UNIVERSIDADE DE LISBOA
Faculdade de Medicina Veterinária

DEVELOPMENT AND INTEGRATION OF ANIMAL-BASED WELFARE INDICATORS, INCLUDING PAIN, IN GOAT FARMS IN PORTUGAL

ANA CATARINA LOPES VIEIRA GODINHO DE MATOS

Orientadores: Prof. Doutor George Thomas Stilwell
Prof. Doutora Mónica Duarte Correia de Oliveira

Tese especialmente elaborada para obtenção do grau de Doutor em Ciências Veterinárias, Especialidade de Clínica

2015
LISBOA

UNIVERSIDADE DE LISBOA

Faculdade de Medicina Veterinária

DEVELOPMENT AND INTEGRATION OF ANIMAL-BASED WELFARE INDICATORS, INCLUDING PAIN, IN GOAT FARMS IN PORTUGAL

ANA CATARINA LOPES VIEIRA GODINHO DE MATOS

Orientadores: Prof. Doutor George Thomas Stilwell
Prof. Doutora Mónica Duarte Correia de Oliveira

Tese especialmente elaborada para obtenção do grau de Doutor em Ciências Veterinárias, Especialidade de Clínica

Júri:

Presidente: Doutor Rui Manuel de Vasconcelos e Horta Caldeira

Vogais:

- Doutora Silvana Mattiello
- Doutor Severiano José Cruz da Rocha e Silva
- Doutor Miguel Luís Mendes Saraiva Lima
- Doutora Mónica Duarte Correia de Oliveira
- Doutor George Thomas Stilwell
- Doutora Maria Leonor Santos Galhardo

2015

LISBOA

Aos meus filhos, ao meu marido e aos meus pais

AGRADECIMENTOS

Ao chegar ao fim deste percurso, é impossível não olhar para trás e agradecer a um conjunto de pessoas e instituições por todo o apoio necessário para a conclusão desta tese de doutoramento.

O meu Obrigada:

Ao meu orientador, Professor George Stilwell, e à minha co-orientadora, Professora Mónica Oliveira. Professor Stiwell, foi uma honra ter sido a sua primeira aluna de doutoramento. Professora Mónica, sendo de uma área científica tão diferente, obrigada por ter aceite orientar a minha tese, aprendi muito consigo. Obrigada por todo o apoio e disponibilidade, por todas as ideias e, também, por todas as críticas.

Ao Telmo Nunes, por toda a partilha de ideias, por toda a orientação e por toda a ajuda. Muito do que fomos conversando está espalhado por estas páginas.

A todos os investigadores do projecto AWIN, e em especial à Professora Silvana Mattiello e à Monica Battini. *Grazie per tutto il lavoro che abbiamo sviluppato insieme. Ho imparato molto da voi.* E também à Francoise Wemelsfelder. *Thank you Francoise for all our inspirational discussions around measurement.*

Aos Professores Graça Pires de Melo, Miguel Saraiva Lima, Rui Bessa e Rui Caldeira, pela constante disponibilidade e comentários sempre úteis. Aos Professores António Monteiro e Severiano R. Silva, pelos comentários e ajuda tão importantes no desenvolvimento da escala visual de condição corporal.

À Susana Brandão, por todo o trabalho que desenvolvemos para a construção da escala visual de condição corporal. Obrigada por teres aceite o desafio e por tanto que aprendi contigo.

A todos os produtores que nos abriram as portas das suas explorações e das suas casas. Sem o seu interesse este trabalho não teria nunca sido possível. E aos colegas Álvaro Lopes, Ana Caria, Ana Rita Simões, Bonacho Costa, Hélder Quintas e Miguel Madeira, pelos contactos que estabeleceram com alguns dos produtores e por nos terem acompanhado nas visitas. Obrigada pelo vosso interesse.

À Direcção Geral de Alimentação e Veterinária, nas pessoas da Dra. Maria Jorge Correia e da Dra. Ana Paula Martins. Ao Queijo Saloio, na pessoa do Enj. João Aleixo. Obrigada por toda a ajuda e por todo o vosso interesse.

Aos colegas Carla Marmelo, Felipe Fernandes de Almeida, Inês Barão, Mónica Fonseca, Sara Martins, Uriel Mateus, por toda a ajuda com a recolha de dados nos diferentes projectos que constituem esta tese. Foi sempre muito divertido trabalhar convosco.

Ao José Chumbo da Virtualmente.pt, obrigada pela ajuda com o desenvolvimento e implementação do web-survey. E obrigada pela paciência quanto te pedia as "listagens".

Ao Pedro Salgado. Obrigada pelo desenho científico. Foi uma colaboração fantástica.

À Inês Ajuda, minha colega de doutoramento. Obrigada pelo apoio e pela partilha. Desejo-te toda a força para a última etapa do teu doutoramento. Tenho a certeza que vais ser bem sucedida!

À Edna Can, a “minha co-orientanda”. Tenho a certeza que aprendi muito mais contigo, do que tu comigo. Obrigada Edna, por tudo!

À Inês Viegas, ao Rodolfo Oliveira Leal e à Ana Paula Resende. Obrigada pela amizade, pela companhia e pelas pausas para café. Oxalá nos venhamos a reencontrar!

Aos meus pais, por todo o apoio. Sem eles nada, desde sempre, tinha sido possível. Ao Miguel, meu marido, obrigada por toda a ajuda e paciência. Eu sei que às vezes foi preciso ter muita! Tenho a certeza que és o professor de *Business Analytics* que mais sabe sobre bem-estar animal. Aos meus filhos, Miguel e Tomás, que dentro da minha barriga visitaram explorações de cabras de Norte a Sul de Portugal.

APOIOS FINANCEIROS

O presente trabalho foi financiado no âmbito do projecto Europeu com a referência "Development, integration and dissemination of animal-based welfare indicators, including pain, in commercially important husbandry species, with special emphasis on small ruminants, equidae and turkeys (KBBE.2010.1.3-03: 7th Framework Programme (# 266213)".

Development and integration of animal-based welfare indicators, including pain, in goat farms in Portugal

ABSTRACT

The development of species-specific protocols for on-farm welfare assessment represents a major concern for the European Union agricultural policy. Proper welfare assessment demands for the use of valid, reliable and feasible animal-based welfare indicators. The literature and policy-makers recognise the need for advancing knowledge in this area. This thesis aims at developing and integrating animal-based indicators in on-farm welfare assessment protocols, focusing on dairy goat farms.

Specifically, this thesis contributes to literature in three areas. First, it contributes to the identification of animal-based welfare indicators that should be included in welfare assessment protocols. We conducted a literature review that allowed for the recognition of the need for future research in the indicators' psychometric properties, such as reliability and feasibility. Secondly, this thesis develops tools to assist the measurement of body condition and lameness. For body condition, we developed a visual body condition scoring system (BSC). Our approach requires minimum animal handling without compromising a valid and reliable individual assessment of the goats. With respect to lameness we developed a web-survey that allowed us to collect observer's ratings of goats lameness condition. Our survey showed that observers were only able to consistently assess severely lame goats, a finding which is important towards the integration of the indicator in assessment protocols. The observers' ratings also showed that the numerical rating scales should only be used considering their ordinal level of measurement. This directs research towards the development of scoring systems with higher levels of measurement, like the modified visual analogue scales. Third, this thesis contributed to the development of a welfare assessment protocol that integrated and tested the two studied indicators (BCS and lameness). Such protocol was implemented in 30 Portuguese farms and provided insights into the main welfare problems affecting intensively kept dairy goats in our country (claw overgrowth, queuing at feeding, very fat animals), which is paramount to improve dairy goats' welfare.

Research conducted for this thesis has practical implications for both welfare assessment research and to the goat industry in general. Ultimately, through the development of adequate assessment tools, it integrates the welfare issue into the food chain, meeting the consumers' expectations in the development of a sustainable food production system.

KEYWORDS: on-farm welfare assessment; dairy goat; body condition score; lameness; web-survey

Desenvolvimento e integração de indicadores de bem-estar animal, incluindo dor, em explorações de cabras em Portugal

RESUMO

A elaboração de protocolos de avaliação de bem-estar específicos para cada espécie pecuária é uma preocupação da política agrícola europeia. A literatura da área de bem-estar animal identifica a criação de instrumentos de medição como o primeiro passo para a elaboração destes protocolos. Esta tese tem como objetivo desenvolver e integrar indicadores para incluir em protocolos de avaliação para utilizar em explorações de cabras de aptidão leiteira.

Esta tese apresenta três contributos para a literatura de bem-estar animal. Em primeiro lugar, contribui para a identificação de indicadores, baseados nos animais, com potencial para integração em protocolos de avaliação. A revisão bibliográfica realizada permitiu reconhecer a necessidade premente de investigação nesta área, dado que a maior parte dos indicadores necessitam de ser testados e validados. Em segundo lugar, esta tese desenvolve ferramentas para apoiar a avaliação da condição corporal e da claudicação. Para a condição corporal foi criado um sistema visual de avaliação considerado válido e repetível, e que apenas necessita de uma breve contenção dos animais para ser utilizado. Relativamente à claudicação foram recolhidas participações de observadores relativamente à observação de vídeos de cabras com diferentes níveis de claudicação. A análise destas observações permitiu concluir que os participantes apenas são consistentes a avaliar os casos mais graves de claudicação, facto importante para a integração do indicador em protocolos de avaliação. As classificações dos observadores mostraram ainda que as escalas numéricas em uso apenas podem ser utilizadas considerando um nível ordinal de medição. Este facto abre o caminho para o desenvolvimento de escalas com níveis mais elevados de medição, como as escalas visuais analógicas modificadas. Em terceiro lugar, esta tese desenvolve um protocolo de avaliação que inclui e testa os indicadores condição corporal e claudicação. Este protocolo permitiu investigar sobre os maiores problemas de bem-estar que afetam as explorações intensivas de leite de cabra em Portugal (sobre crescimento das unhas, filas na manjedoura, animais gordos), sendo esta informação fundamental para analisar como melhorar o bem-estar das cabras de leite.

A investigação conduzida no âmbito desta tese apresenta implicações práticas tanto para o estudo do bem-estar animal, como para a exploração de leite de cabra. O desenvolvimento de ferramentas adequadas de avaliação permite a integração da valoração do bem-estar na cadeia de produção, indo ao encontro das expectativas dos consumidores para a concepção de sistemas mais sustentáveis de produção de alimentos.

PALAVRAS-CHAVE: avaliação de bem-estar animal nas explorações; cabra de leite; condição corporal; claudicação; *web-survey*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABREVIATIONS

ACTH	Adrenocorticotropic Hormone
AD	Avoidance Distance
AIC	Akaike Information Criterion
AWIN	Animal Welfare Indicators
BCS	Body Condition Scoring
CAB	Commonwealth Agricultural Bureaux
CAE	Caprine Arthritis and Encephalitis
COT	Consistency of indicators Over Time
DGAV	Direcção Geral de Alimentação e Veterinária
DGSANCO	General Directorate for the Health and Consumer Protection
DF	Degrees of Freedom
DL	Decreto-lei
EFSA	European Food Safety Authority
EU	European Union
FAO	Food and Agriculture Organization of the United Nations
FAOSTAT	Food and Agriculture Organization of the United Nations Statistics Division
FAWC	Farm Animal Welfare Council
HAR	Human-animal Relationship
ICC	Intra-class Correlation
IP	Internet Protocol
INE	Instituto Nacional de Estatística
NRS	Numerical Rating Scale
PC	Principal Component
PCA	Principal Component Analysis
QBA	Qualitative Behaviour Assessment
RGB-D	Red Green Blue - Depth
RSPCA	Royal Society for the Prevention of Cruelty to Animals
SCC	Somatic Cell Count
SD	Standard Deviation
SE	Standard Error
SIFT	Scale-Invariant Feature Transform
SPSS	Statistical Package for the Social Sciences
SQL	Structured Query Language
TGI	Tiergerechtheitsindex
UK	United Kingdom
USA	United States of America
VAS	Visual analogue scale
WP	Work Package

LIST OF PUBLICATIONS

This thesis was based on the following manuscripts (international peer-reviewed papers):

Battini, M., **Vieira, A.**, Barbieri, S., Ajuda, I., Stilwell, G. & Mattiello, S. (2014). Invited review: Animal-based indicators for on-farm welfare assessment for dairy goats. *Journal of Dairy Science*, 97(11), 6625-6648. http://dx.doi.org/ 10.3168/jds.2013-7493.

