Development and validation of a visual body condition scoring system for dairy goats with picture-based training

A. Vieira,*¹ S. Brandão,†† A. Monteiro,§ I. Ajuda,* and G. Stilwell*

*Animal Behaviour and Welfare Research Laboratory, Centre for Interdisciplinary Research in Animal Health, Faculdade de Medicina Veterinária, Universidade de Lisboa, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal

†Electrical and Computers Engineering, Instituto Superior Técnico, Universidade de Lisboa, Avenida Rovisco Pais, 1049-001 Lisboa, Portugal

‡Electrical and Computers Engineering, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213

§Escola Superior Agrária de Viseu, Instituto Politécnico de Viseu, Quinta da Alagôa, Estrada de Nelas, 3500-606 Viseu, Portugal

ABSTRACT

Body condition scoring (BCS) is the most widely used method to assess changes in body fat reserves, which reflects its high potential to be included in on-farm welfare assessment protocols. Currently used scoring systems in dairy goats require animal restraint for body palpation. In this study, the Animal Welfare Indicators project (AWIN) proposes to overcome this constraint by developing a scoring system based only on visual assessment. The AWIN visual body condition scoring system highlights representative animals from 3 categories: very thin, normal, and very fat, and was built from data sets with photographs of animals scored by a commonly used 6-point scoring system that requires palpation in 2 anatomical regions. Development of the AWIN scoring system required 3 steps: (1) identification and validation of a body region of interest; (2) sketching the region from photographs; and (3) creation of training material. The scoring system's reliability was statistically confirmed. An initial study identified features in the rump region from which we could compute a set of body measurements (i.e., measures based on anatomical references of the rump region) that showed a strong correlation with the assigned BCS. To validate the result, we collected a final data set from 171 goats. To account for variability in animal size and camera position, we mapped a subset of features to a standard template and aligned all the rump images before computing the body measurements. Scientific illustrations were created from the aligned images of animals identified as representative of each category to increase clarity and reproducibility. For training material, we created sketches representing the threshold between consecutive categories. Finally, we conducted 2 field reliability studies. In the first test,

no training was given to 4 observers, whereas in the second, training using the threshold images was delivered to the same observers. In the first experiment, interobserver results was substantial, showing that the visual scoring system is clear and unambiguous. Moreover, results improved after training, reaching almost perfect agreement for the very fat category. The visual body condition scoring system is not only a practical tool for BCS in dairy goats but also shows potential to be fully automated, which would enhance its use in welfare assessment schemes and farm management.

Key words: body condition scoring, dairy goat, visual body condition scoring system, representative image, template alignment

INTRODUCTION

Goats mobilize body fat reserves according to their nutritional status, physiological needs, and availability of adipose tissue (Morand-Fehr, 2005). Fat mobilization leads to metabolic changes that may have a substantial effect on the animal's health, welfare, and production (Caroprese et al., 2009; Roche et al., 2009). Different methods exist to monitor changes in body fat reserves but BCS is the most widespread method across species, being considered simple and repeatable (Ferguson et al., 2006).

For cattle, sheep and goats, the scoring systems most commonly used to assess BCS are numerical rating scales with 5-point (Edmonson et al., 1989; Ferguson et al., 1994), 6-point (Russel et al., 1969; Lowman et al., 1976; Hervieu and Morand-Fehr, 1999; Harwood, 2006), or 8-point (Wildman et al., 1982) scales. These scoring systems are usually divided into intermediate scores (0.25 or 0.5) that result in 13- to 21-point scoring systems. For this reason, observers need considerable training and experience. Moreover, all of these methods are subjective as they rely on the observer's assessment, which is generally based on visual appraisal or palpation of anatomic locations that may demand

Received February 5, 2015.

Accepted May 28, 2015.

¹Corresponding author: ana.lopesvieira@gmail.com

animal restraint, which is time consuming and not easily achieved on large farms (Edmonson et al., 1989; Halachmi et al., 2008).

In goats, the most widespread method is that first published in a preliminary form (Santucci and Maestrini, 1985) and later fully presented by Hervieu and Morand-Fehr (1999). It is a 6-point scoring system with intermediate scores (0.25) that requires palpation of 2 anatomical regions: the sternum and the lumbar vertebrae. The validity and interobserver reliability of this scoring system were confirmed by Santucci et al. (1991).

Body condition score is considered a valid, reliable, and feasible welfare indicator with high potential to be included in on-farm welfare protocols (Winckler et al., 2003; Botreau et al., 2009). One of the main challenges in welfare assessment is to develop scientifically sound on-farm protocols focusing on animal-based indicators (such as BCS) that can be implemented in an acceptable amount of time (Knierim and Winckler, 2009).

The inclusion of traditional BCS systems in on-farm welfare protocol for goats presents 2 challenges: (1) very detailed scoring systems require very experienced assessors; and (2) body palpation requires restraint of goats.

The first challenge has been addressed in other on-farm protocols in goats (Anzuino et al., 2010), as well as in other species (Welfare Quality, 2009a,b), by limiting the scoring systems to 2 or 3 levels; for example, very thin, normal, and very fat. Because our main objective was to ensure feasibility when using BCS as a welfare indicator, this simplification was also selected as our priority.

