Animal (2018), **12:9**, pp 1942–1949 © The Animal Consortium 2018
doi:10.1017/S1751731117003597

animal

Inter-observer reliability of animal-based welfare indicators included in the Animal Welfare Indicators welfare assessment protocol for dairy goats

A. Vieira[1†], M. Battini[2], E. Can[3,4], S. Mattiello[2] and G. Stilwell[3]

[1]*Centre for Management Studies, Instituto Superior Técnico, Universidade de Lisboa, Avenida Rovisco Pais, 1049-001 Lisbon, Portugal;* [2]*Laboratorio di Benessere animale, Etologia applicata e Produzioni sostenibili, Dipartimento di Medicina Veterinaria, Università degli Studi di Milano, Via G. Celoria 10, 20133 Milan Italy;* [3]*Faculdade de Medicina Veterinária, Universidade de Lisboa, Polo Alto da Ajuda, 12 1300-477 Lisbon, Portugal;* [4]*Division of Agricultural and Environmental Sciences, University of Nottingham, Sutton Bonington Campus, College Road, Sutton Bonington, Loughborough LE12 5RD, UK*

(Received 16 February 2017; Accepted 29 November 2017; First published online 8 January 2018)

This study was conducted within the context of the Animal Welfare Indicators (AWIN) project and the underlying scientific motivation for the development of the study was the scarcity of data regarding inter-observer reliability (IOR) of welfare indicators, particularly given the importance of reliability as a further step for developing on-farm welfare assessment protocols. The objective of this study is therefore to evaluate IOR of animal-based indicators (at group and individual-level) of the AWIN welfare assessment protocol (prototype) for dairy goats. In the design of the study, two pairs of observers, one in Portugal and another in Italy, visited 10 farms each and applied the AWIN prototype protocol. Farms in both countries were visited between January and March 2014, and all the observers received the same training before the farm visits were initiated. Data collected during farm visits, and analysed in this study, include group-level and individual-level observations. The results of our study allow us to conclude that most of the group-level indicators presented the highest IOR level ('substantial', 0.85 to 0.99) in both field studies, pointing to a usable set of animal-based welfare indicators that were therefore included in the first level of the final AWIN welfare assessment protocol for dairy goats. Inter-observer reliability of individual-level indicators was lower, but the majority of them still reached 'fair to good' (0.41 to 0.75) and 'excellent' (0.76 to 1) levels. In the paper we explore reasons for the differences found in IOR between the group and individual-level indicators, including how the number of individual-level indicators to be assessed on each animal and the restraining method may have affected the results. Furthermore, we discuss the differences found in the IOR of individual-level indicators in both countries: the Portuguese pair of observers reached a higher level of IOR, when compared with the Italian observers. We argue how the reasons behind these differences may stem from the restraining method applied, or the different background and experience of the observers. Finally, the discussion of the results emphasizes the importance of considering that reliability is not an absolute attribute of an indicator, but derives from an interaction between the indicators, the observers and the situation in which the assessment is taking place. This highlights the importance of further considering the indicators' reliability while developing welfare assessment protocols.

Keywords: inter-observer reliability, dairy goat, on-farm welfare assessment, animal-based indicators, training

Implications

Nowadays, the use of animal-based indicators in welfare assessment schemes is usually preferred. However, their use entails several challenges such as the fulfilment of three requirements (validity, feasibility and reliability) that ensure that the indicators can be successfully used at farm level. Considering that animal-based indicators may be more prone to subjectivity than resource-based indicators, it is paramount to assess their reliability. To our knowledge, no empirical research studies have focused on assessing inter-observer reliability (IOR) of welfare indicators on goats and therefore more studies addressing this topic are strongly requested.

Introduction

Welfare assessment requires a multidimensional approach (Fraser, 1995). There are generally two categories of indicators that can be used to assess animal welfare at farm level: resource and animal-based (Johnsen *et al.*, 2001).

† E-mail: ana.lopesvieira@gmail.com

Traditionally, on-farm welfare assessment focused on the evaluation of resource-based indicators, such as the type of management provided to the animals (e.g. Bartussek, 1999). However, providing good management and resources may not reflect high standards of welfare. In fact, some studies question the validity of resource-based indicators, showing farms with similar production systems demonstrating a huge variation in animal welfare (Mülleder *et al*., 2007). Following these constraints, the interest in assessing the actual welfare state of the animals through animal-based measures has increased (Johnsen *et al*., 2001).