Vieira, A., Brandão, S., Monteiro, A., Ajuda, I., & Stilwell, G. (2015). Development and validation of a visual body condition scoring system for dairy goats with picture-based training. *Journal of Dairy Science*, 98(9), 6597-6608. http://dx.doi.org/ 10.3168/jds.2015-9428

Vieira, A., Oliveira, M.D., Nunes, T. & Stilwell, G. *Design and test of a web-survey for collecting observe's ratings on lameness scoring data*. Manuscript submitted for publication.

Vieira, A., Oliveira, M.D., Nunes, T. & Stilwell, G. (in press) *Making the case for developing alternative lameness scoring systems for dairy goats*. (Accepted at *Applied Animal Behaviour Science)*

Can, E., **Vieira, A.**, Battini, M., Mattiello, A. & Stilwell, G. *On-farm welfare assessment of dairy goat farms using animal-based indicators: The example of 30 commercial Portuguese farms*. Manuscript submitted for publication.

The research work also resulted in the following posters and oral presentations:

Vieira, A., Can, E. & Stilwell, G. (2015, March). *Applying the AWIN welfare assessment prototype to Portuguese dairy goat farms – preliminary results*. Poster session presented at the 7th IDF International Symposium on Sheep, Goat and other non-Cow Milk, Limassol, Cyprus.

Vieira, A., Amaral, R., Nunes, T., Ajuda, I., Stilwell, G. & Oliveira, M. (2014, September). *A multicriteria decision analysis model for the selection of welfare strategies in dairy goat farms*. Poster session presented at the 6th International Conference on the Assessment of Animal Welfare at Farm and Group Level, Clermont-Ferrand, France.

Vieira, A. "A elaboração de um protótipo de avaliação de bem-estar – as principais dificuldades e soluções." Avaliação de bem-estar animal em animais de produção - PsiAnimal. Benavente, IN. 26-27 April 2014. Conference Presentation.

Vieira, A., Battini, M., Mattiello, S. & Stilwell, G. "A proposal for a new welfare assessment protocol for goats". European Regional Meeting on Goats. International Goat Association. Debrecen, Hungary, 8-11 April 2014. Conference Presentation.

Vieira, A., Battini, M., Mattiello, S. & Silwell, G. "Selecção dos indicadores mais apropriados para a avaliação do bem-estar de cabras de leite em regime intensivo". VI Congresso Sociedade Portuguesa de Ciências Veterinárias. Oeiras, IN. 3-5 April 2014. Conference Presentation.

Vieira, A., Battini, M., Ajuda, I., Mattiello, S. & Stilwell, G. 2013. [abstract] [electronic version]. Is the collection of animal-based welfare indicators during milking affected by the order of goats entry into the milking parlor? In *Proceedings UFAW International Animal Welfare Science Symposium, Barcelona, Spain, July, 4-5*, p. 100. Retrieved from http://www.ufaw.org.uk/UFAW%202013%20Symposium%20abstract%20booklet%20final%20v3.pdf

Vieira, A., Battini, M., Ajuda, I & Stilwell, G. "Indicadores de bem-estar em cabras de leite e locais de eleição para a sua avaliação". 2º Congresso da Associação Portuguesa de Terapia do Comportamento e Bem-Estar Animal - PsiAnimal. Lisboa, IN. 29-30 September 2012. Conference Presentation.

Vieira, A., Battini, M., Ajuda, I., Mattiello, S. & Stilwell, G. "Set up of a sampling strategy for the collection of animal-based welfare indicators during milking". XI ICG 2012. Gran Canaria, Spain, IN. 24-27 September 2012. Conference Presentation.

The research work conduted in this thesis also contributed to the following manuscripts (international peer-reviewed papers), posters, and oral presentations:

AWIN (2015). *AWIN welfare assessment protocol for goats*. doi: http://dx.doi.org/ 10.13130/AWIN_goats_2015.

Battini, M., Stilwell, G., **Vieira, A.**, Barbieri, S., Canali, E. & Mattiello, S. (2015) On-farm welfare assessment protocol for adult dairy goats in intensive husbandry systems. *Animals*, 5, 934-950. http://dx.doi.org/10.3390/ani5040393

Battini, M., Peric, T., Ajuda, I., **Vieira, A**., Grosso, L., Barbieri, S., Stilwell, G., Prandi, A., Comin, A., Tubaro, F. & Mattiello, S. (2015). Hair coat condition: A valid and reliable indicator for on-farm welfare assessment in adult dairy goats. *Small Ruminant Research*, 123(2-3), 197–203. http://dx.doi.org/10.1016/j.smallrumres.2014.12.009

Stilwell, G., Battini, M, **Vieira, A.**, Barbieri, S., Canali, E., Zanella, A. & Mattiello, S. "On farm welfare assessment protocol for adult female dairy goats in intensive husbandry systems". 7th IDF International Symposium on sheep, goat and other non-cow milk. Limassol, Cyprus, IN. 23-25 March, 2015. Conference Presentation

Mattiello, S., Battini, M., Barbieri, S., **Vieira, A.,** Stilwell, G., Canali, E. & Zanella, A. (2014). Un protocollo per la valutazione del benessere nella capra: I risultati del progetto AWIN. *Large Animal Review*, 1(4), 90 – 93. Retrieved from http://www.sipaoc.it/documenti/2014/ XXI%20Congresso%20SIPAOC%202014%20%20Atti%0congressuali.pdf

Ajuda, I., **Vieira, A.** & Stilwell, G. (2014). Are there differences in dairy goats claws' temperature, before and after trimming? [electronic version]. In *2014 IEEE International Symposium on Medical Measurements and Applications (MeMeA), Lisbon, Portugal, 11-12 June,* pp. 1–5. http://dx.doi.org/10.1109/MeMeA.2014.6860149.

Ajuda, I., **Vieira, A.** & Stilwell, G. "How does trimming affect the animal and welfare in intensive bred goats with overgrown hooves? – Preliminary results". European Regional Meeting on Goats. International Goat Association. Debrecen, Hungary, IN. 8-11 April 2014. Conference Presentation.

Ajuda, I., **Vieira, A.** & Stilwell, G. (2013, September). *Pain in disbudding kids, is there a practical solution? Preliminary results*. Poster session presented at the 31st World Veterinary Congress (WVC), Prague, Czech Republic.

Ajuda, I., **Vieira, A.**, de Almeida, F.F. & Stilwell, G. (2013, August). *Lameness on dairy goats: What is the level of pain experienced and how can we evaluate it? Preliminary results*. Poster session presented at the 17^{th} International Symposium and 9thInternational Conference on Lameness in Ruminants, Bristol, UK.

Ajuda, I.G., **Vieira, A**., de Almeida, F. & Stilwell, G. "Conformation of the udder, is that a problem in our dairy farms? Preliminary results". Goat Milk Quality conference. Tromsø, Norway, IN. 4 – 6 June 2013. Conference Presentation.

Battini, M., Barbieri, S., Ajuda, I., **Vieira, A.**, Stilwell, G. & Mattiello, S. "Le condizioni del mantello sono un valido indicatore di benessere nella capra?". XX Congresso Nazionale S.I.P.A.O.C. Siracusa, Italy, IN. 26-29 September, 2012. Conference Presentation.

Ajuda, I., **Vieira, A.**, Battini, M. & Stilwell, G. (2012, September). *Deformation and overgrowth of the hooves, is it a problem in intensive dairy goat farms?* Poster session presented at the XI ICG 2012, Gran Canaria, Spain.

Battini, M., **Vieira, A.**, Ajuda, I., Mattiello, S. & Stilwell, G. (2012, September). *Identifying promising animal-based welfare indicators in intensively bred dairy goats*. Poster session presented at the XI ICG 2012, Gran Canaria ,Spain.

"I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the state of Science, whatever the matter may be."

Lord Kelvin, 1883

PART I

General introduction

The general introduction starts with the context and drivers for this research thesis, followed by the objectives and thesis structure. To comprehensively understand the focus of the work conducted in this research, the first part is organized into four more chapters that present a brief literature review. Along chapters two to four, the underlying theoretical and methodological concepts of this research work are presented. The general introduction finishes with chapter five that presents some data on the species under assessment- dairy goats.

CHAPTER 1

Objectives, and thesis structure

Chapter 1 presents the Animal Welfare Indicators project, along with the drivers for this research thesis. It finishes with the overall and main objectives, followed by the thesis structure.

The present PhD thesis aims at developing and integrating animal-based welfare indicators, in dairy goat farms, and was conducted within the scope of a collaborative, large-scale European project with the reference: "Development, integration and dissemination of animal-based welfare indicators, including pain, in commercially important husbandry species, with special emphasis on small ruminants, equidae and turkeys (KBBE.2010.1.3-03: 7th Framework Programme (# 266213)" that started in 2011. We thus start by describing this project and will then frame the objectives of this research thesis.

The Animal Welfare Indicators (AWIN) project engaged 10 Institutions in nine countries and approached animal welfare indicators in four separate, but complementary, Work Packages (WP). Work Package 1 aimed at developing and integrating animal welfare indicators, including pain indicators, in order to set welfare assessment protocols for horses, donkeys, turkeys, sheep and goats. Work Package 2 studied the relationship between diseases, including pain, and animal welfare, WP3 examined the influence of prenatal and early--postnatal environments and handling methods on welfare and health of pregnant females and their offspring. Finally, WP4 was one of the distinctive aspects of the project as it focused on animal welfare education and dissemination of science to farmers, animal owners, stakeholders and other interested parties. The objective was to construct global networks of excellence, being the ultimate example, the creation of the Animal Welfare Science Hub (www.animalwefarehub.com) that incorporates research and learning materials, derived from deliverables of the diverse WPs.

The research conducted in this thesis was part of WP1 (Welfare, including pain, assessment) and addressed, in cooperation with another research team from the University of Milan, the development of an on-farm welfare assessment protocol for dairy goats. In order to meet WP1 objectives regarding the dairy goats, the two Universities designed an action plan that will be briefly described.

The starting point was to identify promising animal-based indicators that could be included in a welfare assessment protocol for goats[1], based on a comprehensive review of the existing scientific literature (whereby it was also reviewed available data on their psychometric proprieties[2]). The criteria for inclusion of these indicators was discussed in meetings with experts[3]. After drafting an initial list of relevant indicators, consensus had to be reached among experts about categories of animals to be assessed. This was extremely important as the welfare of animals in different types of production is highly dependent on the particular details and management of the system.

Taking into account that intensive livestock production is continuously growing across different farmed species, and it is expected to continue to do so (Cronin, Rault & Glatz, 2014), focus was given to this production system[4]. Additionally, the development of goat husbandry is fairly more intense under the extreme settings of very intensive and very extensive systems (Boyazoglu et al., 2005; Morand-Fehr et al., 2004). Moreover, this is the system where concerns from consumers more frequently arise, so we would be tackling and anticipating future demand.

A research action plan to tackle the lack of knowledge concerning the validity, repeatability and feasibility of the indicators was thus drafted within the scope of WP1. In this context, research conducted in the Animal Behaviour and Welfare Lab., Centre for Interdisciplinary Research in Animal Health, of the Faculdade de Medicina Veterinária, focused on the development of indicators that were classified according to the "welfare principles" of "good feeding", "good housing", and "good health", whereas research conducted in the University of Milan focused mainly on behavioural indicators.

The newly developed indicators and the already established ones were integrated into a welfare assessment prototype protocol that was tested for its overall feasibility in 60 goat farms located in Portugal and Italy. Regarding the welfare assessment protocols, the WP1 strategy across species was to propose a stepwise strategy for on-farm welfare assessment, with the protocol suggesting, as a first level, a quick screening comprising a selection of robust and feasible indicators that could be readily recorded. This assessment could then advance into a longer and more in-depth evaluation.