However, the second challenge remained unaddressed. Generally, BCS levels are qualitative scoring systems built with verbal descriptions of how observers should assess visually and palpate different body regions and anatomical features. Most of the time, these verbal descriptors are difficult to elaborate and even more difficult to understand. Recently, different studies in dairy cows and buffalo species have shown that BCS can be obtained from images (Ferguson et al., 2006; Bewley et al., 2008; Negretti et al., 2008). We believe that visual scoring systems that provide a realistic contextual representation of the different levels should be validated as a reliable and replicable method for BCS.

Thus, visual scoring systems should convey the appearance of a representative animal from each category. The definition of a representative for a given category to classify subsequent sets of new images is a common research topic in computer vision. Approaches focus either on the statistical distribution of features (Tu and Yuille, 2004); for example, identifying mean/median images, or on the identification of threshold features;

for example, images that are on the transition between 2 different categories (Burges, 1998). We believe that assessors, after becoming acquainted with the spectrum of possible levels and thresholds, can easily identify the category of a new animal by comparing it with the representative category.

Thus, this study had 3 objectives. First, to provide a systematic approach to the development of a visual body condition scoring system for adult dairy goats; second, to develop a training program using the new scoring system based on the concept of threshold images; and finally, to assess the interobserver reliability under field conditions.

MATERIALS AND METHODS

The general flow for the development of a visual body condition scoring system is presented in Figure 1. The study was organized in 2 stages. The first was the preliminary scoring system development study, and it took place between November and December 2012. The second stage, the final study for scoring system development, occurred between February 2013 and July 2014. Data were analyzed using the base packages of the R statistical language (R Core Team, 2013). All the image-processing tasks were performed in Matlab (The MathWorks Inc., 2012).

Preliminary Study for Scoring System Development

Data Acquisition. We used 32 goats [20 Saanen, 5 Alpine, and 7 crossbreed (Saanen $\times$ Alpine)] from an intensive dairy goat farm to select the body region that would give us the best representation of BCS levels. The goats were housed in pens with straw litter and had occasional access to exterior exercise areas. Diet was composed of a TMR distributed twice a day, and goats were milked twice a day in an automatic milking parlor (average of 2.0 L per goat per day). Kids were separated from their mothers after birth.

Two experienced assessors independently scored the 32 goats using the BCS method developed by Hervieu and Morand-Fehr (1999). If there was any discrepancy, the final score was decided by consensus. Only adult goats were scored, independently of milk yield and days in milk. The BCS median was 3, with a range of 1.50 to 5.00.

Using a camera (Nikon D60; Nikon, Tokyo, Japan) with automatic exposure adjustment and focus, we photographed the front (sternum region) and the back (rump region) of each goat, as these are the anatomical regions where palpation is performed for BCS evaluation (Hervieu and Morand-Fehr, 1999). Environmental conditions (e.g., light and background) and distance

Figure 1. Outline of development of visual body condition scoring system.

to the goats were always similar. For each photograph, an object (e.g., piece of tape) of consistent height and width was placed on the animal to act as a frame of reference.

Body Measurements. For image analysis, we performed a total of 4 and 8 measurements (in mm) of the front and rump regions, respectively (Figure 2). For the front view, we assessed the sternum area (a1), sternum perimeter (p1), smaller width at the level of the fifth rib (w1), larger width at the level of the scapula (w2). In the rump region, we measured the width between tuber sacrale (hip or hook bones; w3), smaller width of the rump region (w4), width between tuber ischia (pin bones; w5), length of the rump region (l1), rump region area (a2), rump region perimeter (p2), lateral virtual area (a3) that corresponds to the negative or positive area that is limited by the line that unites the hipbone and the ipsilateral thurl, lateral virtual perimeter (p3) that corresponds to the line surrounding a3. Digital image analysis was always performed by the same operator, using ImageJ software (ImageJ 1.47t, <http://imagej.nih.gov/ij/>).

Statistical Analysis and Results. We used a stepwise linear regression method with the BCS value assigned by the Hervieu and Morand-Fehr (1999) method as the dependent variable and body measurements as the independent variables. Front and rump view images were analyzed separately.

The stepwise regression of the rump region view had a higher adjusted R^2 , indicating the region whose variables better explained the relationship between BCS and the measurements performed (Table 1). The lateral virtual area (a3) showed the highest significance (Table 1).

Final Study for Scoring System Development

Data Acquisition. We increased the sample to 171 goats, again with a BCS range of 1.50 to 5.00, and 3 as the median. The animals were from 4 different dairy farms and included the following breeds: Saanen ($n = 73$), Alpine ($n = 24$), Murciana ($n = 41$), and cross-breed (Saanen $\times$ Alpine; $n = 33$). The environmental conditions, as well as the BCS assessment method, were the same as described for the preliminary study. Following the results of the preliminary study, we only photographed and used the rump region measurements. We followed the same image collection procedures.

Unbiased Body Measurements and Shape Alignment. To correct the measurements according to different body sizes and camera positioning, we aligned the rump features to a template. The template fixed a posture and scale for the animal, so that the hip and pin bones (the 4 reference features) lay in the same position across photographs, as illustrated in Figure 3. The corrected body measurements were computed using the aligned features.