However, the integration of animal-based indicators in welfare assessment schemes entails several challenges. The first challenge relates to their availability: there are few studied animal-based indicators (Johnsen *et al*., 2001), and they are even fewer if we consider small ruminants (Battini *et al*., 2014; Caroprese *et al*., 2016). The second challenge refers to the fulfilment of three requirements, namely validity, feasibility and reliability that will allow the indicators to be used effectively at farm level (Waiblinger *et al*., 2001). Validity tells us the extent to which an indicator measures what it is supposed to measure, whereas feasibility refers to opportunity of successfully using the indicator during on-farm assessment. Reliability entails a particular challenge, as there is an associated level of subjectivity involved when using animal-based indicators, and so the observers' assessment might be biased by their own concern or level of empathy with the animals (Meagher, 2009). If reliability is poor, then the indicator is probably inappropriate for the evaluation of animal welfare, or should be further refined, namely by improving its definition, ensuring good data recording and better training the observers. These actions have proved to be essential to increase reliability and achieve an accurate and consistent welfare assessment outcome (De Rosa *et al*., 2009). Reliability weighs the amount of random and systematic error (the smaller the error, the more reliable is the measurement), and in a more intuitive way can be defined as 'the ability of scores of a measuring device to differentiate among subjects or objects' (Kottner *et al*., 2011, p. 104). Reliability can be explored by means of inter or intra-observer reliability studies, observer consistency studies and test-retest reliability studies (Scott *et al*., 2001; Streiner and Norman, 2008). In this paper, we will focus on IOR that measures 'the degree to which two or more ratters are able to differentiate among subjects or objects under similar assessment conditions' (Kottner *et al*., 2011, p. 104). Literature frequently use the term agreement as a synonym of reliability, however, they are different concepts. Inter-observer agreement, defined as 'the degree to which two or more ratters achieve identical results under similar assessment conditions' (Kottner *et al*., 2011, p. 104), can be used to assess stability of observations, being useful for data interpretation when compared with reliability measures.

After reliability, validity and feasibility are established, a scoring system has to be defined to operationalize the indicators in order to allow their integration in a welfare assessment protocol (Vieira, 2015). This operationalization is accomplished by choosing a scoring system which will imply different constraints in terms of level of measurement. In welfare assessment, ordinal and continuous scales are the two most common scoring systems (Scott *et al*., 2001) with the former being the most frequently used, as a result of being generally easier to deliver as they are based on a comparability assumption. On the other hand, ordinal scales are, in most cases, artificial constructs and therefore are sometimes associated with difficulties in their interpretation, which makes paramount to assess their reliability (Vieira, 2015). Moreover, there are different studies pointing out the challenge of conducting reliability studies under commercial conditions (Kaufman and Rosenthal, 2009). Therefore, the lack of studies, and especially of studies that provide information for data interpretation, is a drawback when one is considering the integration of certain indicators and their associated scoring systems in a welfare protocol. This is particularly important when considering animal-based indicators, and substantiates the scientific relevance of this paper.

The aim of this study was to evaluate IOR of animal-based indicators (at group and individual-level) of the Animal Welfare Indicators (AWIN) welfare assessment prototype protocol applied to dairy goats in intensive dairy farms. The study was implemented by conducting two field studies where two pairs of observers, one in Portugal and another in Italy, visited 10 farms each and applied the AWIN prototype protocol.

Material and methods

Farm visits

In all, 30 intensive commercial dairy goat farms in Portugal (PT) and 30 in Northern Italy (IT) were visited to apply the AWIN welfare assessment prototype protocol. In PT, the farms were randomly selected from a national database of *Direcção-Geral de Alimentação e Veterinária*, whereas in IT farms were selected with the support of the S.A.T.A. (Technical Advice Service for Farmers) of the Lombardy region, where the majority of intensive dairy goat farms are located and depending on farmers' availability, geographic location and farm size. With the objective of assessing IOR, 10 farms in IT and 10 farms in PT were visited between January and March 2014, with these farms being randomly selected among the 30 farms of each country. In PT, the number of adult dairy goats on each of the 10 selected farms ranged from 80 to 2000 animals, with a mean (± SD) of 470 (±560) goats; in IT, the number of dairy goats ranged from 32 to 912 animals, with a mean (± SD) of 227 (±279) goats. In all farms, goats were kept indoors on straw litter, although some farms also had access to an outdoor grazing or exterior pen, where the goats had the opportunity to exercise. Breeds were mainly Saanen and Alpine, milked twice a day, fed with Total Mixed Ration and with permanent access to water.

Each farm was visited by two observers. The two Italian observers had different background and level of experience

with dairy goats: one was an animal scientist with more than 3 years of experience with dairy goats, whereas the other was a veterinarian with no specific experience with dairy goats. The two Portuguese observers had a common background (veterinary), but different levels of experience: one had more than 3 years of experience working with dairy goats, the other had just graduated from veterinary school. All the observers received equal training before the farm visits were initiated. Each observer received digital training material for each of the indicators in assessment that included a detailed definition of the indicator, how to assess and score it, some examples and a self-evaluation section. Following this all observers received a full week training on the use of the AWIN prototype protocol. The training included theoretical and on-farm practical sessions.