During the course of the project, all WP1 researchers encouraged a participatory approach through the involvement and the collaboration with stakeholders, in order to assure the effectiveness and sustainability of the final assessment protocols. This direct contact between researchers and stakeholders, as pointed by Morand-Fehr et al. (2004), is a main requisite to applied research.

Within the scope of the AWIN project, the research objectives set for this thesis will now be clarified.

---

[1] The reasons why the AWIN project focused on animal-based indicators in detriment of environmental-based ones are discussed on chapter three.
[2] In chapter four extensive attention is given to these basic, but fundamental, concepts.
[3] The result of this work is presented in chapter six.
[4] Additional data to understand the growing interest in this species is presented in chapter five.

1.1 THESIS OVERALL AND MAIN OBJECTIVES

The overall objective of this research thesis is to develop and integrate animal-based welfare indicators, including pain, in dairy goat farms.

In order to accomplish this overall objective, three main objectives were set:

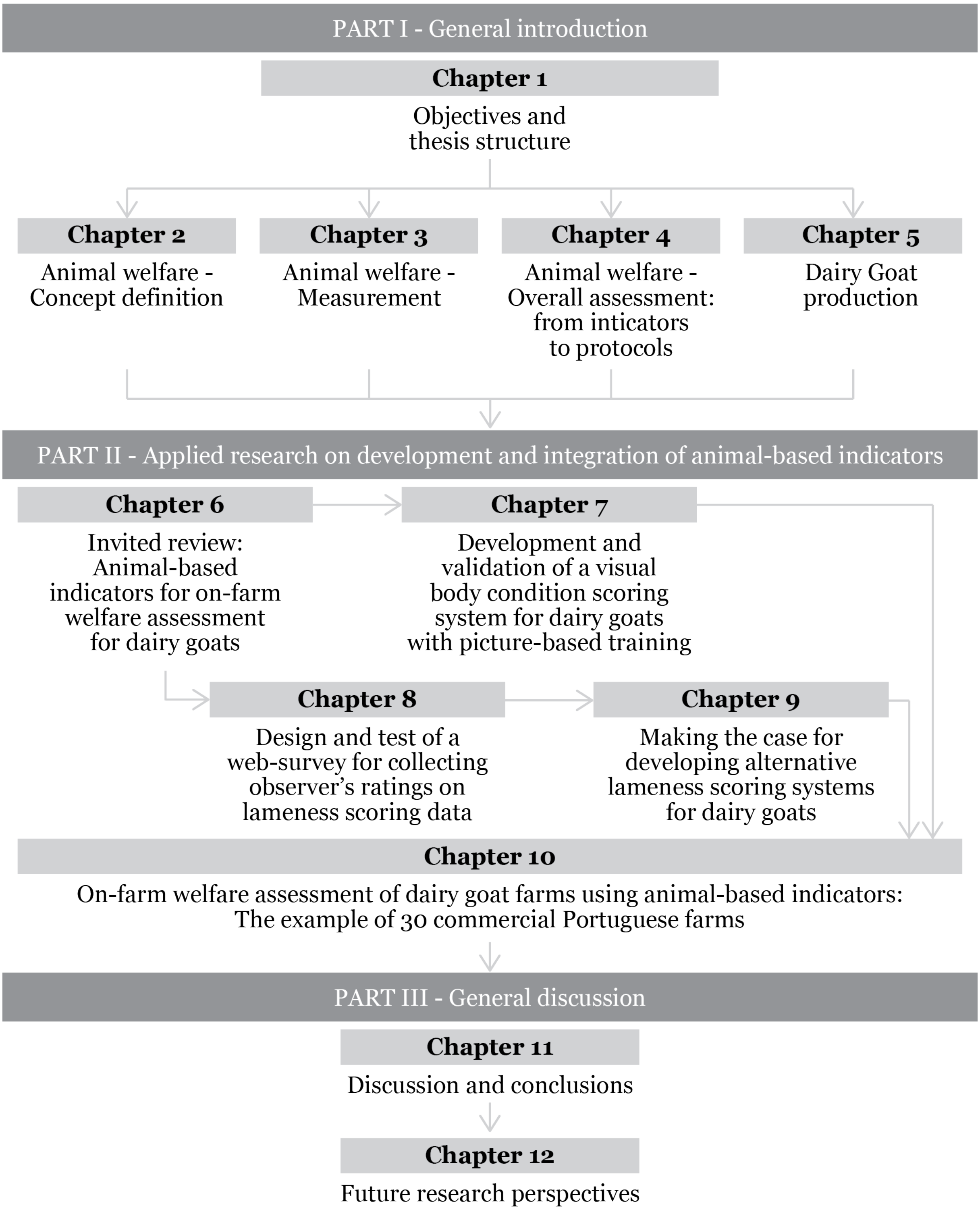
1. The identification – by means of an extensive and comprehensive literature review – of the animal-based welfare indicators that had inherent value to be included in welfare assessment protocols, but needed further scientific support;
2. The development of animal-based indicators based on the assessment of their psychometric proprieties and the level of measurement needed – the research conducted in this thesis focused particularly on the indicators body condition scoring (BCS) and lameness. This was due mainly to three reasons:
   a) the importance associated with these two indicators (Lievaart & Noordhuizen, 2011). Body condition scoring is considered a valid welfare indicator across different species not only for the "good feeding", but also for the "good health" principle. Nevertheless, for goats traditional methods lack the feasibility needed to be included in welfare assessment protocols, as it needs further reliability studies. Lameness, as an important indicator of pain, is frequently addressed in welfare assessment protocols regarding the "good health" principle. However, no scoring systems for goats' lameness, including its psychometric proprieties, have been established;
   b) collaborative studies could be developed in Portugal by having WP2 conducting studies that requested the use of these two indicators. For instance, the lameness work was in part validated by WP2 comprehensive studies on claw overgrowth and deformation, and the work conducted for the development of the visual BCS system demonstrated to be valuable for WP2 studies on pregnancy toxaemia;
   c) the fact that the hosting institution is a Veterinary School with all the available expertise for the indicator's validation. And this PhD thesis had the supervision of a researcher of the Centre for Management Studies of Instituto Superior Técnico that supervised the formal aspects of scoring system development, which represented scientific advances in this field of knowledge.
3. The integration and testing of the animal-based indicators in welfare assessment protocols and their use to point on-farm prevalence and tendencies, as a basis for setting thresholds of acceptability. The 30 farms visited in Portugal were entirely commercial farms, which showed great advantages concerning the validation of indicators and the feasibility of the protocol[5].

[5] The results of the farm visits are presented on chapter ten.

1.2 THESIS STRUCTURE AND SPECIFIC OBJECTIVES

Figure 1 depicts the thesis structure, showing the workflow followed during the research process. The following paragraphs assist in reading Figure 1, introducing briefly each chapter of this research thesis and highlighting each chapter specific objective and contribution to the above mentioned objectives.

Figure 1. Thesis structure



Part I is dedicated to the General introduction and is organized in five chapters. More than introducing the different subjects of research, these chapters provide a brief literature review on the concepts needed to comprehensively understand the research undergone in this thesis. The main and specific objectives of each chapter are as follows:

Chapter 1 presents the research context, the objectives, and thesis structure.

Chapter 2 takes the reader through the emergence of animal welfare as a scientific discipline, highlighting the importance society had in its establishment. Moreover, it introduces and debates the state of the art for the animal welfare concept.

Chapter 3 introduces the two main categories of indicators used to measure animal welfare on farms. It focuses mainly on animal-based indicators and the challenges posed. This review is paramount to explain the options made in the two first main objectives of this thesis.

Chapter 4 elucidates the reader on how animal welfare scientists generally go from indicators to protocols, aiming for an overall assessment of welfare. Furthermore, it introduces and discusses the basic concepts for the indicators development that were behind all the fieldwork conducted during this research, and presents current forms of animal welfare control.

Chapter 5 presents some facts and figures regarding dairy goat farming.

Part II presents the research conducted to fulfil the objectives above mentioned and is divided into five chapters. All the five chapters have been published, are "in press" or have been submitted to peer-reviewed journals.

Chapter 6 presents the result of the first main objective of this thesis. This study was developed in cooperation with the University of Milan. As already mentioned the specific objective was to review promising animal-based indicators that could be used to set up a valid, reliable, feasible, and practical on-farm welfare assessment protocol for dairy goats, centred on the evaluation of lactating animals. Most of the reviewed literature dealt with dairy goats; however, although lactating dairy goats are our main target, papers considering other productive categories (e.g., kids, dry goats) or goats farmed for different purposes, or even other species, were taken into account whenever they provided evidence to support the use of indicators that could be included in an on-farm welfare assessment protocol for lactating dairy goats. We excluded indicators that focused exclusively on resources and management, as well as animal-based indicators that require further laboratory analysis, or routinely collected herd data.

Chapters 7, 8 and **9** cover the second main objective of this research, the development of BCS and lameness as animal-based indicators.

Chapter 7 describes the development of a new scoring system to assess body condition in goats. In this chapter we address the feasibility challenges of including this indicator on on--farm welfare assessment protocols, and present an innovative approach to the development of a visual BCS system for adult dairy goats. Moreover, we present a training program using the new scoring system based on the concept of threshold images, and assess the inter--observer reliability of the scoring system under field conditions.

Chapters 8 and **9** are related with the development and integration of the lameness indicator. In chapter eight we present and discuss the use of web-surveys as an alternative method for data collection in animal lameness scoring. In chapter nine we start by discussing distinct drawbacks that have been associated with different types of scoring systems traditionally used to assess lameness in goats, and then move to assess the grounds for developing a modified VAS to assess lameness in goats. The strategy followed allowed us to identify the level of measurement that could be set to assess lameness considering the constraints dairy goat farms present.

Chapter 10 covers the third main objective of this research, and presents the results of the prevalence of the indicators in the on-farm welfare assessment protocol prototype developed within the AWIN project, considering 30 dairy goat farms visited in Portugal. It was important not only to assess the integration of the indicators developed in the previous objective, but also to gain knowledge on their overall importance in Portuguese farms.

Part III is dedicated to the general discussion and future research perspectives.

Chapter 11 focuses on the main theoretical and methodological implications findings of this research, integrating the knowledge developed throughout all the previous chapters and highlighting the practical implications of this research. Moreover, it discusses some of the limitations of this research work.

Chapter 12 debates some future research perspectives.

CHAPTER 2

Animal welfare – Concept definition

Chapter 2 discusses the emergence of animal welfare as a scientific discipline, and debates the state of the art animal welfare concept. This introductory chapter is important to recognize animal welfare as a multidimensional concept.

Intensive farming has its origin after the II World War (Cronin et al., 2014). In 1945 Europe was in deep food deprivation as its production systems had been destroyed, and its population moved from the fields to the cities where it was not possible to practice a self-sustainable agriculture. These situations led to a scarcity of food from animal origin, which made the price of food very high for the general population. Hence, strategies had to be put in place that would allow producing a large amount of food, at reasonable prices, and without depending too much on human labour.

Farms became larger and more specialized, and there was an enormous increase in total number of production animals and of stock density (Blokhuis, Miele, Veissier & Jones, 2013). Consequently, there were profound changes in facilities and introduction of mechanization and automation, which changed profoundly the management of the animals (Cronin et al., 2014; Keeling, 2005). This trend continued to our days as governments promoted a series of changes in farm structural and enterprise characteristics, encouraging constant rises in the number of animals per farm (Cronin et al., 2014; Winter, Fry & Carruthers, 1998).

In parallel, scientific research, enhanced by governments and supported by industries from the pharmaceutical and nutrition fields, developed knowledge towards the maximization of production, essentially by improving animal health, nutrition and genetics.

For cultural, behavioural, and commercial reasons there was a gap of communication between those that produced food and those who ate it, making consumers too distant from all these changes (Blokhuis et al., 2013). In the sixties, two books started to close this gap: *Silent Spring* by Rachel Carson (1962) and *Animal Machines: The New Farming Industry* by Ruth Harrison (1964) (with preface written by Rachel Carson). The two books had a tremendous impact in society promoting its own conscience and perhaps more importantly the awareness of the scientific community. By a very vivid and clear description of what was production intensification and its consequences, both authors brought these questions to the societal debate. It was the beginning of the consumer-citizen concept.