To align all photographs, we assumed that the goats were motionless and their rear legs were aligned perpendicularly to the body axis, which is a good description of the actual image acquisition conditions. Under these conditions, the 4 reference features—the hip and pin bones—lay in the same plane and we could use a projective transformation to align all the anatomical references to the same template. By assuming that the whole rump was in that plane, which is a valid assumption for an observer far from the rump, we could apply the same transformation to the whole rump, including the remaining features. To determine the transforma-

Table 1. Summary of the stepwise linear regression model applied to the front and rump view measurements (in millimeters)¹

Body measurement ²	Coefficient				Adjusted R ²	Residual SE (df)
	Estimate	SE	t-value	Significance		
Front view					0.39	0.81 (29)
Intercept	4.136	1.397	2.961	0.00606**		
a1	7.387×10^{-7}	2.313×10^{-7}	3.194	0.00337**		
p1	-1.454×10^{-5}	7.812×10^{-6}	-1.861	0.07297	0.88	0.36 (28)
Rump view (Intercept)	5.485	2.204	2.489	0.0190*		
a2	6.554×10^{-5}	2.810×10^{-5}	2.333	0.0271*		
p2	-6.640×10^{-3}	4.106×10^{-3}	-1.617	0.1171		
a3	6.489×10^{-4}	1.018×10^{-4}	6.373	0.0000***		

¹The BCS value assigned by the Hervieu and Morand-Fehr (1999) method was the dependent variable and the body measurements were the independent variables. The front and rump views were compared by means of their adjusted R². The view that has a greater adjusted R² represents the region whose variables better explained the relationship between the BCS and the measurements performed.

²a1 = sternum area, p1 = sternum perimeter, a2 = rump region area, p2 = rump region perimeter, a3 = lateral virtual area.

****P* < 0.001; ***P* < 0.01; **P* < 0.05.

tion between image and template, we first extracted the rump feature positions; that is, their pixel position in terms of column *u* and row *v* in the image. To allow a technician to recover each feature’s pixel coordinates, we developed a Matlab application for manual feature labeling.

Based on the recommendations of Azzaro et al. (2011), 2 technicians performed the labeling process to ensure the final data set was not biased by subjectivity.

Labeling consisted of 2 parts. In the first, we annotated 11 points in the rump region, as illustrated in Figure 2 (rump region). In the second part, we an-

Figure 2. Measurements and their relation with body features. The panel on the left (front view) shows the following measurements: sternum area (a1), sternum perimeter (p1), smaller width at the level of the fifth rib (w1), larger width at the level of the scapula (w2). The center panel (rump region) shows the width between tuber sacrale (hip or hook bones; w3), smaller width of the rump region (w4), width between tuber ischia (pin bones; w5), length of the rump region (l1), rump region area (a2), and rump region perimeter (p2). The panel on the right (lateral rump region) shows the lateral virtual area (a3) and the lateral virtual perimeter (p3). The lateral virtual area (a3) corresponds to the negative or positive area limited by the line that unites the hipbone and the ipsilateral thurl; the lateral virtual perimeter (p3) corresponds to the line surrounding a3.

Aligned Images

Figure 3. Scheme showing the shape alignment process of different body sizes and camera poses using the 4 reference features in the animal rump.

notated the lateral virtual area (a_3) and the lateral virtual perimeter (p_3) as illustrated in Figure 2 (lateral rump region).

Following the labeling process, each feature f in each image was associated with a position in the image represented by coordinates (u_f, v_f) . We used the Matlab function *cp2tform* to find the projective transformation, T_p , that, when applied to the coordinates of the 4 reference features in the original image, transformed them to the coordinates of the 4 reference features in the template. We estimated the unbiased body measurements by applying the same transformation to all remaining features. To create an aligned image; that is, an image where the referent features are aligned with the template, we used T_p to align images using the Matlab function *imtransform*.

Validation of Visual Body Condition Scoring System. We evaluated the relationship between BCS

and the unbiased body measurements using a stepwise linear regression. To explore relationships above the linear, we first computed the second- and third-order monomial of the measurements. We then applied a stepwise linear regression between these monomials and the BCS, which is still minimum variance estimator.

Development of Representative Images. We grouped all the animals into BCS categories: 0 = very thin ($BCS \leq 2$), 1 = normal ($BCS > 2$ and ≤ 3.5), and 2 = very fat ($BCS > 3.5$). Further details on these BCS categories are available in the Discussion section. We designed the categories to be mutually exclusive, so that each assessed goat fell into only one category. From the 171 goats, 38 were very thin, 91 were normal, and 42 were very fat.

We chose the representative animals based on the center of each category defined on the estimated BCS $\hat{b}$; that is, for each animal in a given category we com-

puted the estimated BCS from the unbiased body measurements and the third-order polynomial estimated during the scoring system validation. The center of the category, $\hat{b}_c$, corresponds to the estimated scores' average for that category. Thus, animal a_c at the center of a category, and therefore its best representative, is the animal with an estimated BCS closest to $\hat{b}_c$. To avoid sampling bias, we selected 10 more animals with similar visual distribution features; that is, we chose a set of animals a_i with the minimum distance to a_c , where the distance is defined as $d(a_c, a_i) = \|m_c - m_i\|$, where m_c and m_i are the unbiased body measurements.

To create a single sketch, we combined the 10 images of each category into a single image. We did so by looking at the distribution of each pixel value over the whole set of images. For example, let $c_i(u, v)$ be the color of the pixel (u, v) in image 1. To estimate the color the pixel (u, v) in the sketch, we first collect the colors of all pixels (u, v) in the set of 10 images: $s(u, v) = \{c_1(u, v), c_2(u, v), \dots, c_{10}(u, v)\}$. Then we computed $c_{\text{sketch}} = \text{median}\{s(u, v)\}$. To minimize the effect of color variation in the different breeds, we first converted all colors to grayscale, using Matlab *function rgb2gray*, and ensured that the distribution of colors was similar, using Matlab *function histeq*.