On-farm assessment of animal-based indicators

In all, 25 animal-based indicators, classified in accordance with the four principles and 12 criteria developed by Welfare Quality® (Botreau *et al*., 2007) were assessed either at group-level (14 indicators) or at individual-level (11 indicators). Detailed descriptive criteria used to assess each animal-based indicator and the order of data collection, are presented in Battini *et al*. (2016) and Can *et al*. (2016). However, in order to understand the type of analyses employed for the reliability study, the type of data collection is briefly presented herewith. The data collection began with group-level observations of the following indicators: number of goats 'Improperly disbudded', 'Queuing at feeding', 'Queuing at drinking', with poor 'Hair coat condition', 'Oblivion', 'Abnormal lying', with signs of thermal stress (either 'Shivering' or 'Panting') and 'Kneeling at the feeding rack'. For all these indicators, the total number of goats in the pen presenting these conditions was recorded, and transformed into percentages out of the total number of goats in the assessed pen. Immediately after this assessment, 'Qualitative Behaviour Assessment' was conducted; subsequently, the observer entered the pen to perform both 'Latency to the first contact' and 'Avoidance distance tests', and finally 'Severe lameness' and 'Kneeling in the pen'. The individual-level assessment focused on 11 indicators: 'Body Condition Score (BCS)', 'Udder asymmetry', 'Cleanliness' (hindquarters, lower legs and udder), 'Lesions' (hindquarters, lower legs, body, neck and head), 'Knee calluses', 'Abscesses' (hindquarters, udder, body, neck and head), 'Overgrown claws', 'Discharges' (ocular, nasal or vulvar), and 'Faecal soiling'. All these indicators were recorded on the same animals (with both sides (left and right) being considered when relevant, e.g. for body, legs, claws) and were scored using a binary assessment system, except for 'BCS' and 'Knee calluses' that had three assessment scores. In order to speed up the individual assessment (Battini *et al*., 2016), whenever possible animals were restrained either at the feeding rack (in four Italian farms) or, depending on farmers' availability, at the milking parlour (in one Portuguese farm). In all other cases (nine Portuguese and six Italian farms), the animals had to be manually restrained inside the pen. The individual goats were selected based on a systematic sampling and after being assessed each goat was marked with an animal marker.

Only one pen was assessed in each farm. This decision was supported by the results of preliminary observations carried out in order to determine the optimal sampling strategy on the number of pens to be assessed: these observations showed the presence of significant differences between different farms, but almost no variation between pens in the same farm (Vieira *et al*., 2012). However, although we did not expect to find a great variation between pens, we decided to sample the pen considered as presenting the potentially highest risk for welfare. This selection takes into consideration the following aspects: highest density, lower feeding space/animal ratio, and lower drinking place/animal ratio, presence of both horned and hornless goats in the same pen. If all pens were in similar conditions, one random pen was selected. As the selection of the pen at higher risk may increase the prevalence of some indicators, and may therefore also affect the results on reliability, we excluded from assessment infirmary, culling, quarantine or maternity pens.

The number of goats to be assessed was proportional to the pen size, with percentages ranging from the totality of the goats in the pen ($\leqslant$15 animals) to a minimum of 25% of goats in the same pen (>150 animals), assuming a 50% prevalence, and considering a 90% interval of confidence and an accuracy of 10%. All data for each farm was collected on the same day by both assessors. The assessors applied the AWIN welfare assessment prototype protocol simultaneously and individually, not interfering or interacting with each other during all application of the protocol. Observers were new to the farms and had not previously performed any similar assessment on those same animals.

Statistical analysis

To assess IOR for continuous data (group-level indicators), intra-class correlations (ICCs) coefficients were calculated with a two-way mixed effects model (Shrout and Fleiss, 1979), that is, the subjects in the study were considered to be random but the observers were not random effects. There are several ICC variants that may be selected based on the nature of the study and the type of agreement the researcher wishes to estimate (Shrout and Fleiss, 1979; Hallgren, 2012). For the present study, we estimated ICC in terms of absolute agreement, as we wanted to take into account if good IOR is characterized by scores that are similar in absolute value. Estimates for ICC were interpreted using Shrout guidelines (Shrout, 1998): 0.0 to 0.10 = virtually none; 0.11 to 0.40 = slight; 0.41 to 0.60 = fair; 0.61 to 0.80 = moderate; 0.81 to 1.0 = substantial.

To assess IOR for categorical data (individual-level indicators), κ and weighted κ (κ_w) coefficients (Cohen, 1968) were calculated. κ consists of a measure of 'true' agreement that reflects the proportion of agreement fully chance corrected. Weighted κ (κ_w) penalizes disagreements in terms of their seriousness, whereas unweighted κ handles all disagreements equally not taking order of categories into

account, thus, being inappropriate for ordinal scales (Cohen, 1968). The quadratic weighting scheme, where disagreement weights are proportional to the square of the deviation of individual ratings (Brenner and Kliebsch, 1996), was used. We followed Fleiss thresholds for k: 0 to 0.40 = poor; 0.41 to 0.75 = fair to good; and 0.76 to 1 = excellent; and Landis and Koch for κ_w: <0 = poor; 0.00 to 0.20 = slight; 0.21 to 0.40 = fair; 0.41 to 0.60 = moderate; 0.61 to 0.80 = substantial; and 0.81 to 1 = almost perfect (Streiner and Norman, 2008). For individual-level indicators, the proportion of overall agreement, as a measure of inter-observer agreement was also calculated by dividing the number of agreements (both positive and negative agreements) by the total number of agreements and disagreements (Uebersax, 2014). A value around 75% is suggestive of good agreement (Burn and Weir, 2011). Data were analysed using the packages of the R statistical language.

Results

Group-level observations were performed on a total of 1518 adult dairy goats (734 in PT and 784 in IT), and individual observations were carried out on a total of 703 adult dairy goats (360 in PT and 343 in IT). Tables 1 and 2 report the number of cases observed in the 10 Portuguese and IT dairy goat farms (means values for the population of farms) for the indicators considered for the IOR study: 95% confidence intervals for the study population.