Particularly the book *Animal Machines: The New Farming Industry* began an animal welfare debate that culminated in 1965 when Professor Roger Brambell chaired a technical committee assigned by the British Government to examine these issues (Keeling, 2005). The "Report of the Technical Committee to Enquire into the Welfare of Animals Kept under Intensive Livestock Husbandry Systems", also known as "The Brambell Report", (Brambell Committee, 1965) established the "Brambell's Five Freedoms", stating that animals should have the freedom "to stand up, lie down, turn around, groom themselves and stretch their limbs". The Committee not only recommended that new general and specific legislation was needed to protect the welfare of farm animals, but also that a statutory advisory committee on farm animal welfare should be created. As a result, the Farm Animal Welfare Advisory Committee (FAWAC) was created in 1967, to monitor the livestock production sector, and statutory provisions for the welfare of livestock were considered in the Agriculture (Miscellaneous Provisions) Act 1968. Later, in 1979, the Farm Animal Welfare Council (FAWC) replaces the FAWAC and codifies the "Brambell's Five Freedoms" into the "Five Freedoms" concept[6], the basis for all future animal welfare assessment.

Thus, to fully understand the current concept of animal welfare, it is very important to consider that animal welfare, as a scientific discipline, owes its existence to society ethical concerns and not to scientists' curiosity or findings (Fraser, Weary, Pajor & Milligan, 1997; Keeling, 2005). For this reason, animal welfare analysis is composed of two perspectives: an ethical (focusing on values and judgment), and a scientific one (focused on facts), and these two should be clearly separated (Broom, 1991). In the research conducted in this thesis we follow the scientific component perspective of animal welfare.

2.1 ANIMAL WELFARE AS A MULTIDIMENSIONAL CONCEPT

Animal welfare science has developed in the last years giving special attention to welfare indicators (Fraser, 2006; Veissier, Butterworth, Bock & Roe, 2008), and to the development of species-specific protocols for on-farm welfare assessment.

Measurement is an fundamental component of scientific research (Streiner & Norman, 2008), and one of the rules before measuring something is having a clear definition of what is being measured (Nunnally & Bernstein, 1994). Temple Grandin usually says "you can't manage what you can't measure", also thinking of welfare. So, it is clear that the concept of what is being measured has a strong influence on the type of research that is conducted

[6] Initially, 1) Freedom from hunger and thirst, by ready access to water and a diet to maintain health and vigour; 2) Freedom from discomfort, by providing an appropriate environment; 3) Freedom from pain, injury and disease, by prevention or rapid diagnosis and treatment; 4) Freedom to express normal behaviour, by providing sufficient space, proper facilities and appropriate company of the animals' own kind; 5) Freedom from fear and distress, by ensuring conditions and treatment that avoid mental suffering.

(Duncan & Fraser, 1997; reviewed by Fraser et al., 1997). For this reason before moving on to the more methodological and theoretical concepts underlying this research, it is important to clearly state the “animal welfare” definition adopted.

Being a concept intrinsically related to culture and beliefs, the interpretation of what is animal welfare differs widely between countries and people (Vanhonacker, Verbeke, Van Poucke & Tuyttens, 2007), with science being essential to objectively define it (Fraser et al., 1997). However, even scientists have presented different definitions. From literature it is possible to identify at least three main approaches to define animal welfare.

The first approach focuses on “natural-living” concerns and highlights the importance of animals living in harmony with their environment, hence having the ability to express natural genetically encoded behaviour patterns, including those acquired through their normal ontogenic development (Fraser et al., 1997; Rollin, 1993). This approach is very close to consumers’ beliefs, and therefore aspects allowing animals to meet their behavioural (or ethological) needs, by providing proper environmental conditions, are considered welfare requisites (Fraser, 2006). Nevertheless, scientists have highlighted the importance of not considering the full range of behaviours as a guide to assess animal welfare, as there are behaviours that, although natural, are considered to show suffering (e.g., flight reactions) (Fraser et al., 1997; Hughes & Duncan, 1988).

The second approach centres its attention on animal’s “feelings and emotions”, pointing to more psychological aspects of welfare, like suffering and fear, as negative emotions, or pleasure and comfort, as positive ones (Fraser et al., 1997). This approach is based on the belief that non-human animals can experience these emotional states, and considers that welfare is mostly dependent on what the animal “feels” (Duncan, 1993, 1996). This approach has been accepted and integrated in European legislative and administrative provisions as a key principle ever since animals are considered to be “sentient beings” (The Treaty of Lisbon, 2009). Therefore, animal welfare is given the same type of footing as other key principles such as human health or sustainable development. There are several studies that show it is possible to measure animals’ emotions. However, there are no studied direct and objective indicators to assess an animal’s emotional state. Therefore researchers have to rely on assessing behavioural or physiological changes, which are frequently associated with cognitive bias from observers and may therefore present problems of interpretation (Fraser, 2009; Mendl, Burman, Parker & Paul, 2009).

The third welfare approach is based on the concept of an adequate “biological function”, and describes the welfare of an individual as “its state as regards its attempts to cope with its environment” (Broom, 1986, p. 524). It relies on the assumption that an animal to be able to overcome a challenge that affects it’s functioning, has to be able to activate coping mechanisms that can be either physiological or behavioural, at a determined “biologic cost”, sometimes referred as "stress". The notion of a good level of welfare implies that an adequate balance is set between different coping strategies and the fulfilment of biological needs, as the imbalance may lead to the occurrence of disease (Broom & Johnson, 1993). However, the establishment of a link between welfare and indicators of biological function is not always easy or straightforward, as it is not always guaranteed that the fact that an animal is not capable to answering to an aversive situation leads to the manifestation of diseases or injuries (Broom & Johnson, 1993).

All these approaches have their own drawbacks and cannot by themselves define animal welfare; considering only one approach can even be seen as a hindered concept that can therefore limit the overall aim of animal welfare research (Fraser et al., 1997). Hence, these approaches should be considered as intertwined, leading to the need to assess animal welfare in a more "holistic" approach (Broom, 1998). The FAWC followed this more "holistic" approach and integrated all three of the above mentioned concerns when adopting and revising the "Five Freedoms" concept by the Brambell Committee in 1965 (Farm Animal Welfare Council [FAWC]; 2009). The "Five Freedoms" concept is still today a reference for the study and assessment of animal welfare in production animals (Webster, 2001), and should be viewed as a general guideline pointing the way to improve animal welfare (Rushen, Butterworth & Swanson, 2011).

In 2008, the Welfare Quality project re-elaborated the concept of the "Five Freedoms" (FAWC, 2009) and defined four main areas of concern ("Welfare Principles"), which were then split into twelve criteria ("Welfare criteria") (Table 1) (Blokhuis, Veissier, Miele & Jones, 2010; Rushen et al., 2011), each of which corresponded to a key welfare dimension as a result of a fruitful science-society dialogue around the welfare of farm animals. Accordingly, criteria are independent of each other and form an exhaustive, but minimal list (Blokhuis et al., 2010; Botreau, Veissier, Butterworth, Bracke & Keeling, 2007). The Welfare Quality concept of animal welfare establishes a hierarchical aggregation procedure to produce an overall assessment of animal welfare (Botreau, Capdeville, Perny & Veissier 2008).

Table 1. Welfare principles and criteria, as defined by the Welfare Quality project

Principles	Welfare Criteria
Good Feeding	Absence of prolonged hunger
	Absence of prolonged thirst
Good Housing	Comfort around resting
	Thermal comfort
	Ease of movement
Good Health	Absence of injuries
	Absence of disease
	Absence of pain induced by management procedures
Appropriate Behaviour	Expression of social behaviours
	Expression of other behaviours
	Good human-animal relationship
	Positive emotional state

Derived from what was written above, nowadays the state of the art considers the animal welfare concept to be a multidimensional one, requiring that all welfare key criteria are included and interconnected to determine the overall welfare status (Botreau, et al., 2008; Mason & Mendl, 1993). The need to consider multiple criteria when evaluating on-farm animal welfare makes the development of welfare assessment protocols a difficult task. For this reason the development of feasible indicators is one of the main challenges in welfare assessment research (Knierim & Winckler, 2009). The work conducted in this thesis on the development of the BCS and lameness indicators aimed to advance knowledge in that area.

It is also important to mention that the animal welfare definition is not static and is subjected to permanent changes as society and scientific knowledge advances. A recent example of this was the work conducted by Tuyttens, Vanhonacker, Van Poucke & Verbeke (2010). This team of researchers assessed the quantitative correspondence between the operational definition of animal welfare, having defined the concept in terms of four "welfare principles" and 12 "welfare criteria", and the opinion of different stakeholders from Belgium. The results showed that the Belgium stakeholders (that were not present in the focus groups that lead to the Welfare Quality operational definition) broadly agreed with the concept division on the four "welfare principles" and 12 "welfare criteria", but had questions regarding the aggregation procedures not considering all principles to be equally important (Tuyttens et al., 2010). Therefore reflecting public concerns on an animal welfare definition is a challenge, especially considering the diversity of cultural background and socio-economic development among European countries. Even within each country citizens have different perceptions and understanding of what is animal production and welfare.

Still, the concept defined by the Welfare Quality project is nowadays the definition that represents the more generalised consensus of what is animal welfare, both to society and to the scientific community [7] (Rushen et al., 2011). For this reason the Welfare Quality framework was followed to identify the most promising welfare indicators to assess dairy goats' welfare[8].

[7] The literature review conducted on chapter six followed this state of the art definition.

[8] Further information regarding the indicators is presented on chapter six.

CHAPTER 3

Animal welfare – Measurement

Chapter 3 introduces the two main categories of welfare indicators used to assess animal welfare on farms. It gives special emphasis to the challenges associated with developing animal-based indicators that have triggered some of the research conducted in this thesis.

As discussed in the previous chapter, animal welfare is a multidimensional concept. Therefore, to have a proper overall assessment, all the key concerns have to be considered (together with their interconnectedness), which turns animal welfare assessment into a complex multicriteria evaluation challenge (Botreau et al., 2008). When structuring a multicriteria model, the first step consists in defining, following theoretical and practical rules, a set of criteria that are considered relevant (Bana e Costa & Beinat, 2005; Botreau, Veissier, et al., 2007). The definition of such criteria was conducted by the Welfare Quality project, and represents the starting point for the research developed in this thesis.[9]

After the "welfare criteria" are defined, the next step is to ensure their operationalization through different indicators (Bana e Costa & Beinat, 2005; Botreau, Veissier & Perny, 2009). The association of an indicator to a "welfare criterion" is usually a matter of choosing from the pool of indicators available the ones that are considered most adequate (i.e., that have the proprieties that will be described in the beginning of chapter four). It is also important to acknowledge that the same indicator may provide information related to different "welfare criteria" (Blokhuis et al., 2010).

Two main categories of indicators have been used to assess on-farm animal welfare: environmental-based (resource and management) and animal-based (Rushen et al., 2011). In the next sub-chapters we detail some characteristics of the two categories of indicators. The objective is to help the reader to understand the option to use animal-based indicators in detriment of environmental-based ones.

[9] In chapter two, while exploring the animal welfare concept, the welfare quality framework of assessment is detailed.

3.1 ENVIRONMENTAL-BASED (RESOURCE AND MANAGEMENT) INDICATORS

Environmental-based indicators (also referred to as input-based indicators) describe the features of the environment in which the animals are inserted, therefore focusing on what is available to the animals in terms of resources (e.g., food, shelter, type of floor, space allowance, access to pasture), and management choices (e.g., how and when they are fed, how and if they are moved and mixed with other animals, what and how painful procedures are executed) (Johnsen, Johannesson & Sandøe, 2001; Rushen et al., 2011; Sandøe, Munksgaard, Badsgard & Jensen, 1997).

Resource and management-based indicators are considered to have a low level of subjectivity, and are usually associated with a high level of feasibility (they require little training and are not time consuming, therefore also have low costs associated), and reliability (usually having high inter and intra-observer reliability) (Capdeville & Veissier, 2001; Johnsen et al., 2001; Whay, 2007). Moreover, the assessment of environmental-based indicators can be seen as a way to prevent welfare problems, therefore considering these type of indicators more as risk factors than direct measures of animal welfare (Canali & Keeling, 2010; Rushen et al., 2011).