Scientific Illustration for Final Visual Body Condition Scoring System. We gave the final sketches to a professional scientific illustrator and, after a conjunct analysis of several preliminary drawings, a final visual body condition scoring system was developed.

Training Program Development for Observers

Our training program occurred between the months of September and November 2013 and applied the idea of threshold images. We defined the threshold as the interval of overlapping estimated body measurements between 2 consecutive categories. The animals that were close to the center of the interval were used as threshold representatives. We determined that an animal was in the threshold interval if its estimated BCS was higher or lower than those of a consecutive category. For example, consider an animal a_i in the very thin category with an estimated BCS $\hat{b}_i$. Consider also that the minimum estimated BCS in the normal category is $\hat{b}_m$. So, a_i is in the threshold region if and only if $\hat{b}_i > \hat{b}_m$. We selected the representative animals and constructed the sketch images for the threshold animals using the same approach we used for each category. We then integrated these sketches into a training program developed for potential scoring system users.

Interobserver Reliability Study

The goats used in the analysis of interobserver reliability were kept in the same husbandry and management conditions as described above. Each individual goat was observed simultaneously by 4 observers. Observers 1 and 3 were veterinarians with 20 and 5 yr of experience, respectively, in clinical and research work with small ruminants, and they used BCS routinely in their studies in pregnancy toxemia. Observers 2 and 4 were master's degree veterinary students involved in research with dairy goats for their theses. Observers 1 and 3 had previous experience in BCS with the Hervieu and Morand-Fehr (1999) method, but observers 2 and 4 did not. The standing goats were assessed from behind, ensuring there was good visualization of the rump region. The observers were positioned caudally and slightly dorsally to each animal. The assessment of each goat took approximately 1 min. The observers were not allowed to discuss their scorings during the observations.

Considering an expected interobserver reliability (p_1) of 0.8, acceptable (p_0) at 0.6 or higher, with $\alpha = 0.05$ and $\beta = 0.2$, we estimated a minimum number of 40 observations to complete the studies (Walter et al., 1998). Previous studies have shown that substantial score variation between observers with different experience is to be expected (Ferguson et al., 1994; Kristensen et al., 2006) so analysis was done separately per observer experience; 2 observers were considered for the calculation of the sample size needed for the reliability study. This study included 2 experiments.

Experiment 1. The day of field testing took place in December 2013. Interobserver agreement was assessed before training in the newly developed visual body condition scoring system, meaning that no information on threshold images was given. We used 45 dairy goats (very thin = 6; normal = 19; very fat = 20). The breeds were Saanen ($n = 26$), Alpine ($n = 7$), and crossbreed (Saanen $\times$ Alpine; $n = 12$).

Experiment 2. The experiment was performed during one day of field testing in February 2014. This second experiment was identical to the first except that the observers were trained before scoring the goats. The training was given by the scoring system developers and consisted of a theoretical presentation with careful explanation on how to assess, such as position of the observer and distance from the goat. Instruction on how to use the visual body condition scoring system was aided by showing sketches of the representative and threshold images and multiple photos of the different BCS categories and threshold animals. This was followed by a self-assessment test (the visual body

condition training program is available upon request from the authors). A practical session with live animals was also conducted during which the scores were compared and discussed. After the training, we used 49 dairy goats (very thin = 4; normal = 36; very fat = 9). The breeds were Saanen ($n = 25$), Alpine ($n = 9$), and crossbreed (Saanen $\times$ Alpine); $n = 15$).

Agreement (proportion of exact agreement and disagreement, by 1 and 2 points) and reliability (unweighted and weighted kappa) measures were calculated. The proportion of agreement assesses the ability of observers to give an identical BCS, and it is calculated by dividing the number of agreements by the total number of agreements and disagreements (Martin and Bateson, 2007). A value around 75% is generally considered indicative of good agreement (Burn and Weir, 2011).

Reliability measures the ability of observers to differentiate among the different levels of a scoring system (Streiner and Norman, 2008; Kottner et al., 2011). The weighted kappa (k_w) coefficient indicates the proportion of agreement beyond that expected by chance and penalizes disagreements in terms of their seriousness (unweighted kappa, k , treats all disagreements equally); it is, therefore, appropriate for ordinal scoring systems (Cohen, 1968; Lantz, 1997). As advised by Streiner and Norman (2008), we followed the quadratic weighting scheme, where the disagreement weights were based on the square of the amount of discrepancy. We followed Fleiss thresholds for k : 0–0.40 = poor; 0.41–0.75 = fair to good; and 0.76–1 = excellent; and Landis and Koch for k_w : <0 = poor; 0.00–0.20 = slight; 0.21–0.40 = fair; 0.41–0.60 = moderate; 0.61–0.80 = substantial; and 0.81–1 = almost perfect (Streiner and Norman, 2008).

Data were analyzed using the base packages of the R statistical language (R Core Team, 2013) and the specialized “psych” R package (Revelle, 2014).

RESULTS AND DISCUSSION

Preliminary Study for Scoring System Development

In our preliminary study, the stepwise regression of the “rump region view” had a higher R^2 , indicating that the rump region variables better explained the relationship between BCS and the measurements performed. These results are in line with studies in dairy cows (Ferguson et al., 2006) and buffaloes (Negretti et al., 2008), which found that BCS could be assigned from visualization of digital photographs taken from the rear of the animals. Moreover, in the Edmonson et al. (1989) study, the overall score was most closely related to the scores of the pelvic and tail-head areas of the cow.