Group-level observations

The ICC for 'Improperly disbudded', 'Queuing at feeding', 'Hair coat condition' and 'Shivering score 1' showed the highest IOR level (substantial, 0.85 to 0.99) between observers and in both field studies (Table 3). The indicators 'Abnormal lying', 'Kneeling at the feeding rack' and 'Severe lameness' could only be computed for the IT study but also showed the highest IOR level (substantial, 0.85 to 0.92) (Table 3). The indicator 'Queuing at drinking' also presented the highest IOR level (0.99) in the PT study, but a moderate (0.67) level for the IT study (Table 3). The indicators 'Oblivion', 'Shivering score 2' and 'Panting score 1 and 2' could not be computed due to the low number, or absence, of recorded cases (see Table 1).

Individual-level observations

With the exception of the indicator 'Lesions_head', the proportion of overall agreement in both field studies, was above 75%, which is an indication of a good agreement among observers. In PT, this good agreement was followed by the highest IOR level (excellent, 0.80 to 0.95) for the indicators: 'Udder asymmetry', 'Cleanliness – hindquarters', 'Cleanliness – lower legs', 'Abscesses – udder', 'Abscesses – body', 'Overgrown claws', 'Ocular discharge' and 'Faecal soiling'. In IT, the same indicators, with the exception of 'Abscesses – body', presented the second highest IOR level (fair to good, 0.44 to 0.63) (Table 4). Overall, in both countries, the different indicators under the general description of 'Lesions' presented the lowest levels of agreement, being in some cases of poor agreement. In IT, where individual assessment was carried out in four farms at the feeding rack and in six farms using manual restraining, a higher IOR level for most indicators was achieved when animals were manually restrained (Table 5).

For both three-level indicators, 'BCS' and 'Knee calluses', proportion of overall agreement in both countries was above 75%, showing a good agreement among observers (Table 6). Regarding k_w, a substantial IOR level (0.79) was achieved in PT for both indicators, whereas in IT a moderate IOR level

Table 1 *Animal-based indicator's prevalence (group-level observations) observed in the 10 Portuguese (PT) and 10 Italian (IT) dairy goat farms*

	PT		IT	
Animal-based indicators	Mean (%)	95% CI	Mean (%)	95% CI
Improperly disbudded	19.5	16.7 <CI <22.6	16.3	13.8 <CI <19.1
Queuing at feeding	4.6	3.3 <CI <6.5	11.6	9.5 <CI <14.1
Queuing at drinking	3.4	2.3 <CI <5.1	1.9	1.1 <CI <3.2
Hair coat condition	15.7	13.1 <CI <18.5	24.6	21.7 <CI <27.8
Oblivion	0.4	0.1 <CI <1.3	0	–
Abnormal lying	0.4	0.1 <CI <1.3	1.3	0.6 <CI <2.4
Shivering				
Score 1	3.0	1.9 <CI <4.6	1.9	1.1 <CI <3.2
Score 2	0	–	0	–
Panting				
Score 1	0.7	0.2 <CI <1.7	0.1	0.01 <CI <0.8
Score 2	0	–	0	–
Kneeling at the feeding rack	0	–	0.4	0.1 <CI <1.0
Kneeling in the pen	1.2	0.6 <CI <2.4	0.5	0.2 <CI <1.0
Severe lameness	2.3	1.4 <CI <3.8	1.4	0.7 <CI <2.6

CI = confidence intervals.
Means values for the population of farms and 95% CI for the study population.

Table 2 *Animal-based indicator's prevalence (individual-level observations) observed in the 10 Portuguese (PT) and 10 Italian (IT) dairy goat farms*

	PT		IT	
Animal-based indicators	Mean (%)	95% CI	Mean (%)	95% CI
BCS				
Very thin	2.5	1.2 <CI <4.9	15.2	11.6 <CI <19.5
Very fat	14.7	11.3 <CI <18.9	5.0	3.0 <CI <8.0
Udder asymmetry	6.1	3.9 <CI <9.2	6.7	4.4 <CI <10.0
Cleanliness				
Hindquarter	20.3	16.3 <CI <24.9	36.4	31.4 <CI <41.8
Lower legs	24.4	20.2 <CI <50.0	38.8	33.6 <CI <44.1
Udder	2.5	1.2 <CI <4.9	5.5	3.4 <CI <8.6
Lesions				
Hindquarter	12.8	9.6 <CI <16.8	0.9	0.2 <CI <2.7
Lower legs	11.4	8.4 <CI <15.2	2.3	1.1 <CI <4.7
Body	14.4	11.1 <CI <18.6	4.7	2.8 <CI <7.6
Neck	19.2	15.3 <CI <23.7	2.6	1.3 <CI <5.1
Head	20.6	16.6 <CI <25.2	21.9	17.7 <CI <26.7
Knee calluses				
Score 1	90.6	86.9 <CI <93.3	95.0	92.0 <CI <97.0
Score 2	3.1	1.6 <CI <5.6	3.8	2.1 <CI <6.6
Abscesses				
Hindquarter	0.6	0.1 <CI <2.2	0	–
Body	3.3	1.8 <CI <5.9	5.8	3.7 <CI <9.0
Udder	2.2	1.0 <CI <4.5	3.4	1.9 <CI <6.2
Neck	3.3	1.8 <CI <5.9	4.1	2.3 <CI <6.9
Head	6.9	4.6 <CI <25.2	2.9	1.5 <CI <5.5
Overgrown claws	40.6	35.5 <CI <45.8	54.5	49.1 <CI <59.9
Ocular discharge	2.2	1.0 <CI <4.5	0.9	0.2 <CI <2.7
Nasal discharge	0.8	0.2 <CI <2.6	7.0	4.6 <CI <10.3
Vulvar discharge	0.3	0.01 <CI <1.8	0.9	0.2 <CI <2.8
Faecal soiling	6.9	4.6 <CI <10.2	21.6	17.4 <CI <26.4

CI = confidence intervals; BCS = body condition score.
Means values for the population of farms and 95% CI for the study population.