For the reasons presented above, most of the monitoring systems that were developed in Europe since 2001 are largely based on environmental-based indicators (Johnsen et al., 2001), e.g., the Animal Needs Index developed by Bartussek (1999), the Freedom Food Schemes (Main, Webster & Green, 2001), as well as the majority of national and European legislation that has been produced (Veissier et al., 2008).

However, in order to adopt environmental-based indicators, scientific evidence should exist to support the assertion that a given housing or management routine is unquestionably a hazard affecting welfare. Conversely, some studies do not sustain environmental-based indicators as valid, showing farms with equivalent production systems demonstrating enormous variation in animal welfare (Mülleder, Troxler, Laaha & Waiblinger, 2007; Sandøe et al., 1997). Furthermore, individuals with different genetic backgrounds (e.g., different breeds) may respond differently to the same environment. For example, Mattiello, Battini, Andreoli & Barbieri (2011) showed how dairy cows from different breeds presented different levels of welfare under similar environmental conditions. The validity of using this type of indicators can be questioned due mainly to their indirect relation with animal welfare, and for all the multiple interactions that can be established with other resource and management features and the animal itself (Waiblinger, Knierim & Winckler, 2001).

This type of studies highlight two important findings: 1) how assessments based on environmental-based indicators can fail to assess the welfare status of animals in a given farm, as the link between specific environmental indicators and the animals' welfare status is not completely clear (Blokhuis et al., 2013); and, 2) how, only by using animal-based indicators (e.g., behavioural and health indicators), it is possible to assess the animals' response to living in a particular farm, and so to properly assess animal welfare at herd level (Rushen et al., 2011).

These points show the disadvantages of using environmental-based indicators and support the decision to focus on animal-based indicators for welfare assessment. This has being highlighted by multiple research teams (e.g., Welfare Quality) and there are several current studies that focus on the development and integration of animal-based indicators in welfare assessment protocols (Main, Kent, Wemelsfelder, Ofner & Tuyttens, 2003; Rushen et al., 2011; Webster, 2005; Whay, Main, Green & Webster, 2003). This was also the strategy followed in the AWIN project and consequently in this thesis.

3.2 ANIMAL-BASED INDICATORS

Animal-based indicators (also referred as outcome-based indicators) give a more direct assessment of the animal's welfare as, by observing the animal directly, they measure the way in which the animal itself responds to the environmental conditions to which it is subjected (Johnsen et al., 2001; Rushen et al., 2011; Whay et al., 2003), therefore allowing for comparisons of the welfare of animals kept in different types of farming conditions.

There are four main categories of animal-based indicators: health, physiological, behavioural, and record-based indicators.

The most commonly addressed examples of health indicators are injuries (e.g., skin injuries and locomotion disorders) or diseases (e.g., presence of external abscesses, presence of nasal or ocular discharges), both of which can be directly assessed by looking at the animals. All health problems are of economic importance, thus there is an array of studies focusing on health indicators, which makes this category the most quoted. For example, for goats see Battini et al. (2014).

Physiological indicators, including some blood parameters, heart-rate and hormones (e.g., hypothalamic pituitary adrenal axis activity), have a lot of limitations to be used in on-farm welfare assessments as they are difficult to interpret, not always associated with poor welfare, and have several feasibility issues (e.g., methods can be invasive, usually require animals' restraining, require special training and material, being therefore costly, the results are not readily available) (Capdeville & Veissier, 2001; Winckler, 2006).

Behavioural indicators, either in the form of tests performed (e.g., avoidance distance) or in observations (e.g., looking for signs of lameness or abnormal behaviours such as stereotypies or self-mutilation) are very commonly used in welfare assessment schemes. Viñuela-Fernández, Jones, Welsh & Fleetwood-Walker (2007) even point out that behavioural indicators are the most frequently used indicators to assess pain at farm-level.

Some of the advantages attributed to these indicators are that they are non-invasive, non--intrusive, and they many times mirror the animal's health and what they want (Dawkins, 2004). However, the interpretation of this type of indicators can be challenging as they are frequently associated with a high level of subjectivity (Meagher, 2009). Moreover, they present several constraints in terms of feasibility as some of the indicators are expressed for a short period of time (e.g., stereotypical behaviours) and can even be supressed by the presence of the observer.

Record-based indicators (e.g., reproduction indexes, milk yield, longevity) can be suggestive of potential welfare problems, and therefore have the potential to be used in early warning systems (Colditz, Ferguson, Collins, Matthews & Hemsworth, 2014; Kelly, More, Blake & Hanlon, 2011). However, it is also true that the production on a farm can be satisfactory even if some animals are in poor conditions (Botreau, Bracke, et al., 2007), which makes performance indicators unspecific and therefore difficult to interpret. Another disadvantage is that the availability of these indicators at farm level depends greatly on the type of farmer and country. If in some countries health and production databases are readily available (Johnsen et al., 2001), there are countries where this does not happen, which makes comparisons between farms and countries difficult.

Still, there is a lot of research being conducted with this type of indicators as they may guarantee a first level welfare assessment without the need to physically visit the facilities (de Vries, Bokkers, Dijkstra, van Schaik & de Boer, 2011; Sandgren, Lindberg & Keeling, 2009).

Several studies on individual welfare indicators have shown how animal-based indicators can provide a more direct assessment of animal welfare. For example, studies on lameness in different species (Christodoulopoulos, 2009; Eze, 2002; Flower & Weary, 2009; Hill et al., 1997; Lima, Pascoal, Stilwell & Hjerpe 2012b) have shown lameness as an important indicator of pain with strong impact on milk yield, and fertility, and also contributing to pregnancy toxaemia and neonatal diseases.

Recently, different welfare assessment protocols have been developed considering animal--based indicators. Examples are the protocols developed by the Welfare Quality project. This project constructed a multicriteria evaluation model for welfare assessment (farms, slaughterhouses) that was applied to different species (e.g., cattle, pigs and poultry) and that is considered a pioneer in developing animal-based indicators (Blokhuis et al., 2010; Rushen et al., 2011).

The European Union (EU) Strategy for the Protection and Welfare of Animals 2012- 2015 is also recommending animal-based indicators as more suitable for animal welfare assessment (European Commission, 2012). The EU strategy identified four main areas of action among which are: introducing science-based animal-based welfare indicators as a potential way to simplify the legal framework and encourage competitiveness of livestock producers, and developing benchmarking voluntary schemes by better informing the general public on animal welfare issues.

3.3 CHALLENGES OF ANIMAL-BASED INDICATORS

The integration of animal-based indicators in welfare assessment schemes entails several challenges. For instance, there are few available indicators centred directly on the animals (Johnsen et al., 2001), and they are even fewer if we consider small ruminants (Battini et al., 2014; Caroprese, Casamassima, Rassu, Napolitano & Sevi, 2009; Phythian et al., 2011). Therefore, when developing a welfare assessment protocol centred on animal-based indicators there is a strong need for further research, validation or adjustments.

Animal-based indicators have several advantages in terms of validity over environmental-based indicators. However, when considering validation, there are indicators for which the relation with welfare is simple and straightforward (e.g., the examples already described for injuries and diseases mainly linked to pain), but for others finding a correlation is much more difficult. For example, for indicators related to human-animal relationship and to positive emotional state, the link is much more difficult to establish due to their relationship with husbandry conditions, either in terms of housing or management, and animals' characteristics like breed, age or even individual temperament (Waiblinger et al., 2006; Yeates & Main, 2008).

Other challenges are related to reliability as there is an associated level of subjectivity involved when using animal-based indicators, and so the observers' assessment might be biased by his/her own concern for the animals (Meagher, 2009).

Additionally, the recording of animal-based indicators is strongly influenced by the species being assessed and the housing conditions. Taking goats as an example: this specie is considered to be gregarious and with a strong anti-predator sense which makes animal-based individual observations time-consuming and difficult. Whereas environmental-based observations are much more straightforward, fast to collect, and can be performed at any time. For these reasons, the development of more feasible ways of assessing the indicators is paramount. An example of how research can advance in this sense is the work developed in this thesis with the BCS indicator. In chapter seven we detail how we developed a visual scoring system to assess BCS that replaced the need to score BCS by palpation.

Another possible strategy to increase the feasibility of the indicators is by implementing an effective sampling strategy that reduces the observations to some individuals or groups of individuals inside a unit. There are several problems associated with sampling strategies. The first is that the implementation of these strategies require previous knowledge and determination of parameters such as expected prevalence, level of precision, and the confidence level of the estimation that are not frequently available (Endres, Lobeck-

-Luchterhand, Espejo & Tucker, 2014). The second is that the sampling is influenced by the recording location. For example, if a protocol is designed to collect information during milking, it has to take into account that some of the animal-based indicators (e.g., lameness) might influence the milking order (Main et al., 2010), and therefore the sampling strategy has to take this into consideration.

In our research project we addressed some of the challenges associated with the development of animal-based indicators by developing the BCS (chapter seven) and lameness (chapter eight) assessment. Furthermore, we took into account validity, reliability and feasibility concepts when drawing our experimental designs. In the next chapter we detail these concepts, providing the reasoning for the development of this research.

CHAPTER 4

Animal Welfare - Overall assessment: from indicators to protocols

Chapter 4 introduces and discusses the basic concepts underlying the indicators development that provided a theoretical basis for all the fieldwork conducted during this research. Furthermore, it illustrates how to move from the indicators level to an overall assessment of welfare.

After the "welfare criteria" are defined, the next step is to ensure its operationalization with different indicators (Bana e Costa & Beinat, 2005; Botreau et al., 2009). The association of an indicator to a "welfare criterion" is usually a matter of choosing, from the pool of indicators available, the ones that are considered most adequate (i.e., that have the proprieties that will be described in the beginning of chapter four). It is also important to acknowledge that the same indicator may provide information related to different "welfare criteria" (Blokhuis et al., 2010).

To develop a protocol that delivers an overall assessment of animal welfare different indicators need to be integrated into one assessment protocol (Fraser, 2006; Webster, 2005). There are two fundamental questions underlying this process. The first one is that the indicators need to meet given proprieties to be considered suitable to be integrated in the protocol. The second is that the choice of indicators, and associated level of measurement, needs to consider the aim of the protocol. In the two next sub-chapters we will debate these issues.

4.1 DEVELOPING WELFARE INDICATORS – WHAT ARE THE BASICS?

To be able to properly select indicators for an on-farm welfare protocol, it is paramount to take into account their psychometric proprieties. Therefore, this chapter explores the concepts of reliability, validity, and feasibility. After these properties are respected, the next step is to define how to operationalize the indicators, that is, how to define a scoring scale.

4.1.1 Reliability

Reliability is an essential requisite for scientific measurement. It weighs the amount of random and systematic error (the smaller the error, the more reliable is the measurement), and is usually assessed before validity, as it is important to first guarantee that an indicator has the ability to measure something in a reproducible way (Meagher, 2009; Streiner & Norman, 2008). Overall, it estimates the ability of observers to differentiate each case among the different levels of a scoring system (Kottner et al., 2011; Streiner & Norman, 2008).

In this research thesis, reliability is assessed under the general description of consistency, that is, the reproducibility of a given indicator when it is gathered in different occasions (Streiner & Norman, 2008). This form of reliability can be explored by means of inter-observer reliability studies (measures the degree of agreement between different observers), intra-observer reliability studies, or observer consistency studies (measures the degree of agreement between observations made by the same observer on two different occasions), and test-retest reliability studies (refers to the agreement between observations performed on the same individual, or farm, on two different occasions separated by a time interval) (Scott, Nolan & Fitzpatrick, 2001; Streiner & Norman, 2008). Some authors consider that in the case of welfare assessment the intra-observer reliability is a part of test-retest reliability (de Passillé & Rushen, 2005; Temple, Manteca, Dalmau & Velarde, 2013).

