamieson J, Bond A, Atkinson C, Main DCJ. 2016. Animal welfare initiatives improve feather cover of cage-free laying hens in the UK. Animal Welfare. 25(2):243-253.

Pandolfi F, Stoddart K, Wainwright N, Kyriazakis I, Edwards SA. 2017. The "Real Welfare' scheme: benchmarking welfare outcomes for commercially farmed pigs. Animal. 11(10):1816-1824.

Pasture for Life. 2018. The best animal welfare. The Pasture-fed Livestock Association; [accessed 2018 15 June]. https://www.pastureforlife.org/why-pasture/the-best-animal-welfare/.

Red Tractor. 2013. Red Tractor Annual review. London, UK.

Red Tractor. 2014. Red Tractor Annual Review. London, UK.

Red Tractor. 2015. Red Tractor Annual Review. London, UK.

Red Tractor. 2017. Red Tractor Assurance for Farms- Dairy Standards. London, UK.

Robbins JA, von Keyserlingk MAG, Fraser D, Weary DM. 2016. INVITED REVIEW: Farm size and animal welfare. Journal of Animal Science. 94(12):5439-5455.

Runhaar H. 2021. Four critical conditions for agroecological transitions in Europe. International Journal of Agricultural Sustainability. 19(3-4):227-233.

Solano L, Barkema HW, Pajor EA, Mason S, LeBlanc SJ, Nash CGR, Haley DB, Pellerin D, Rushen J, de Passille AM et al. 2016. Associations between lying behavior and lameness in Canadian Holstein-Friesian cows housed in freestall barns. Journal of Dairy Science. 99(3):2086-2101.

Thomsen PT, Baadsgaard NP. 2006. Intra- and inter-observer agreement of a protocol for clinical examination of dairy cows [Article]. Preventive Veterinary Medicine. 75(1-2):133-139. English.

Whitfield S, Marshall A. 2017. Defining and delivering "sustainable' agriculture in the UK after Brexit: interdisciplinary lessons from experiences of agricultural reform. International Journal of Agricultural Sustainability. 15(5):501-513.

Wood S. 2006. Generalized Additive Models. An Introduction with R. New York: Chapman and Hall.

Table 1: Protocol for welfare outcome measures assessed on 10 cows and associated database entry



Table 2: Stock numbers on UK dairy farms (2013/14-2015/16)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Year 1 | | | Year 2 | | | Year 3 | | |
|  | 6738 farm audits | | | 7083 farm audits | | | 6068 farm audits | | |
| Stock numbers | n farms | % farms | mean (range) | n farms | % farms | mean (range) | n farms | % farms | mean (range) |
| Cows in milk | 6328 | 93.9% | 128 (0-2600) | 7023 | 99.2% | 135 (0-4110) | 6008 | 99.0% | 131 (0-1863) |
| Dry cows | 6284 | 93.3% | 21 (0-550) | 6994 | 98.7% | 23 (0-1400) | 5963 | 98.3% | 23 (0-1200) |
| Youngstock | 6131 | 91.0% | 97 (0-3028) | 6848 | 96.7% | 104 (0-1800) | 5814 | 95.8% | 106 (0-3500) |
| Stock bulls | 5333 | 79.1% | 2 (0-25) | 6242 | 88.1% | 2 (0-27) | 5235 | 86.3% | 2 (0-30) |