(0.46) was found for the BCS indicator and a fair IOR level (0.27) was found for the indicator 'knee calluses'.

Overall, and focusing on the group-level indicators that could be computed in both field studies, the two pair of observers reached the highest IOR level (substantial, 0.85 to 0.99) in all the group-level indicators (with the exception of 'Queuing at drinking'). However, when considering most of the individual-level indicators, the Portuguese pair of observers reached a higher level of IOR (excellent, 0.80 to 0.95) when compared with the Italian observers ('fair to good', 0.44 to 0.63). Regarding the individual-level indicators, the majority reached high (fair to good and excellent) IOR levels in both countries.

Discussion

Inter-observer reliability was measured in this study by conducting two field studies where reliability (ICC, *k*, and k_w) and agreement (proportion of overall agreement for individual-level indicators) measures were assessed between two pairs of observers, following a similar approach found in previous studies, as Mullan *et al*. (2011) and Phythian *et al*. (2013). The number of observers (two) for each assessed farm in this study may be considered small when compared with other studies (e.g. Mullan *et al*., 2011), however, it was determined due to feasibility constraints. Moreover, this constraint was overcome by the replication of the experiment in two different countries.

High levels of IOR support the selection of indicators being evaluated in an on-farm welfare assessment context (Hewetson *et al*., 2006; Kaler *et al*., 2009; Meagher, 2009; Phythian *et al*., 2012). This is particularly important when considering animal-based indicators, as indicators taken on animals are more prone to variation.

Most of the group-level indicators, namely 'Improperly disbudded', 'Queuing at feeding', 'Hair coat condition', and 'Shivering score 1' showed the highest level of IOR between observers in both field studies. These results show that the assessment was highly reliable pointing to a usable set of animal-based welfare indicators that were therefore included in the first level of the final AWIN welfare assessment protocol for dairy goats. The training received was appropriate

Table 3 *Inter-observer reliability for group-level observations*

	ICC (95% CI)	
Animal-based indicator	PT	IT
Improperly disbudded	0.99 (0.98 <CI <1)	0.87 (0.59 <CI <0.97)
Queuing at feeding	0.89 (0.62 <CI <0.97)	0.99 (0.99 <CI <1)
Queuing at drinking	0.99 (0.96 <CI <1)	0.67 (0.15 <CI <0.90)
Hair coat condition	0.85 (0.51 <CI <0.96)	0.95 (0.83 <CI <0.99)
Oblivion	n.a.	n.a.
Abnormal lying	n.a.	0.92 (0.71 <CI <0.98)
Shivering		
Score 1	0.99 (0.99 <CI <1)	0.88 (0.61 <CI <0.97)
Score 2	n.a.	n.a.
Panting		
Score 1	n.a.	n.a.
Score 2	n.a.	n.a.
Kneeling at the feeding rack	n.a.	0.89 (0.64 <CI <0.97)
Kneeling in the pen	n.a.	0.55 (-0.13 <CI <0.87)
Severe lameness	n.a.	0.85 (0.82 <CI <0.88)

ICC = intra-class correlation; CI = confidence intervals; n.a. = not possible to compute the information; PT = Portuguese; IT = Italian.
The table presents ICC computed from the assessments performed simultaneously by two observers in 10 PT and 10 IT farms, whereas applying the Animal Welfare Indicators welfare assessment prototype protocol to dairy goats.
ICC: 0.0 to 0.10 = virtually none; 0.11 to 0.40 = slight; 0.41 to 0.60 = fair; 0.61 to 0.80 = moderate; 0.81 to 1.0 = substantial (Shrout, 1998).

for the application of these indicators, overcoming differences in experience and background of the observers, which re-enforces other studies that highlight the importance of training or 'calibration' meetings (Ruddat *et al.*, 2014).

Regarding the individual-level indicators, the majority reached high (fair to good and excellent) IOR levels in both countries, and were above the maximum lower limit of 0.4 established by the Welfare Quality® project. However, IOR of individual-level indicators was lower when compared with the group-level indicators. Overall and considering the two field studies, there are several possible reasons to explain this difference. First of all, the number of individual-level indicators to be assessed on each animal was high, but the restraining time had to be kept to a minimum in order to speed up the total observation time and limit disturbance. This may have affected the reliability of each individual-level indicator. A further possible reason is the fact that individual observation of the animals is more challenging in terms of feasibility, as there is the need to restrain the animals, in what sometimes are sub-optimal conditions. In fact, during the field studies the restraining method was affected by farm characteristics and was different in IT and PT. Our results suggest that the restraining method can actually affect the reliability of the results and may help explaining the better

Table 4 *Agreement and inter-observer reliability for individual-level observations*