Thus, we demonstrated a strong relationship between dairy goat rump body measurements and BCS. This

result allowed us to identify which part of the goat we should use in the visual scoring system and where we should focus our final, and larger, validation study.

Final Study for Scoring System Development

The main objective of including BCS in on-farm welfare assessment schemes is to identify animals at risk of nutritional deficiency or excess, corresponding to very thin animals and very fat animals, that are predisposed to metabolic diseases and poor welfare; hence, the method does not need to be very detailed (Winckler et al., 2003). In a study in dairy cows, although the mean differed between observers, assessors were consistent in their identification of very thin and very fat animals (Ferguson et al., 2006).

The recommendations for BCS of dairy goats under intensive conditions are >2.25 and <3.5 at dry-off; >2.75 and ≤ 3.5 at parturition; and >2 at peak lactation (Smith and Sherman, 2009). Anzuino et al. (2010), in their on-farm welfare assessment, established cut-off points for obviously thin as ≤ 1.5 and ≥ 4 for very fat. However, these cut-off values correspond to extreme conditions representing only emaciated or obese animals, excluding some that might also be considered at high risk in terms of nutritional balance.

In fact, when using the Hervieu and Morand-Fehr (1999) scoring system, evidence suggests that goats with scores of 2.0 will show fertility problems (Santucci et al., 1991). Moreover, McGregor and Butler (2008) observed a correlation between BCS and mortality in goats, finding that mortality increased rapidly in cold weather conditions when BCS was <2.0, with no mortality at BCS ≥ 2.5 . For these reasons, we established the cut-off of ≤ 2 for the very thin goats. On the opposite side of the scoring system, we considered very fat goats as those that scored >3.5, as BCS above this value correspond to a higher risk for the development of pregnancy toxemia and other health problems (Smith and Sherman, 2009).

In the final study, the animals' BCS variability increased considerably, representing the reality on many dairy farms. We also increased the number of images and included animals in the BCS extremes, which is not common in the majority of farm animal body condition research (Bewley et al., 2008). We thus had animals from different breeds and with a wide range of sizes and BCS. To avoid biasing the results due to changes in body size that may not be a consequence of BCS, we mapped the rump features to a standard template. This was done using a projective transformation between the rump plane and the template. We could not use a simple scaling of feature positions as in Azzaro et al. (2011), because we had distortion from changes in the camera

Figure 4. Scheme representing the distortion introduced from changes in the camera orientation with respect to rump plane.

orientation with respect to rump plane. As depicted in Figure 4, the camera orientation introduces different scale factors over the horizontal and vertical axes of the rump region. As animals have different heights, it is not possible to find a fixed setup that would ensure a fixed orientation between the rump plane and the camera, in contrast to what was suggested by Azzaro et al. (2011), where the features of interest are all in a plane parallel to the floor, independent of the animal's characteristics.

To align the animals to the template, we manually annotated them, which ensured an accuracy that would not have been possible if we had used automatic feature matching such as SIFT (scale-invariant feature transform; Lowe, 2004).

Using the corrected features in the extended data set, we inferred that the unbiased body measurements and the BCS were related by a third-order polynomial, as demonstrated by the highest adjusted R^2 (Table 2). We could now use the third-order relation to parameterize animals in the data set using the estimated BCS $\hat{b}$. The parameterization defines an order over the whole set of animals, as if we had mapped the set of all animals into a line. This line includes information on both BCS and rump measurements. Thus, we defined the representative of each category and the threshold regions between categories based on the distribution of the animals over this line. In particular, we defined, as the most relevant animal of each category, the one closest to the $\hat{b}$ distribution mean.

When selecting a group of images close to the representative, and because rump shape has high variability, we needed to ensure that we chose a set of images where the rump of the animals presented adequate overlap. We thus chose a set of animals whose appearance was similar to that of the representative. Having shown the relationship between the rump features and the anatomical characteristics we wanted to highlight, we could use rump features to define similarity between animals instead of a more common texture or shape

descriptors (Belongie et al., 2002). We finally created representative sketches and scientific illustrations for each category (Figure 5). The estimated BCS for the illustrations correspond to 1.75, 2.48, and 3.70 for the very thin, normal, and very fat categories, respectively. We obtained an estimated BCS of 2.10 and 3.14 for the respective threshold image sketches.

We found that the use of scientific illustrations increased clarity, reproducibility, and repeatability. For the final scoring system (Figure 6), we also included some verbal descriptors to each level of assessment. This is frequently used with visual descriptors (Bana e Costa and Beinat, 2005) to aid assessors making their decision.

Training Development for Observers and Interobserver Reliability Study

A benefit of our study was the development of training material on the use of the visual body condition scoring system in goats. The integration of threshold images (Figure 5) greatly benefited the observers' training, as can be seen by the results of the interobserver reliability study (Table 3).