Table 3: Farm characteristics of UK dairy farms (2013/14-2015/16)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Year 1 | | Year 2 | | Year 3 | |
|  | n farms | % | n farms | % | n farms | % |
| Housing system (more than one answer can be given) | | | | | | |
| Response rate | 6559 | 97.3% | 6895 | 97.3% | 5935 | 97.8% |
| cubicles | 5935 | 90.5% | 6211 | 90.1% | 5322 | 89.7% |
| loose yards | 3796 | 57.9% | 3648 | 53.1% | 3261 | 54.5% |
| other | 40 | 0.6% | 852 | 12.4% | 747 | 12.5% |
| Outdoor loafing | | | | | | |
| Response rate | 4374 | 64.9% | 5069 | 71.6% | 4037 | 66.5% |
| yes | 2660 | 60.8% | 2896 | 57.1% | 2404 | 59.5% |
| no | 1714 | 39.2% | 2173 | 42.9% | 1633 | 40.5% |
| Feeding system (more than one answer can be given) | | | | | | |
| Response rate | 6575 | 97.6% | 6878 | 97.7% | 5953 | 97.4% |
| parlour feeding | 5264 | 80.1% | 5557 | 80.8% | 5023 | 84.4% |
| semi or total tmr | 2560 | 38.9% | 3173 | 46.1% | 2664 | 44.8% |
| self / easy feed silage | 1753 | 26.7% | 1816 | 26.4% | 1837 | 30.9% |
| forage in trough / passageway(not mixed) | 857 | 13.0% | 1379 | 20.0% | 1815 | 30.5% |
| out of parlour feeders | 587 | 8.9% | 492 | 7.2% | 459 | 7.7% |
| midday feed | 61 | 0.9% | 62 | 0.9% | 49 | 0.8% |
| other | 428 | 6.5% | 829 | 12.1% | 689 | 11.6% |
| Predominant breed (more than one answer can be given) | | | | | | |
| Response rate | 6584 | 97.7% | 6884 | 97.8% | 6023 | 98.6% |
| Holstein Friesian | 3212 | 48.8% | 3987 | 57.9% | 4444 | 73.8% |
| British Friesian | 1157 | 17.6% | 1667 | 24.2% | 1714 | 28.5% |
| Jersey | 529 | 8.0% | 618 | 9.0% | 522 | 8.7% |
| Ayrshire | 307 | 4.7% | 382 | 5.5% | 297 | 4.9% |
| Montbeliarde | 210 | 3.2% | 270 | 3.9% | 248 | 4.1% |
| Brown Swiss | 176 | 2.7% | 206 | 3.0% | 153 | 2.5% |
| Shorthorn | 146 | 2.2% | 173 | 2.5% | 136 | 2.3% |
| Guernsey | 66 | 1.0% | 62 | 0.9% | 45 | 0.7% |
| other | 710 | 10.8% | 1176 | 17.1% | 1013 | 16.8% |
| Parlour type (more than one answer can be given) | | | | | | |
| Response rate | 6586 | 97.7% | 6912 | 98.2% | 6029 | 98.7% |
| herringbone | 5269 | 80.0% | 5492 | 79.5% | 4771 | 79.1% |
| abreast | 793 | 12.0% | 779 | 11.3% | 628 | 10.4% |
| Automatic Milking Unit (robot) | 213 | 3.2% | 251 | 3.6% | 268 | 4.4% |
| rotary | 114 | 1.7% | 158 | 2.3% | 136 | 2.3% |
| cowshed/ byre | 93 | 1.4% | 149 | 2.2% | 175 | 2.9% |
| tandem | 83 | 1.3% | 93 | 1.3% | 87 | 1.4% |
| parallel | 6 | 0.1% | 24 | 0.3% | 33 | 0.5% |
| other | 335 | 5.1% | 797 | 11.5% | 679 | 11.3% |

Table 4: The means amd confidence intervals of the welfare outcome measure prevalences for each of the three years in the study (2013/14- 2015/16)



Table 5: The means and credible intervals of the odds ratio of a cow being recorded as affected by welfare outcome measures for farm characteristics compared with a comparator.