	Agreement		Reliability	
	Proportion of overall agreement (%) (95% CI)		κ (95% CI)	
Animal-based indicators	PT	IT	PT	IT
Udder asymmetry	99.44 (97.79 <CI <99.90)	92.42 (88.96 <CI <94.89)	0.95 (0.88 <CI <1)	0.44 (0.26 <CI <0.62)
Cleanliness				
Hindquarter	93.61 (90.43 <CI <95.82)	79.01 (74.23 <CI <83.12)	0.79 (0.71 <CI <0.87)	0.58 (0.50 <CI <0.67)
Lower legs	93.06 (89.79 <CI <95.37)	78.13 (73.31 <CI <82.32)	0.80 (0.73 <CI <0.88)	0.57 (0.49 <CI <0.65)
Udder	97.78 (95.50 <CI <98.96)	95.63 (92.73 <CI <97.44)	0.59 (0.33 <CI <0.85)	0.64 (0.48 <CI <0.81)
Lesions				
Hindquarter	91.39 (87.88 <CI <93.98)	97.67 (95.28 <CI <98.91)	0.67 (0.56 <CI <0.77)	0.19 (-0.14 <CI <0.52)
Lower legs	91.67 (88.19 <CI <94.22)	96.50 (93.80 <CI <98.10)	0.45 (0.28 <CI <0.61)	0.13 (-0.13 <CI <0.38)
Body	83.89 (79.59 <CI <87.45)	91.25 (87.62 <CI <93.93)	0.52 (0.42 <CI <0.63)	0.36 (0.18 <CI <0.53)
Neck	85.00 (80.79 <CI <88.44)	95.63 (92.73 <CI <97.44)	0.55 (0.44 <CI <0.65)	0.26 (0.01 <CI <0.52)
Head	86.67 (82.61 <CI <89.92)	64.43 (59.08 <CI <69.45)	0.63 (0.53 <CI <0.72)	0.27 (0.18 <CI <0.36)
Abscesses				
Hindquarter	99.44 (97.79 <CI <99.90)	100 (98.62 <CI <1)	n.a.	n.a.
Udder	99.72 (98.22 <CI <99.98)	95.63 (92.73 <CI <97.44)	0.93 (0.80 <CI <1)	0.52 (0.31 <CI <0.73)
Body	98.89 (96.98 <CI <99.64)	93.29 (89.97 <CI <95.61)	0.84 (0.69 <CI <0.99)	0.37 (0.17 <CI <0.58)
Neck	98.06 (95.86 <CI <99.15)	95.04 (92.03 <CI <96.99)	0.66 (0.42 <CI <0.89)	0.39 (0.16 <CI <0.62)
Head	95.00 (92.07 <CI <96.92)	97.38 (94.90 <CI <98.71)	0.50 (0.31 <CI <0.70)	0.60 (0.35 <CI <0.84)
Overgrown claws	95.56 (92.74 <CI <97.35)	82.22 (77.66 <CI <86.03)	0.91 (0.86 <CI <0.95)	0.64 (0.55 <CI <0.72)
Ocular discharge	99.44 (97.78 <CI <99.90)	98.54 (96.43 <CI <99.46)	0.89 (0.73 <CI <1)	0.44 (0.03 <CI <0.84)
Nasal discharge	98.89 (96.98 <CI <99.64)	91.25 (87.62 <CI <93.93)	0.49 (0.07 <CI <0.92)	0.38 (0.20 <CI <0.55)
Vulvar discharge	99.72 (98.22 <CI <99.99)	98.25 (96.04 <CI <99.29)	n.a.	n.a.
Faecal soiling	99.17 (97.38 <CI <99.78)	86.59 (82.41 <CI <89.92)	0.93 (0.86 <CI <1)	0.63 (0.54 <CI <0.73)

CI = confidence intervals; n.a. = not possible to compute the information; PT = Portuguese; IT = Italian.
The table presents the proportion of overall agreement and *k* scores computed from the assessments performed simultaneously by two observers in 10 Portuguese (PT) and 10 Italian (IT) farms, whereas applying the Animal Welfare Indicators welfare assessment prototype protocol to dairy goats.
k: 0 to 0.40 = poor; 0.41 to 0.75 = fair to good; and 0.76 to 1 = excellent (Streiner and Norman, 2008).

results achieved in PT for individual-level indicators. In PT animals were manually restrained in almost all the farms, whereas in IT many goats were assessed at the feeding rack. When the observers had the opportunity to manually restrain the goats, they probably had a better view of the whole body than when the animals were at the feeding rack. This is supported by the fact that in IT the IOR levels were higher when goats were manually restrained than when they were at the feeding rack. Under this last condition, the observers had a good view of the animal from behind, and this can probably justify the higher IOR levels reached for indicators mainly related to the udder.

Inter-observer reliability differences between countries, with the Portuguese pair of observers reaching a higher level of IOR in most of the individual-level indicators when compared with the Italian observers, may also be explained by the background and experience of the observers, with the common background of the Portuguese observers probably potentiating these results. Another reason for the difference of reliability results between the two countries is purely statistical, as the interpretation of κ values must take into consideration the prevalence of the assessed indicator in the study population (Hoehler, 2000).