Before training, the percentage of exact agreement was lower for the 2 experienced observers (67%), prob-

Table 2. Summary of the adjusted R^2 and residual standard error for the 3 analyzed models used to evaluate the relationship between BCS and unbiased body measurements using a stepwise linear regression

Model ¹	Adjusted R^2	Residual SE (df)
First-order polynomial	0.70	0.46 (164)
Second-order polynomial	0.73	0.44 (160)
Third-order polynomial	0.74	0.43 (160)

¹In the first model we explored a linear relation (first-order polynomial); we then explored relations above the linear, computing the second- (second-order polynomial) and third-order monomial (third-order polynomial) of the measurements. A higher adjusted R^2 was used as the criterion for determining the model that best explains the relationship between BCS and unbiased body measurements

Figure 5. Representative images for the BCS categories (very thin, normal, and very fat) and the respective scientific illustrations; we also include the threshold images that delimit each category: threshold I marks the separation between very thin and normal, and threshold II between normal and very fat. Each representative image was created by combining the images of several animals of each category. The procedures for selecting the animals from each category and for combining are described in subsection Development of Representative Images. Both threshold images were created using the procedure described in subsection Training Program Development for Observers.

ably due to biases of previous experience with other BCS methods. In contrast, the inexperienced observers had a very good level of agreement (82%; Table 3). After training, exact agreement increased in both experienced and inexperienced observers, reaching very similar levels (86 and 82%, respectively; Table 3). Moreover, before training, the pair of experienced observers had agreement of 100% on the very thin goats, 47% on the normal goats, and 63% on the very fat goats. After training, the pair kept the 100% agreement for the very thin goats and increased to 69 and 89% agreement for, respectively, the normal and very fat goats (data not shown). The inexperienced observers reached agreement of 100% after training for both the very thin and very fat goats, having started at 83 and 68%, respectively (data not shown).

Both pairs, before and after training, already had very high percentages of exact agreement on the very thin goats; therefore, this category seems to be self-evident. However, training improved agreement in both pairs of observers when scoring the very fat and normal goats, showing the importance of training, especially on the threshold images, to clearly identify and distinguish normal from very fat goats.

For the experienced observers, the k and k_w coefficients improved with the training, reaching almost perfect agreement with the k_w coefficient. The greater increase in k compared with that of k_w indicates that more scores became absolutely equal after the training.

The inexperienced observers did not show any improvement in their k and k_w coefficients after training. Training the inexperienced observers seemed to

Table 3. Interobserver agreement in the evaluation of the visual body condition scoring system before and after training of observers, respectively

Observer ¹	Before training (n = 45)				After training (n = 49)			
	Exact agreement, % (no.)	One-point difference, % (no.)	k^2 (95% CI)	k_w^2 (95% CI)	Exact agreement, % (no.)	One-point difference, % (no.)	k (95% CI)	k_w (95% CI)
Observers 1 and 3	66.7 (30)	100 (45)	0.49 (0.29–0.69)	0.70 (0.55–0.86)	85.7 (42)	100 (49)	0.75 (0.58–0.92)	0.83 (0.70–0.95)
Observers 2 and 4	82.2 (37)	100 (45)	0.70 (0.51–0.89)	0.80 (0.66–0.94)	81.6 (40)	100 (49)	0.70 (0.51–0.88)	0.80 (0.67–0.93)

¹Observers 1 and 3 were experienced; observers 2 and 4 were inexperienced.
² k = kappa score; k_w = weighted kappa score.

have only a marginal effect on their k and k_w values; however, their values were already considered almost perfect before training, so there was not much room for improvement. This could indicate that the visual scoring system was by itself clear and unambiguous, and that the inexperienced observers were very open to new information and interested in the new method, whereas the experienced observers had to be more convinced on

	Very thin	Normal	Very fat
General condition	Raw or slightly-raw boned goat, with backbone and some ribs visible.	Backbone not prominent but still visible and ribs difficult to assess visually.	Backbone and ribs not visible. Goat has a rounded appearance, sometimes with abdominal fat deposits visible.
Rump region
	Hip and pin bones are prominent. The line that connects the hip bone and the thurl assumes a markedly concave shape. There is little muscle and/or fat between the skin and bone structures.	Hip and pin bones still visible, but not prominent. The line that connects the hip bone the thurl assumes a slightly concave or straight shape. It is possible to realize some muscle and/or fat between the skin and bone structures.	Hip and pin bones are difficult to identify. The line that connects the hip bone the thurl assumes a slightly or markedly convex shape. All the rump region is coated by muscle and fat, contributing to the rounded appearance of the goat.

Figure 6. Final visual body condition scoring system.

the usefulness of the tool. This positive effect of training for BCS assessment purpose was also reported for BCS in dairy cows (Kristensen et al., 2006), and hence it seems important to enhance interobserver reliability. In the context of welfare assessment, where the training of different observers with different levels of experience is a common situation, the development of training materials that could be made available online is valuable.

Construction of the visual body condition scoring system developed here was possible because the visual cues on the rump region reflected the animals' BCS. However, the extent to which rump visual cues reflect the BCS of each animal depends on several factors, including breed. The resulting large variability in the relationship between visual cues and BCS allows only for separation of animals into the 3 categories we presented. To obtain a more discriminative scoring system, we would have to decrease the variability by identifying each underlying factor. We could then create dedicated scoring systems by separating the animals according to those factors. We might achieve several more discriminative visual scoring systems, but overall, they would be more difficult to use. The result would compromise the main objective that lead to the creation of this scoring system—rapid assessment of a large number of animals.

CONCLUSIONS

This study provides a practical tool for evaluating BCS in dairy goats by reducing subjectivity. The AWIN visual body condition scoring system for dairy goats was assessed for validity and repeatability, and we demonstrated that it is simple, repeatable, and expeditious compared with conventional methods. It is also less stressful to the goats and can be used by inexperienced observers with substantial levels of interobserver reliability. The AWIN scoring system is not intended to replace current detailed scores (especially considering the normal category) that are valuable for effective production monitoring, but rather is able to provide a quick first-level assessment. Further studies are needed to set the maximum number of very thin and very fat animals that would trigger a second-level assessment. We proposed a systematic method for the creation of visual scoring systems with the potential to be applied in other contexts, and introduced the concept of threshold images to be used in the development of training programs.