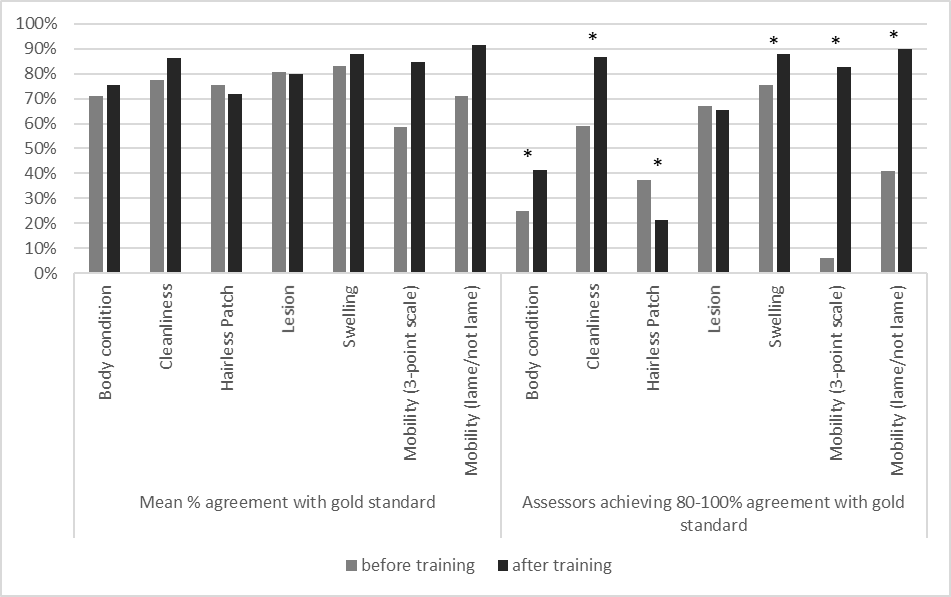


Figure 1: The mean percentage agreement and proportion of assessors achieving 80-100% agreement with a gold standard for body condition (thin/moderate/fat cows), cleanliness (clean/dirty), hairless patch, lesion, swelling (all present/absent) and mobility (using a 3-point scale (not lame/lame/severely lame) and also combining lame categories to give a binary score of not lame/lame). \*significant difference (P<0.05)

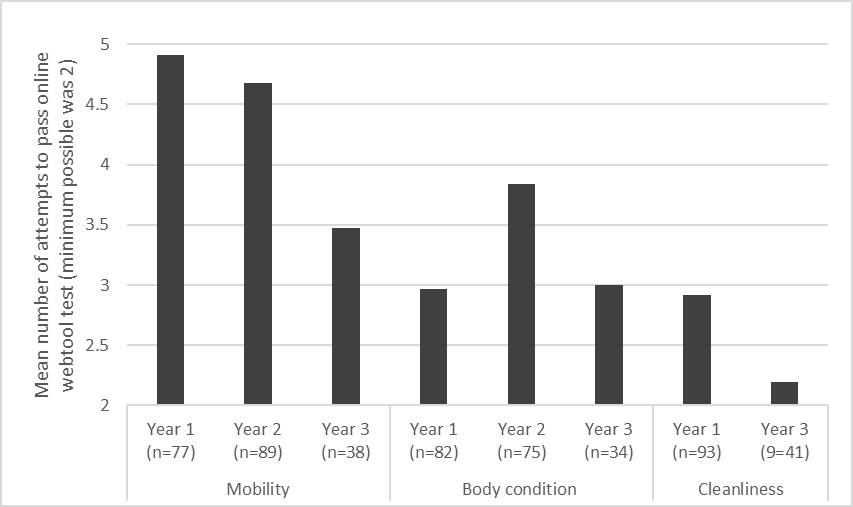


Figure 2: The mean number of attempts assessors took to pass an online standardisation test where they were required to achieve 8/10 agreement with a gold standard twice in a row for mobility (3-point scale, not lame/lame/severely lame), body condition (thin/moderate/fat) and cleanliness (clean/dirty). The minimum number of attempts possible was 2. \* significant difference (p<0.05)