Common inter-and intra-observer reliability measures are the kappa coefficient, (k), the weighted kappa (k_w), and intra-class correlation (ICC). The kappa coefficient (k) is a measure of "true" agreement that assesses the proportion of agreement beyond that expected by chance, handling all disagreements equally (Sim & Wright, 2005) and is used for categorical items. Weighted kappa (k_w) penalizes disagreements in terms of their seriousness, thus, being appropriate for ordinal scales (Cohen, 1968). In k_w the weighting scheme has to be chosen, being the quadratic weighting scheme, where disagreement weights are proportional to the square of the deviation of individual ratings (Cohen, 1968; Sim & Wright, 2005; Streiner & Norman, 2008). For data interpretation, Fleiss and Landis and Koch thresholds are commonly used (Streiner & Norman, 2008). Fleiss thresholds for k are: 0 - 0.40 = poor; 0.41–0.75 = fair to good; and 0.76–1 = excellent; and Landis and Koch thresholds for k_w are: < 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 & Norman, 2008). The Welfare Quality project accepted as the maximum lower limit 0.4, however, as de Passillé & Rushen (2005) mentioned in their study, if there are economic consequences for farmers, higher levels of reliability should be considered as lower limits. Streiner & Norman (2008) consider a minimum of 0.5.

ICC compares the variance of different measurements of the same subject with the total variance across all measurements and subjects, integrating the magnitude of the disagreement to evaluate inter-observer reliability estimates (larger magnitude disagreements lead to lower ICC), and it can be used for ordinal, interval, and ratio variables (Hallgren, 2012; Weir, 2005). Estimates for ICC are interpreted using the following guidelines: 0-10% = virtually none; 11–40% = slight; 41–60% = fair; 61–80% = moderate; 81–100% = substantial agreement (Shrout, 1998). There are several ICC variants that must be considered based on the nature of the study and the type of agreement the researcher wishes to estimate, as "one-way" or "two-way model" (for considering how observers are selected for the study); how the inter-observer reliability is going to be characterized – if "absolute agreement" or "consistency"; the researcher must stipulate the unit of analysis that the ICC results apply to, i.e. averages of ratings scored by several observers or based on ratings provided by a single observer; and finally if the observers are considered to be random or fixed effects (Hallgren, 2012; Shrout & Fleiss, 1979).

Agreement (proportion of exact agreement and disagreement) is another measure frequently used to assess stability of observations, defined as the ability of observers to give identical scores, and is calculated by dividing the number of agreements by the total number of agreements and disagreements (Martin & Bateson, 2007). Although, it is not considered a reliability measure, it is useful for data interpretation when compared with the previous described measures.

It should be recognized that if in a study we have high levels of inter-observer reliability, the intra-observer reliability is sure to be high too, therefore one can exclude the demonstration of the latter from the experimental designs (Streiner & Norman, 2008). However, it is paramount to include them in case that inter-observer reliability is low, as we will not be sure if the reason is due to differences within or between observers, or even both (Streiner & Norman, 2008).

In welfare assessment schemes the stability of results contributes to the reduction of costs (Knierim & Winckler, 2009; Sørensen, Rousing, Møller, Bonde & Hegelund, 2007). Having indicators that do not change significantly over a long period of time allows for a reduction of farm visits, while keeping the results reliability. Literature points to different methods to assess test-retest reliability (Lensink, van Reenen, Engel, Rodenburg & Veissier, 2003; Plesch, Broerkens, Laister, Winckler & Knierim, 2010; Rousing & Waiblinger, 2004), however the most accurate form seems to be the one applied by Temple et al. (2013). Temple et al. (2013) started by performing a Wilcoxon signed rank test to test whether the prevalences obtained during two farm visits were significantly different at the 0.05-level. Subsequently, in order to overcome the inconsistencies between methodologies observed in different studies as reported by Costa-Santos et al. (2011), Dohoo, Martin & Stryhn (2009), and De Vet, Terwee, Knol & Bouter (2006) three statistical methods were used to assess consistency of animal-based indicators over time. At first, a spearman's rank correlations between the two visits was calculated (Martin & Bateson, 2007). Subsequently, ICC was used to determine reliability, with a two-way mixed effects model (Shrout & Fleiss, 1979), i.e., the subjects in the study are considered to be random but the observers (raters) are not random effects, with absolute agreement and single-measures being estimated. Finally, the limits of agreement between visits for the animal-based indicators were determined following the Bland & Altman (1986) method, where limits of agreement are estimated on the difference of prevalences between the two separated visits and the standard deviation of these differences. Low test-retest reliability results may have three meanings: the event under assessment might have changed; the scoring method may be unreliable; the "test is reactive" in the sense that the first assessment might influence the second (Streiner & Norman, 2008).

Reliability can also be explored by focusing on cardinal consistency literature. In decision analysis studies, inconsistency represents a conflict between the decision-maker assessment and a fixed model (González-Pachón, Diaz-Balteiro & Romero, 2014). This is also the implicit definition of inconsistency in inter-observer reliability studies: verifying whether the observer's assessment matches a numerical rating scale (NRS) model, ascertaining for ordinal consistency. In NRS there is ordinal consistency between the different levels of the scale, which is generally easy to deliver by the observers as it merely requires a comparability assumption. However, it is associated with difficulties in aggregating scales (Arrow, 1951). Similarly to ordinal consistency, cardinal consistency allows not only for assessing the order of the answers, but also the intensity of the assessment (Keeney, 1976). Although this information is more difficult to deliver by the observers, it makes the aggregation of scales easier (Keeney & Kirkwood, 1975).

As reliability is a pre-requisite for validity, high levels of observer reliability offer confirmation of the validity of the indicators (Hewetson, Christley, Hunt & Voute, 2006; Kaler, Wassink & Green, 2009; Meagher, 2009), and are necessary to show the objectivity of indicators being assessed by different observers in an on-farm welfare assessment context (Phythian et al., 2012). This is particularly important when considering animal-based indicators, as indicators taken on animals are more prone to variation (Temple et al., 2013). Therefore, if reliability of an indicator is poor then it is probably inappropriate for animal welfare monitoring (De Rosa, Grasso, Pacelli, Napolitano & Winckler, 2009)

It is also important to consider that reliability is not an absolute propriety of an indicator – it is an interaction between the instrument of assessment, the group of observers using the instrument and the situation in which the test is taking place (Streiner & Norman, 2008). Hence, when considering reliability test results or developing reliability studies, one has to take into account the way data is collected. The instrument categories must be explicitly defined and the practice, experience and fatigue of the observers must be considered (Martin & Bateson, 2007). Moreover, an adequate sample has to be determined. The sample size determination has to take into account both the expected and acceptable inter-observer reliability, and also the levels of both α and β (generally, $\alpha = 0.05$ and $\beta = 0.2$), to be able to estimate a minimum number of observations to complete the studies (Walter, Eliasziw & Donner, 1998).

While it is true that under commercial conditions the development of reliability studies may be challenging (Kaufman & Rosenthal, 2009; Schlageter-Tello et al., 2014; Streiner & Norman, 2008), the lack of studies, and especially the lack of studies that provide information for data interpretation, is a drawback when one is exploring indicators with the objective of integrating them in a welfare protocol.

Finally, improving descriptors definition (either by refining definitions or by merging several detailed descriptors), ensuring good data recording and training, have proved to be essential to increase reliability (Brenninkmeyer et al., 2007; Gibbons, Vasseur, Rushen & Passille, 2012; Knierim & Winckler, 2009; March, Brinkmann & Winkler, 2007).

4.1.2 Validity

Validity will determine if an indicator is measuring what it is intended to measure, as this is fundamental to be able to say that an indicator is providing information to the question being asked (Cronbach & Meehl, 1955; Martin & Bateson, 2007; Streiner & Norman, 2008).

There are different approaches to assess validity, but ultimately they all relate to three settings. The first is when one wants to check the validity of an indicator or test for which there are already other scoring methods. In this approach, usually described as criterion validation, the researcher administers the indicator or test simultaneously with the already existing scoring system and determines how strong is the correlation between the two (it is assumed that the correlation should fall between 0.4-0.8, as if it is lower, it means that the two instruments are probably measuring different constructs) (Streiner & Norman, 2008). The already existing scoring system should be, as much as possible, used and acknowledged in the field of study as the "gold-standard" (Meagher, 2009; Streiner & Norman, 2008). Since in the case of overall welfare assessment there is no "gold standard", criterion validity must be demonstrated by comparison to other valid and reliable methodologies (Cohen et al., 1996), being much harder to assess than reliability (De Passillé & Rushen, 2005).

Criterion validation can be divided into concurrent and predictive validation. In the most commonly used concurrent validation studies, the correlation between the new and already available scoring system (here named criterion measure), or between the indicators and other measures to which the latter is theoretically related, is established by administrating them both simultaneously; in predictive validation studies the criterion measure will only be available in the future (e.g., diagnostic tests in which we must wait for the appearance of symptoms), therefore the aim is to have answers earlier than current scoring system (Kamphaus & Frick, 2005; Meagher, 2009; Streiner & Norman, 2008).

The second context in which validity is usually assessed is when no other measures exist, and is known as construct validity (Streiner & Norman, 2008). In this case the validity studies aim to establish a link between the attribute being measured (an effect, e.g., level of lameness) and some other attribute (a treatment, e.g., floor type), by means of a hypothesis or construct (Scott et al., 2001; Streiner & Norman, 2008). The defined hypothesis is then tested by means of an experimental design where the researcher tries to find if the attributes to which the construct is related, are identified as convergent – convergent validity, or as divergent – discriminant validity (Meagher, 2009; Streiner & Norman, 2008). It is however important to point that construct validation is an on-going process, as in the majority of cases one study is not enough to univocally establish validity (Streiner & Norman, 2008).

The third form of validity, frequently seen in welfare and behavioural studies, is content validity, also referred as face validity. In this type of validation, an indicator or test is assessed by having one or more experts assessing them by a subjective judgment that is believed to be valid (Scott et al., 2001; Scott, Fitzpatrick, Nolan, Reid & Wiseman, 2003; Streiner & Norman, 2008).

As reviewed by Botreau, Bonde, et al. (2007) welfare indicators have been validated in different ways, either by comparing them with more refined methods (criterion validity), by investigating treatments effects (construct validity), or by using experts' judgment (content validity). However, it is important to keep in mind that an overall animal welfare assessment system may only be appropriate if the indicators that integrate it have established validity, therefore the use and correct application of validation experimental designs should receive ample attention from the scientific community.

4.1.3 Feasibility

Feasibility refers to the potential practicality and viability of a measurement procedure (Martin & Bateson, 2007). Regarding on-farm welfare assessment feasibility, Knierim & Winckler (2009) identify two main constraints that have to be taken into account. The first is the time constraint. The possibility to measure an indicator in a limited time is a driver of animal-based indicators and measurements development. It is the reason why environmental indicators have had such a predominant space in welfare indicators selection (Johnsen et al., 2001), and why some behavioural indicators (e.g., resting patterns) and those indicators that need further processing after collection (e.g., some physiological indicators), are frequently excluded (Spoolder, De Rosa, Hörning, Waiblinger & Wemelsfelder, 2003). It is important to mention that a survey should be first conducted among stakeholders to determine a maximum acceptable amount of time for an on-farm assessment.

The second constraint is cost (Knierim & Winckler, 2009). More than considering costs associated with specific equipment or consumables, this constraint is influenced again by time taken to perform the assessment, as well as time taken to train assessors, and how that reflects in terms of men-hours. Therefore, an important step for integrating welfare indicators in protocols is developing a sample strategy that allows for the decrease in the the number of animals to be assessed (Sørensen et al., 2007).

4.2 SCORING SYSTEM DEVELOPMENT

After reliability, validity and feasibility is established, a scoring system has to be defined to operationalize the indicators in order to allow their integration in a welfare assessment scheme. This operationalization is accomplished by choosing a scoring system. As already discussed, the assessment objective (e.g., legislative or advisory) usually determines the indicators to be integrated. However, it is also true that the same indicator can be integrated at different levels of measurement. For example, the assessment of only the extreme BCS cases (very thin and very fat animals) may be enough for legislative purposes, but perhaps not for advisory ones where a more sensitive scale is needed. Also, when integrating indicators it is important to have in mind the different options in terms of scoring systems to ensure operationalization. Moreover, it is important to understand that the choice of a given scoring system implies different constraints in terms of level of measurement. In this section we discuss these options, with this discussion showing the scope and relevance for developing the study on lameness scoring system development in chapter nine.