ACKNOWLEDGMENTS

The authors thank João Paulo Costeira (Universidade de Lisboa, Lisbon, Portugal), Manuela Veloso

(Carnegie Mellon University, Pittsburgh, PA), Rui Caldeira (Universidade de Lisboa), Severiano R. Silva (Universidade de Trás-os-Montes e Alto Douro, Vila real, Portugal), and Telmo Nunes (Universidade de Lisboa) for their valuable comments and helpful discussion throughout the study, the farmers who allowed us to visit their farms, and the students who helped us with data collection. We also acknowledge Pedro Salgado (Universidade de Lisboa) for the scientific illustration of the final visual body condition scoring system. The Animal Welfare Indicators (AWIN) project (FP7-KBBE-2010-4-266213) has received funding from the European Union Seventh Framework Programme for research, technological development, and demonstration.

REFERENCES

- Anzuino, K., N. J. Bell, K. J. Bazeley, and C. J. Nicol. 2010. Assessment of welfare on 24 commercial UK dairy goat farms based on direct observations. *Vet. Rec.* 167:774–780. <http://dx.doi.org/10.1136/vr.c5892>.
- Azzaro, G., M. Caccamo, J. D. Ferguson, S. Battiato, G. M. Farinella, G. C. Guarnera, G. Puglisi, R. Petriglieri, and G. Licitra. 2011. Objective estimation of body condition score by modeling cow body shape from digital images. *J. Dairy Sci.* 94:2126–2137. <http://dx.doi.org/10.3168/jds.2010-3467>.
- Bana e Costa, C.A., and E. Beinat. 2005. Model-structuring in public decision-aiding. Working paper. Pages 12–25. London School of Economics and Political Science, London, UK.
- Belongie, S., J. Malik, and J. Puzicha. 2002. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.* 24:509–522. <http://dx.doi.org/10.1109/34.993558>.
- Bewley, J. M., A. M. Peacock, O. Lewis, R. E. Boyce, D. J. Roberts, M. P. Coffey, S. J. Kenyon, and M. M. Schutz. 2008. Potential for estimation of body condition scores in dairy cattle from digital images. *J. Dairy Sci.* 91:3439–3453. <http://dx.doi.org/10.3168/jds.2007-0836>.
- Botreau, R., I. Veissier, and P. Perny. 2009. Overall assessment of animal welfare: strategy adopted in Welfare Quality®. *Anim. Welf.* 18:363–370.
- Burges, C. J. C. 1998. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.* 2:121–167. <http://dx.doi.org/10.1023/A:1009715923555>.
- Burn, C. C., and A. A. S. Weir. 2011. Using prevalence indices to aid interpretation and comparison of agreement ratings between two or more observers. *Vet. J.* 188:166–170. <http://dx.doi.org/10.1016/j.tvjl.2010.04.021>.
- Caroprese, M., D. Casamassima, S. P. G. Rassu, F. Napolitano, and A. Sevi. 2009. Monitoring the on-farm welfare of sheep and goats. *Ital. J. Anim. Sci.* 8:343–354. <http://dx.doi.org/10.4081/ijas.2009.s1.343>.
- Cohen, J. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.* 70:213–220.
- Edmonson, A. J., I. J. Lean, L. D. Weaver, T. Farver, and G. Webster. 1989. A body condition scoring chart for Holstein dairy cows. *J. Dairy Sci.* 72:68–78. [http://dx.doi.org/10.3168/jds.S0022-0302\(89\)79081-0](http://dx.doi.org/10.3168/jds.S0022-0302(89)79081-0).
- Ferguson, J. D., G. Azzaro, and G. Licitra. 2006. Body condition assessment using digital images. *J. Dairy Sci.* 89:3833–3841. [http://dx.doi.org/10.3168/jds.S0022-0302\(06\)72425-0](http://dx.doi.org/10.3168/jds.S0022-0302(06)72425-0).
- Ferguson, J. D., D. T. Galligan, and N. Thomsen. 1994. Principal descriptors of body condition score in Holstein cows. *J. Dairy Sci.* 77:2695–2703. [http://dx.doi.org/10.3168/jds.S0022-0302\(94\)77212-X](http://dx.doi.org/10.3168/jds.S0022-0302(94)77212-X).