By depicting the differences between the individual and group-level indicators and the differences between the two field studies, our results support that reliability is not an indicator's absolute attribute; reliability is rather an elaborated interaction between the indicator itself, the observers performing the assessment and the situation in which the assessment takes place (Streiner and Norman, 2008).

Table 5 *Inter-observer reliability for individual-level observations (*k *scores) computed from the assessments performed simultaneously by two observers in 10 Italian (IT) farms, whereas applying the Animal Welfare Indicators welfare assessment prototype protocol to dairy goats at the Feeding Rack (FR) and by Manually Restraining the goats (MR)*

	Reliability	
	κ (95% CI)	
Animal-based indicators	FR	MR
Udder asymmetry	0.57 (0.3 < CI < 0.84)	0.35 (0.12 < CI < 0.58)
Cleanliness		
Hindquarter	0.46 (0.34 < CI < 0.59)	0.68 (0.57 < CI < 0.78)
Lower legs	0.5 (0.38 < CI < 0.62)	0.63 (0.53 < CI < 0.74)
Udder	0.67 (0.48 < CI < 0.86)	0.53 (0.17 < CI < 0.89)
Lesions		
Hindquarter	n.a.	0.24 (−0.17 < CI < 0.64)
Lower legs	n.a.	0.27 (−0.17 < CI < 0.72)
Body	0.41 (0.12 < CI < 0.71)	0.33 (0.11 < CI < 0.55)
Neck	0.23 (−0.16 < CI < 0.63)	0.28 (−0.05 < CI < 0.62)
Head	0.22 (0.078 < CI < 0.37)	0.29 (0.18 < CI < 0.4)
Abscesses		
Hindquarter	n.a.	n.a.
Udder	0.38 (0.072 < CI < 0.69)	0.65 (0.39 < CI < 0.91)
Body	0.29 (0.0 < CI < 0.58)	0.45 (0.17 < CI < 0.72)
Neck	0.36 (0.074 < CI < 0.64)	0.43 (0.02 < CI < 0.84)
Head	0.59 (0.23 < CI < 0.95)	0.6 (0.28 < CI < 0.92)
Overgrown claws	0.55 (0.42 < CI < 0.69)	0.7 (0.6 < CI < 0.8)
Ocular discharge	n.a.	0.43 (0.025 < CI < 0.84)
Nasal discharge	0.28 (−0.05 < CI < 0.61)	0.4 (0.19 < CI < 0.61)
Vulvar discharge	n.a.	n.a.
Faecal soiling	0.42 (0.2 < CI < 0.65)	0.67 (0.56 < CI < 0.78)

CI = confidence intervals; n.a.: not possible to compute the information.
k: 0 to 0.40 = poor; 0.41 to 0.75 = fair to good; and 0.76 to 1 = excellent (Streiner and Norman, 2008).

Conclusions

Most of the group-level indicators included in the AWIN prototype protocol for dairy goats presented the highest IOR level both in Portugal and Italy, which was paramount for considering their inclusion in the first-level assessment of the AWIN final protocol. The IOR of individual-level indicators was lower, however in most cases the IOR levels reached were still lying between the 'fair to good' and 'excellent' thresholds, with the differences found between Portugal and Italy being mostly a result of the restraining method applied. Building on these results the AWIN prototype protocol for dairy goats was refined, which lead to the publication of the final AWIN on-farm welfare assessment protocol for adult dairy goats in intensive production systems.

Acknowledgements

The authors would like to thank the farmers who allowed us to visit their farms and two anonymous referees for their thorough and insightful comments on earlier versions of this paper.

Table 6 *Agreement and inter-observer reliability for the indicators body condition score (BCS) and Knee calluses (three scores)*

	Agreement		Reliability	
	Proportion of overall agreement (%) (95% CI)		κ_w (95% CI)	
Animal-based indicators	PT	IT	PT	IT
BCS	93.61 (90.43 <CI <95.82)	79.01 (74.23 <CI <83.12)	0.79 (0.70 <CI <0.88)	0.46 (0.34 <CI <0.58)
Knee calluses	96.11 (93.41 <CI <97.78)	90.96 (87.29 <CI <93.68)	0.79 (0.68 <CI <0.90)	0.27 (0.09 <CI <0.45)

CI = confidence intervals; PT = Portuguese; IT = Italian.
The table presents the proportion of overall agreement and weighted κ scores (k_w) computed from the assessments performed simultaneously by two observers in 10 PT and 10 IT farms, whereas applying the Animal Welfare Indicators welfare assessment prototype protocol to dairy goats.
k_w: 0 = poor; 0.00 to 0.20 = slight; 0.21 to 0.40 = fair; 0.41 to 0.60 = moderate; 0.61 to 0.80 = substantial; and 0.81 to 1 = almost perfect (Streiner and Norman, 2008).

The present document results from the Animal Welfare Indicators (AWIN) Project, which has been co-financed by the European Commission, within the VII Framework Program (FP7-KBBE-2010-4, grant no. 266213).

References

Bartussek H 1999. A review of the animal needs index (ANI) for the assessment of animals' well-being in the housing systems for Austrian proprietary products and legislation. Livestock Production Science 61, 179–192.

Battini M, Barbieri S, Vieira A, Stilwell G and Mattiello S 2016. Results of testing the prototype of the AWIN welfare assessment protocol for dairy goats in 30 intensive farms in Northern Italy. Italian Journal of Animal Science 15, 283–293.