There are different types of scoring systems available, and when selecting the type of scale the researcher should take into account the kind of possible answers: categorical or continuous (Streiner & Norman, 2008). In welfare assessment two types of scoring systems are frequently used: ordinal and continuous scales (Scott et al., 2001). Ordinal scales, also called NRSs, are explicit scoring systems in which each individual is scored accordingly to different descriptors that correspond to a whole number (Gaynor & Muir, 2008; Scott et al., 2001; Streiner & Norman, 2008). In welfare assessment, ordinal scales are the scoring systems most frequently used, due to being generally easy to deliver as it merely requires a comparability assumption. However, being in most cases artificial constructs, they are associated with difficulties in their use and interpretation (Arrow, 1951). This is particularly evident if we use them to assess variables that are continuous by nature, such as lameness or pain levels. There are several problems associated with ignoring the continuous nature of variables. The first one is that by itself it introduces error in the answers, as different people will interpret differently the descriptors. Another source of error will be the limited choice of descriptors usually available. The third problem is the associated loss of efficiency and sensibility of the scoring system, with potential loss of valuable information (Streiner & Norman, 2008). These problems are frequently prevented by the use of continuous scoring systems.

The most used continuous scoring systems in animal welfare assessment is the visual analogue scale (VAS) (Meagher, 2009). The VAS is a nonverbal scale allowing for the recording of clinical events, such as lameness, in a continuous way. The VASs were developed for use in pain assessment in humans (Scott & Huskisson, 1976) but are also used to measure a variety of subjective phenomena in the behavioural and social sciences, and are considered to have potential value for the measurement of different clinical phenomena (Wewers & Lowe, 1990). The VAS has the advantage of not imposing a choice of descriptors, being possible to score a change even if a change between categories would not occur, and therefore overcoming some drawbacks associated with ordinal scales (Averbuch & Katzper, 2004; Paul-Dauphin, Guillemin, Virion & Briançon, 1999; Welsh, Gettinby & Nolan, 1993). However, VASs are generally viewed as being too subjective, with low user-acceptance, and difficult to use in farm conditions (Engel, Bruin, Andre & Buist, 2003; Kaler et al., 2009). Current research in scale development has recently been focusing on the development of modified VASs to assess different health indicators. An example of some studies aiming for the development of such scales for lameness assessment are Nalon et al. (2014), and Tuyttens, Sprenger, Van Nuffel, Maertens & Van Dongen (2009) for dairy cows.

A modified VAS is a VAS that besides being anchored on its extreme limits, it presents the thresholds of a categorical scale specified on the VAS continuum, dividing it into segments that are complemented with descriptors. The objective is to help observers understand the scale and make consistent choices, therefore increasing the inter-observer reliability by reducing data variation. The modified VAS keeps the advantages of the VAS, such as allowing for the indicator to be recorded in a continuous way, but showing a higher resolution and being more sensible to smaller changes than NRS (Averbuch & Katzper, 2004; Nalon et al., 2014; Tuyttens et al., 2009). This was possible to observe in two different studies in lameness scoring systems – one in cows (Tuyttens et al., 2009) and another in sows (Nalon et al., 2014). In both studies inter-observer reliability was higher with the modified VAS, and the studies participants reported their preference for the modified VAS.

These studies using the modified VASs, split the continuum into equal ranges placing (underlying) NRS descriptors as thresholds to these ranges for which a text box composed by different lameness and posture signs is added. However, it is imperative to assess if the distribution of the NRS descriptors in the VAS is in fact evenly spread, and research should focus on investigating the distribution of different lameness and posture signs along the scales' continuum[10].

[10] The study presented on chapter nine focuses on these questions.

When choosing a type of scoring system one should also consider the level of measurement it delivers. Measurement, or the means by which we operationalize a measure, involve "rules for assigning symbols to objects so as to (1) represent quantities of attributes numerically (scaling) or (2) define whether the objects fall in the same or different categories with respect to a given attribute (classification)" (Nunnally & Bernstein, 1994, p.1). Four levels of measurement can be distinguished: nominal, ordinal, interval, and ratio.

The nominal level of measurement is considered when observations are assigned to mutually exclusive, qualitative categories using a nominal (categorical) scale, and it is the most unrestricted form of measurement (Martin & Bateson, 2007; Stevens, 1946). The only permissible statistics with this kind of level of measurement are determination of the number of cases, mode and contingency correlations (Stevens, 1946).

The ordinal level of measurement considers situations where the observations can be ordered along a scale accordingly to some common propriety using an ordinal scale (Martin & Bateson, 2007). As in ordinal scales the distance between categories is unknown, only a limited number of calculations (e.g., non-parametric statistics) can be performed (Merbitz & Morris, 1989; Streiner & Norman, 2008).

The interval level of measurement considers the situation in which scores can be placed on a scale in a way that the intervals between categories are known, meaning that the difference between two points of the scale can be quantified. However, the zero point and the unit of measurement are random for an interval scale (Martin & Bateson, 2007). This allows for mode, median, arithmetic mean and standard deviation to be calculated (Stevens, 1946).

The ratio level of measurement is used whenever it is possible to determine a meaningful zero in a scale, and it represents the highest level of measurement (Martin & Bateson, 2007; Streiner & Norman, 2008) adding the calculation of a coefficient of variation to the mathematical operations.

4.3 Integration indicators - What is the aim the protocol?

As enumerated by Main (2009) there are four broad circumstances, each having their specific objectives, in which welfare assessment systems can be applied: research, legislative requirements (non-voluntary), certification systems (voluntary) and advisory/management schemes. Bellow we focus on the last three circumstances, as some of the characteristics of these forms of control, like the welfare indicators' selection, development, and integration will be influenced by the welfare assessment protocol particular application.

Regarding legislative requirements, animal welfare attracts more attention in Europe than in other regions (Dalla Villa, Matthews, Alessandrini, Messori & Migliorati, 2014), and since the year 2000 the legislation on animal welfare and its complexity has increased (Schnettler, Vidal, Silva, Vallejos & Sepúlveda, 2009)[11], and it is expected to continue to increase as there is growing evidence that animal welfare has a strong impact on both food safety and quality (Blokhuis, Keeling, Gavinelli & Serratosa, 2008; Horgan & Gavinelli, 2006).

In the EU there are two main ways in which legislation regarding animal welfare is formulated. The first one is via the General Directorate for the Health and Consumer Protection (DGSANCO), a department of the European Commission – the executive body of the EU. The DGSANCO, with its scientific committee (the scientific committee on Animal Health and Animal Welfare is a panel of experts working with the European Food Safety Authority [EFSA]), delivers recommendations on how to ensure production animals' welfare. These recommendations may be drafted into directives that are submitted to the Council of Ministers of the EU and may be adopted as a Council Directive if approved (Veissier et al., 2008). The EU directives are then transposed into national laws in each EU country.

The other way in which legislation might arise is via specific initiatives that are formulated by institutions, such as the Council of Europe, to which the EU is a signatory member. The Council of Europe published the European Convention for the protection of animals kept for farming purposes, which was translated into an EU directive (Directive 98/58/EC) that covers the minimum animal welfare standards for farmed animals. The directive 98/58/EC was then translated into Portuguese regulation by the decree number 64/2000 from April 22 (*Decreto-Lei (DL) n. º 64/2000, de 22 de Abril*). The *Direcção-Geral de Alimentação e Veterinária* (DGAV), which is the competent national authority, ensures that the owners or keepers of animals fulfil this legislation.

The majority of legal documents being produced consider farm animal species together, not making differentiations between specific species needs, and usually relying on minimal requirements to ensure animals' needs (Veissier et al., 2008). There are however specific legislation available in Portugal for laying hens (*DL n. º 72-F/2003, de 14 de Abril*), swine (*DL n. º 135/2003, de 28 de Junho*), broilers (*DL n. º 79/2010, de 25 de Junho*), and veal calves up to six months (*DL n. º 48/2011, de 10 de Fevereiro*). Small ruminants, particularly goats, lack specific legislation, which can be identified as a need and justify the opportunity of research projects financed by the EU on this subject, namely the AWIN project. The different pieces of legislation specify requirements focused on housing and husbandry aspects – that is environmental-based indicators[12] – which do not deliver an overall status of animal welfare and have resulted in an inflexible legislation (Blokhuis et al., 2013). Considering that legislation is produced at an European level, it is very difficult to produce legislation detailing different production practices throughout European countries, as these countries are subjected to different farming, cultural and environmental conditions. For this reason it is

[11] This introduction only considers legislative documents for animals kept for farming purposes. Transport and slaughter legislation is not mentioned.

[12] For further discussion on this issue, see sub-chapter 3.1.

paramount to focus future research on animal-based indicators[13] that can overcome these difficulties and allows for the development of legislative documents that can be readily applied across different European countries. This is what European farmers expect from new legislation on animal welfare (Bock, 2009).

Furthermore, while it is true that new legislation documents represent popular measures by governments, it is important to acknowledge that the new measures have to be properly enforced in order to make them effective (Ingenbleek, Immink, Spoolder, Bokma & Keeling, 2012). For this reason, considering the development of indicators for legislative purposes, an adequate choice of feasible indicators has to be pursued.

Even considering what was mentioned in terms of legislative requirements, animal welfare enforcement has been done mainly by non-state actors like different retailers, non--governmental organizations and industries (Dalla Villa et al., 2014; Veissier et al., 2008). As discussed by Blokhuis et al. (2008) retailers and farmers have recognized that to meet consumers' demands in terms of animal welfare they should incorporate strategies into their business to enhance and be transparent regarding the welfare status of animals providing their products. The enforcement of these rules has been done by an array of methods, namely, different non-mandatory welfare codes and guidelines, information and education programmes, assurance programmes of corporate customers and their associations, and product differentiation through certification and labelling programmes (Fraser, 2006; Ingenbleek et al., 2012; Main et al., 2001). Examples of these market-based instruments are Globalgap, Label Rouge EU label, Free Range, Freedom Food, Whole foods (Ingenbleek et al., 2012; Main et al., 2001). There are many cases where policies implemented by non-state actors are even stricter that national or European legislation, with the specific aim of targeting a consumer market niche, or protecting a geographically specific market (Veissier et al., 2008). Certification schemes are based on formal requirements that farmers have to fulfil. It is important to remember that farmers are usually motivated to keep the standards and engage in these schemes, as there is a strong advising and management component (Main et al., 2001). Therefore it is reasonable to expect farmers collaboration when applying different protocols.

Welfare assessment schemes for advisory purposes, by which the farmer receives feedback regarding risk factors and advice regarding improvement of welfare standards, usually have fewer time restrictions, as it is expected that the information gained will deliver some financial benefit to the farmer. Therefore, a higher level of complexity can be expected.

However, a long and complex protocol will require a lot of time from the assessor, which will make the certification scheme more expensive (Main et al., 2001). In this view, these types of assessment constraints are close to those suggested for the legislative systems.

[13] For further discussion on this issue, see sub-chapter 3.2.

It is fair to say that there is a large degree of interrelationships between these types of assessment tools; therefore they often share common grounds. For example, they all should include given psychometric proprieties, but also be implemented on-farm by assessors with a varying degree of training, respect time constraints and sensitive in the sense that they can point to the causes of impaired welfare (Waiblinger et al., 2001).

In this introduction, aggregation methods of all the different indicators into one assessment protocol are not going to be detailed, as it is not the aim of this research thesis to develop an overall protocol to assess welfare. However, it is important to mention the general characteristics an aggregation method should have, so as to have these in mind when developing related studies. Moreover, it is important to consider the challenge that is to assign weights to each criterion within a protocol and how this weight assigning it is not merely a technical process, as it also takes into account ethical assumptions and societal concerns and values (Sandøe et al., 1997; Spoolder et al., 2003). Science has the responsability to formalise judgments made by societal groups (Botreau, Bracke, et al., 2007). The integration method needs to be clear in the sense that it can be easily explained to stakeholders and it has to be repeatable in a way users trust their results and are encouraged to improve animal welfare (Botreau, Bonde, et al., 2007). Additionally, the method should be feasible to be able to be used routinely on large number of animal units (Botreau, Bonde, et al., 2007).