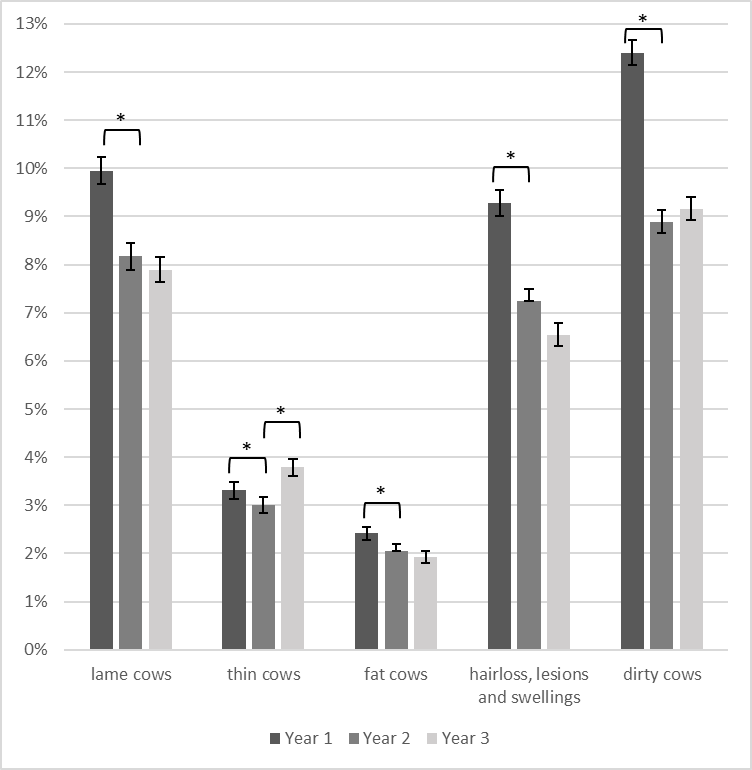


Figure 3: Prevalences of welfare outcome assessments conducted on a sample of 10 cows on UK dairy farms from October 2013-September 2016. \*significant difference (p<0.05)

Figure 4: Odds ratios of the likelihood of a cow being recorded as affected by welfare outcome measures across the years of assessment, and with the changing seasons during the course of any one year and in relation to herd size.

SEE FIGURE ON FINAL PAGE OF SUBMISSION

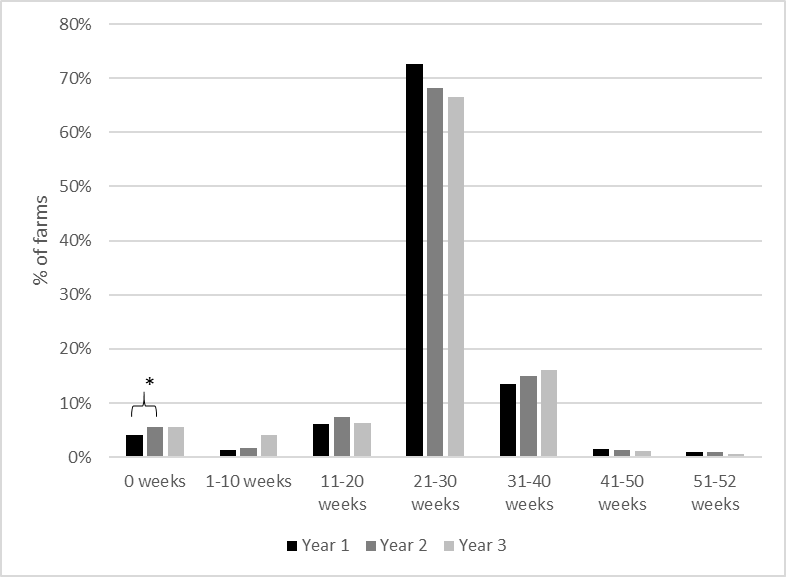


Figure 5: The farmer-estimated number of weeks cows spent at grass over the preceding 12 months. \*significant difference (p<0.05)

Suppplementary material

**Pre and post training questionnaire**

**Materials and Methods**

Prior to attending the training day, assessors had been sent a pre-training questionnaire and were given time on arrival to complete it if they hadn’t already done so. The pre-training questionnaire asked assessors to:

1. Estimate the proportion of the audit they spend looking at livestock, buildings/parlour and records/health plan, and indicate which of these aspects they think the farmer prefers
2. Offer their three biggest welfare issues for UK dairy cattle
3. Rate their confidence in raising a welfare problem during an audit on a scale from 1-10, and then suggest anything that would help improve their confidence
4. Rate their confidence on a scale from 1-10 for five areas:
   1. identifying affected cows
   2. understanding the causes of the problem
   3. directing to sources of possible advice
   4. knowing related standards
   5. raising a non-compliance, for each of six welfare outcome measures: (a) mobility, (b) body condition, (c) swellings, (d) skin lesions, (e) hairless patches, (f) cleanliness
5. Rate on a scale from 1-10 how beneficial they expect the inclusion of the welfare outcome assessments into the assurance audit to be for farmers, assessors and the assurance scheme
6. Estimate how many minutes the welfare outcome assessment would take in the audit

The post training questionnaire was completed at the end of the training day and contained identical questions to the pre-training questionnaire for all but section (1) described above. In addition, it contained open questions about the most and least useful aspects of the training course and areas for improvement in the training, and asked assessors to identify which of four methods they would prefer further training to be delivered to them.

Results

The pre-training questionnaire was responded to by 80 assessors (minimum response for a question was 63 assessors) and the post-training questionnaire by 96 assessors (minimum response for a closed question was 88 assessors and for an open question was 34 assessors). Prior to training the mean proportion of time estimated by assessors spent during an annual audit for records was 47% (range 25-85%), for buildings 28% (range 10-50%), and for livestock 23% (range 4-50%), although 90% of assessors suggested that farmers prefer the livestock element of the audit. Building inspection and records were suggested to be preferred by 42% and 4% of assessors respectively, whilst 19% of assessors reported farmers had no preference (assessors could tick more than one answer). The welfare issues most likely to be included in the list of the top three facing UK dairy cattle were lameness, mastitis and fertility, suggested by 91%, 58% and 21% of assessors before training and by 97%, 63% and 15% of assessors after training respectively. Other welfare issues suggested by more than 10% of assessors in either questionnaire were: tuberculosis, poor body condition, incorrect nutrition, disease and poor housing.

Prior to training, the mean confidence scores assigned by assessors for their ability to identify the causes of, direct to sources of advice about, understand the Red Tractor standards related to welfare outcomes, and raise a non-compliance for each (mobility, body condition, hairless patch, lesion, swelling, cleanliness) ranged between 6.3 and 8.3 out of 10. Following training, the mean confidence scores ranged from 7.6 to 9.1 and confidence scores were significantly higher (p<0.05) for all 5 categories of all 6 measures apart from understanding the causes of dirt (mean pre- and post-training ratings of 8.2 vs 8.4, p=0.057). Female assessors reported significantly lower confidence scores prior to training for identifying lame cows (mean 7.7 vs 8.2, p=0.004), thin/fat cows (mean 7.3 vs 8.0, p=0.022), and swellings (mean 7.3 vs 8.1, p=0.019), and directing to sources of advice to reduce lameness (mean 5.3 vs 7.0, p=0.048) and thin/fat cows (5.5 vs 7.0, p=0.042). However, following training there was no significant difference in confidence between male and female assessors for any category or welfare measure.

The assessors’ estimate of the duration of welfare outcome assessment was significantly lower post training (mean 27.7 minutes) compared to pre-training (mean 32.0 minutes) (p=0.007). The benefits of including welfare outcomes were rated significantly higher for farmers post training (mean 7.6 vs 6.6/10 pre-, p=0.004) and Red Tractor (8.4 vs 7.5/10 pre-, p=0.009) with no significant difference in benefits to assessors (post 6.1 vs 5.9/10 pre-, p=0.360).

Following training, assessors were asked to rank the most useful parts of the training. In practice some assessors ranked and others rated the 5 options, hence quantitative evaluation is not possible. Nevertheless, qualitatively it was clear that on-farm practical assessments in groups and pairs were felt to be the most beneficial by far, compared to the presentations, tests and explanations of the measures. Future support and training was most commonly thought to be useful if it was in the form of laminated on-farm documents, then an online training tool, DVD of examples and one-to-one on-farm training, (by 57, 52, 44 and 34 assessors respectively).