Battini M, Vieira A, Barbieri S, Ajuda I, Stilwell G and Mattiello S 2014. Invited review: animal-based indicators for on-farm welfare assessment for dairy goats. Journal of Dairy Science 97, 6625–6648.

Botreau R, Veissier I, Butterworth A, Bracke M and Keeling L 2007. Definition of criteria for overall assessment of animal welfare. Animal Welfare 16, 225–228.

Brenner H and Kliebsch U 1996. Dependence of weighted kappa coefficients on the number of categories. Epidemiology 7, 199–202.

Burn CC and Weir AAS 2011. Using prevalence indices to aid interpretation and comparison of agreement ratings between two or more observers. Veterinary Journal 188, 166–170.

Can E, Vieira A, Battini M, Mattiello S and Stilwell G 2016. On-farm welfare assessment of dairy goat farms using animal-based indicators: the example of 30 commercial farms in Portugal. Acta Agriculturae Scandinavica, Section A – Animal Science 66, 43–55.

Caroprese M, Napolitano F, Mattiello S, Fthenakis GC, Ribó O and Sevi A 2016. On-farm welfare monitoring of small ruminants. Small Ruminant Research 135, 20–25.

Cohen J 1968. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin 70, 213–220.

De Rosa G, Grasso F, Pacelli C, Napolitano F and Winckler C 2009. The welfare of dairy buffalo. Italian Journal of Animal Science 8, 103–116.

Fraser D 1995. Science, values and animal welfare: exploring the 'inextricable connection'. Animal Welfare 4, 103–117.

Hallgren KA 2012. Computing inter-rater reliability for observational data: an overview and tutorial. Tutorials in Quantitative Methods for Psychology 8, 23–34.

Hewetson M, Christley RM, Hunt ID and Voute LC 2006. Investigations of the reliability of observational gait analysis for the assessment of lameness in horses. The Veterinary Record 158, 852–857.

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

Johnsen PF, Johannesson T and Sandøe P 2001. Assessment of farm animal welfare at herd level: many goals, many methods. Acta Agriculturae Scandinavica, Section A – Animal Science 51, 26–33.

Kaler J, Wassink GJ and Green LE 2009. The inter- and intra-observer reliability of a locomotion scoring scale for sheep. Veterinary Journal 180, 189–194.

Kaufman AB and Rosenthal R 2009. Can you believe my eyes? The importance of interobserver reliability statistics in observations of animal behaviour. Animal Behaviour 78, 1487–1491.

Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, Roberts C, Shoukri M and Streiner DL 2011. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. Journal of Clinical Epidemiology 64, 96–106.

Meagher RK 2009. Observer ratings: validity and value as a tool for animal welfare research. Applied Animal Behaviour Science 119, 1–14.

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

Mülleder C, Troxler J, Laaha G and Waiblinger S 2007. Can environmental variables replace some animal-based parameters in welfare assessment of dairy cows? Animal Welfare 16, 153–156.

Phythian CJ, Cripps PJ, Michalopoulou E, Jones PH, Grove-White D, Clarkson MJ, Winter AC, Stubbings LA and Duncan JS 2012. Reliability of indicators of sheep welfare assessed by a group observation method. Veterinary Journal 193, 257–263.

Phythian CJ, Toft N, Cripps PJ, Michalopoulou E, Winter AC, Jones PH, Grove-White D and Duncan JS 2013. Inter-observer agreement, diagnostic sensitivity and specificity of animal-based indicators of young lamb welfare. Animal: An International Journal of Animal Bioscience 7, 1182–1190.

Ruddat I, Scholz B, Bergmann S, Buehring A-L, Fischer S, Manton A, Prengel D, Rauch E, Steiner S, Wiedmann S, Kreienbrock L and Campe A 2014. Statistical tools to improve assessing agreement between several observers. Animal: An International Journal of Animal Bioscience 8, 643–649.

Scott EM, Nolan AM and Fitzpatrick JL 2001. Conceptual and methodological issues related to welfare assessment: a framework for measurement. Acta Agriculturae Scandinavica, Section A – Animal Science 51, 5–10.

Shrout PE 1998. Measurement reliability and agreement in psychiatry. Statistical Methods in Medical Research 7, 301–317.

Shrout PE and Fleiss JL 1979. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 89, 420–428.

Streiner DL and Norman GR 2008. Health Measurement Scales: A practical guide to their development and use, 4th Edition. Oxford University Press, New York, USA.

Uebersax J 2014. Raw agreement indices. Retrieved on 11 September 2017 from http://www.john-uebersax.com/stat/raw.htm.

Vieira A, Battini M, Ajuda I, Mattiello S and Stilwell G 2012. Set up of a sampling strategy for the collection of animal-based welfare indicators during milking. In Proceeding of the XI International Conference on Goats, 23–27 September 2012, Las Palmas, Gran Canaria, Spain, p. 51.

Vieira A 2015. Development and integration of animal-based welfare indicators, including pain, in goat farms in Portugal. PhD thesis, Universidade de Lisboa, Lisboa, Portugal.

Waiblinger S, Knierim U and Winckler C 2001. The development of an epidemiologically based on-farm welfare assessment system for use with dairy cows. Acta Agriculturae Scandinavica, Section A – Animal Science 51, 73–77.