- Halachmi, I., P. Polak, D. J. Roberts, and M. Klopčič. 2008. Cow body shape and automation of condition scoring. *J. Dairy Sci.* 91:4444–4451. <http://dx.doi.org/10.3168/jds.2007-0785>.
- Harwood, D. 2006. The Normal Goat—Feeding and Management. Pages 15–23 in *Goat Health and Welfare: A Veterinary Guide*. The Crowood Press Ltd., Ramsbury, UK.
- Hervieu, J., and P. Morand-Fehr. 1999. Comment noter l'état corporel des chèvres. *Réussir Chèvre Rev. Elev. Chèvres Ist. L'Elev.* 231:26–32.
- Knierim, U., and C. Winckler. 2009. On-farm welfare assessment in cattle: Validity, reliability and feasibility issues and future perspectives with special regard to the Welfare Quality approach. *Anim. Welf.* 18:451–458.
- Kottner, J., L. Audigé, S. Brorson, A. Donner, B. J. Gajewski, A. Hróbjartsson, C. Roberts, M. Shoukri, and D. L. Streiner. 2011. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J. Clin. Epidemiol.* 64:96–106. <http://dx.doi.org/10.1016/j.jclinepi.2010.03.002>.
- Kristensen, E., L. Dueholm, D. Vink, J. E. Andersen, E. B. Jakobsen, S. Illum-Nielsen, F. A. Petersen, and C. Enevoldsen. 2006. Within- and across-person uniformity of body condition scoring in Danish Holstein cattle. *J. Dairy Sci.* 89:3721–3728. [http://dx.doi.org/10.3168/jds.S0022-0302\(06\)72413-4](http://dx.doi.org/10.3168/jds.S0022-0302(06)72413-4).
- Lantz, C. A. 1997. Application and evaluation of the kappa statistic in the design and interpretation of chiropractic clinical research. *J. Manipulative Physiol. Ther.* 20:521–528.
- Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* 60:91–110. <http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94>.
- Lowman, B. G., N. A. Scott, and S. H. Somerville. 1976. Condition Scoring of Cattle. School of Agriculture, Edinburgh, UK.
- Martin, P., and P. Bateson. 2007. *Measuring Behaviour: An Introductory Guide*. Cambridge University Press, Cambridge, UK.
- McGregor, B. A., and K. L. Butler. 2008. Relationship of body condition score, live weight, stocking rate and grazing system to the mortality of Angora goats from hypothermia and their use in the assessment of welfare risks. *Aust. Vet. J.* 86:12–17. <http://dx.doi.org/10.1111/j.1751-0813.2007.00249.x>.
- Morand-Fehr, P. 2005. Recent developments in goat nutrition and application: A review. *Small Rumin. Res.* 60:25–43. <http://dx.doi.org/10.1016/j.smallrumres.2005.06.004>.
- Negretti, P., G. Bianconi, S. Bartocci, S. Terramocchia, and M. Verna. 2008. Determination of live weight and body condition score in lactating Mediterranean buffalo by visual image analysis. *Livest. Sci.* 113:1–7. <http://dx.doi.org/10.1016/j.livsci.2007.05.018>.
- R Core Team. 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org>.
- Revelle, W. 2014. *psych: Procedures for personality and psychological research*. Version 1.4.8. Northwestern University, Evanston, IL. <http://CRAN.R-project.org/package=psych>.
- Roche, J. R., N. C. Friggens, J. K. Kay, M. W. Fisher, K. J. Stafford, and D. P. Berry. 2009. Invited review: Body condition score and its association with dairy cow productivity, health, and welfare. *J. Dairy Sci.* 92:5769–5801. <http://dx.doi.org/10.3168/jds.2009-2431>.
- Russel, A. J. F., J. M. Doney, and R. G. Gunn. 1969. Subjective assessment of body fat in live sheep. *J. Agric. Sci.* 72:451–454. <http://dx.doi.org/10.1017/S0021859600024874>.
- Santucci, P., A. Branca, M. Napoleone, R. Bouche, G. Aumont, F. Poisot, and G. Alexandre. 1991. Body condition scoring of goats in extensive conditions. Pages 240–255 in *Goat Nutrition*. Wageningen Academic Publishers, Wageningen, the Netherlands.
- Santucci, P., and O. Maestrini. 1985. Body composition of dairy goats in extensive systems of production: method of estimation. *Ann. Zootech.* 34:473–474.
- Smith, M. C., and D. M. Sherman. 2009. Nutrition and metabolic diseases. Pages 733–786 in *Goat Medicine*. John Wiley & Sons, Ames, IA.
- Streiner, D. L., and G. R. Norman. 2008. *Health Measurement Scales: A Practical Guide to their Development and Use*. 4th ed. Oxford University Press, Oxford, UK.
- The MathWorks Inc. 2012. *MATLAB and Statistics Toolbox Release*. The MathWorks Inc., Natick, MA.
- Tu, Z., and A. L. Yuille. 2004. Shape matching and recognition—Using generative models and informative features. Pages 195–209 in *Computer Vision—ECCV 2004*. T. Pajdla and J. Matas, ed. Springer, Berlin, Germany.
- Walter, S. D., M. Eliasziw, and A. Donner. 1998. Sample size and optimal designs for reliability studies. *Stat. Med.* 17:101–110.
- Welfare Quality. 2009a. *Welfare Quality assessment protocol for pigs*. Welfare Quality Consortium, Lelystad, the Netherlands.
- Welfare Quality. 2009b. *Welfare Quality assessment protocol for cattle*. Welfare Quality Consortium, Lelystad, the Netherlands.
- Wildman, E. E., G. M. Jones, P. E. Wagner, R. L. Boman, H. F. Troutt Jr., and T. N. Lesch. 1982. A dairy cow body condition scoring system and its relationship to selected production characteristics. *J. Dairy Sci.* 65:495–501. [http://dx.doi.org/10.3168/jds.S0022-0302\(82\)82223-6](http://dx.doi.org/10.3168/jds.S0022-0302(82)82223-6).
- Winckler, C., J. Capdeville, G. Gebresenbet, B. Hörning, U. Roiha, M. Tosi, and S. Waiblinger. 2003. Selection of parameters for on-farm welfare-assessment protocols in cattle and buffalo. *Anim. Welf.* 12:619–624.