4.3.1 The AWIN protocol

Accordingly to the issues already described above and considering that the AWIN project aimed to fill a gap in welfare assessment of dairy goats both at the legislative and certification or advisory level, a two level approach was adopted for the development of the AWIN welfare assessment protocol for dairy goats. In a first level, the protocol conducts a quick screening, consisting of a selection of 12 robust, feasible, and, in some cases, multi-criteria animal-based indicators (e.g., queuing at feeding, hair coat condition, lameness), covering all the principles and criteria developed by Welfare Quality (see Table 1). All indicators are animal-based, except for bedding. Detailed information on description, assessment and method of scoring of each indicator can be found in the AWIN welfare assessment protocol for goats (AWIN, 2015). In this first level, indicators are collected and recorded as group data and individual restraint of the animals is not required. Also, in this level only one pen has to be assessed. In order to increase the sensitivity of the assessment, the pen considered as presenting the potentially greatest risk for welfare is to be selected.

Whenever there is a noncompliance with the current legislation, or the within-farm proportion of animals with no signs of the most important or prevalent welfare problems (abscesses, improper disbudding, poor hair coat condition, severe lameness, queuing at feeding, queuing at drinking) is lower than the proportion of animals observed in the worst 5% of the farms of the reference population, or when no goats touch the assessor during the 300 s of the "latency to the first contact test", the second level assessment is applied.

This second level consists of a more comprehensive and in depth assessment that requires individual assessment, but still keeping feasibility especially by ensuring that it is conducted in a reasonable amount of time. In this second level, 18 indicators are evaluated and, if more than one pen is present, additional pens should be evaluated.

For the evaluation of indicators on individual animals, goats need to be restrained and handled, the choice of where to assess the animals should be agreed with the farmer. The individual selection of the animals follows a sampling scheme developed by the project and presented in detail in the species protocol (AWIN, 2015).

The outcome of the assessment is presented as the proportion of animals that do not present welfare problems. An objective descriptive output is generated, where all the indicators (except for Qualitative Behaviour Assessment) are shown. The position of the assessed farm is compared to the median value of the reference population, with all data used to calculate the proportion of each indicator being weighted according to the number of goats on the farm.

The main aim of this thesis was not the development of the dairy goat welfare protocol, for this reason we only presented a brief but necessary description to understand the options undertaken during the development of the indicators, as detailed in Part II of this thesis.

CHAPTER 5

Dairy goat production

Chapter 5 presents some facts and figures that sustain the growing importance attributed to intensive dairy goat production, justifying the focus that is given in this research to this production system.

The domesticated goat (*Capra hircus*) can be found throughout the world, except for the polar regions (Webster, 2011). According to Food and Agriculture Organization of the United Nations the largest number of goats in the world is present in developing countries, with Asia (59.4%) and Africa (35%) leading the list, and Europe having only 1.6% of world goat population (Food and Agriculture Organization of the United Nations [FAO]; 2015). However, even considering Europe's small goat population (16.5 million), it detains the most organized programs for selection, processing and commercialization of goat milk, producing 13.2% of world's goat milk (Castel, Ruiz, Mena & Sánchez-Rodríguez, 2010; Dubeuf, 2010). Dairy goat farming is of paramount importance to the economies of the Mediterranean countries, like, France, Greece, Italy, Portugal, and Spain (Escareño et al., 2013; Pirisi, Lauret & Dubeuf, 2007).

In the last decade the goat population in Europe has diminished 2.5%, although the milk production has increased (Boyazoglu et al., 2005; Castel et al., 2010). This is especially due to production in countries like France and Spain where the number of goats has increased (3.4%, and 8.8% respectively), accompanied by an increase in milk production of 23.5% in France and 15.8% in Spain. These increases are mainly attributed to improvements in farm management, genetic improvements and technological advances (Castel et al., 2010).

In 2011 the European Commission published a study on the impact of the Common Agricultural Policy on the sheep and goat sector. In this report it was identified that there is tendency for the concentration of animals in fewer and larger farms, accompanied by an increase of milk production (European Commission, 2011).

In Portugal the latest statistics available are from 2013 and present a total of 398 thousand goats, distributed mainly in Alentejo, Centre and North inland regions (Instituto Nacional de Estatística, 2015). The generality of farms are small with an average of 12 goats per farm, however, the farms with less than 10 animals are decreasing, whilst the number of farms with more than 100 goats is increasing (Instituto Nacional de Estatística, 2011). Furthermore, this last category of farms concentrated 41% of the total goat population in 2009, while ten years before it only congregated 34% (Instituto Nacional de Estatística, 2011), which emphasizes the changes in farm practices. Accordingly to DGAV (DGAV; personal communication, 2015) there are 269 farms considered to be under intensive production, with a total of 52 199 goats.

In Portugal the goat population is more meat than dairy oriented, with the dairy goat sector population (n=149,000) representing around 35% of the total population and 36% of the farms (Instituto Nacional de Estatística, 2011). In these over 30 million litres were produced in 2012 (FAO, 2015). Changes in husbandry condition, towards a more intensive systems, may have some impact on health and welfare. However, goats are active and inquisitive animals, that seem to have adapted very well to these relatively new systems found on the developed world (Webster, 2011).

Compared with the media attention given to other farm animals, such as cattle, poultry, or pigs, goats' welfare have been disregarded by the general public as they are still associated with mountain living and extensive systems, and recognized by their several adaptation mechanisms to harsh environments, therefore being kept out of close scrutiny (Caroprese et al., 2009). Maybe as a result of this there has been much less research on goats than in other production species (Dubeuf, Morand-Fehr & Rubino, 2004). However, the overall recognition of this long-underestimated species has developed, enhancing its importance in the production sector (Boyazoglu et al., 2005; Morand-Fehr et al., 2004). This increase in importance has been accompanied by a financial compensation with the price of goat milk being higher when compared to dairy cows milk (Boyazoglu et al., 2005). For these reasons it is vital to keep the positive image that the goat sector has maintained during the years (Boyazoglu et al., 2005). Keeping an idea of good welfare associated with this production system is therefore extremely important.

The research on goat welfare has been focusing on few particular aspects of welfare (e.g., studies on lameness (Christodoulopoulos, 2009; Eze, 2002), udder health (Mavrogianni, Menzies, Fragkou & Fthenakis, 2011; Mavrogianni, Alexopoulos & Fthenakis, 2004), and avoidance distance tests (Mattiello et al., 2010). However, other researchers have also aimed to review several issues that may be useful to develop on-farm welfare monitoring systems for small ruminants (Caroprese et al., 2009), or establish protocols for assessing and monitoring the overall welfare of farmed dairy goats (Anzuino, Bell, Bazeley & Nicol, 2010; Muri, Stubsjoen & Valle, 2013).

A recent study has investigated the effects of a disease eradication programme by means of an on-farm welfare assessment protocol, proving how this protocol could evaluate the benefits of such programme in terms of animal welfare (Muri, Leine & Valle, 2015). Despite this there are no official on-farm protocols for dairy or meat goats' welfare assessment, as there is no specific EU legislation that only considers the welfare of goats.

PART II

Applied research on development and integration of animal-based indicators

Part II covers all the research conducted during this thesis. It starts by a literature review on goat animal-based welfare indicators that may be suitable for on-farm welfare assessment, and that is therefore relevant for this thesis. Then it explores the research developed for the development of body condition score and lameness as valid, reliable and feasible animal-based welfare indicators. Finally, it presents the results of the application of the protocol prototype developed within the AWIN project on 30 dairy at farms in Portugal.

CHAPTER 6

Invited review: Animal-based indicators for on-farm welfare assessment for dairy goats

More than focusing on some aspects of animal-based indicators for goat welfare that have been widely studied, or in the clinical description of the indicators, chapter 6 discusses indicators that need to be further developed in order to be integrated in a welfare assessment protocol. Moreover, this chapter highlights some indicators that require further research in terms of validity, reliability, or feasibility; justifying the work developed in this thesis.

This chapter has been published as: Battini, M., Vieira, A., Barbieri, S., Ajuda, I., Stilwell, G. & Mattiello, S. (2014). Invited review: Animal-based indicators for on-farm welfare assessment for dairy goats. *Journal of Dairy Science*, 97(11), 6625-6648. http://dx.doi.org/ 10.3168/jds.2013-7493.

A. Vieira contribution focused mainly on reviewing and drafting the manuscript on some of the published indicators that were classified according to the welfare principles “good feeding”, “good housing”, and “good health”. Articles reviewed were thoroughly analysed to extract information on animal-based indicators' validity, reliability, and on-farm feasibility.

This paper reviews animal-based welfare indicators in order to develop a valid, reliable and feasible on-farm welfare assessment protocol for dairy goats. The indicators were considered in the light of the four accepted principles (good feeding, good housing, good health, appropriate behaviour) subdivided into the 12 criteria developed by European Welfare Quality program. We will only examine the practical indicators to be used on-farm, excluding those requiring the use of specific instruments or laboratory analysis and those that are recorded at the slaughterhouse. Body condition score, hair coat condition and queuing at the feed barrier or at the drinker seem the most promising indicators for the assessment of the “good feeding” principle. As to “good housing”, some indicators were considered promising for assessing “comfort around resting” (e.g., resting in contact with a wall) or “thermal comfort” (e.g., panting score for the detection of heat stress and shivering score for the detection of cold stress). Several indicators related to “good health”, such as lameness, claw overgrowth, presence of external abscesses and hair coat condition, were identified. As to the “appropriate behaviour” principle, different criteria have been identified: agonistic behaviour is largely used as the “expression of social behaviour” criterion, but it is often not feasible for on-farm assessment. Latency to first contact and the avoidance distance test can be used as criteria for assessing the quality of the

human-animal relationship. Qualitative behaviour assessment seems to be a promising indicator for addressing the "positive emotional state" criterion. Promising indicators were identified for most of the considered criteria; however, no valid indicator has been identified for "expression of other behaviours". Interobserver reliability has rarely been assessed and warrants further attention; in contrast, short-term intraobserver reliability is frequently assessed, and some studies consider mid- and long-term reliability. The feasibility of most of the reviewed indicators in commercial farms still needs to be carefully evaluated, as several studies were performed under experimental conditions. Our review highlights some aspects of goat welfare that have been widely studied, but some indicators need to be investigated further and drafted before being included in a valid, reliable and feasible welfare assessment protocol. The indicators selected and examined may be an invaluable starting point for the development of an on-farm welfare assessment protocol for dairy goats.

6.1 INTRODUCTION

Consumer demand for assurance schemes of high quality animal products, in terms of health, safety and respect of animal welfare, has been increasing over the last few decades. In response to this demand, the assessment of animal welfare at farm level has become one of the most debated issues in the field of animal husbandry. This topic has been widely discussed at the international level, and species-specific protocols for on-farm welfare assessment are presently a major concern worldwide and for the European Union (EU) agricultural policy (Blokhuis et al., 2013).

Welfare assessment requires a multidimensional approach (Mason & Mendl, 1993), corresponding to a multi-criteria evaluation issue, and it should aim at determining the actual welfare of animals, including both physical and mental state (EFSA, 2012). Different indicators need to be included in efficient welfare assessment schemes, as all are important and they cannot compensate for each other (Blokhuis et al., 2010).

In 2008, the EU Welfare Quality project re-elaborated the concept of the "Five Freedoms" of animals (Brambell Committee, 1965) and defined four main areas of animal needs ("Welfare Principles"), which were then split into twelve independent criteria (Blokhuis et al., 2010; Rushen et al., 2011), each of which corresponded to a key welfare question. Welfare principles and criteria are as follows:

1. Good feeding: absence of prolonged hunger, absence of prolonged thirst;
2. Good housing: comfort around resting, thermal comfort, ease of movement;
3. Good health: absence of injuries, absence of disease, absence of pain induced by management procedures;
4. Appropriate behaviour: expression of social behaviours, expression of other behaviours, good human-animal relationship, positive emotional state.

Each criterion includes specific indicators that may be used to assess each component of welfare (Rushen et al., 2011). Although the same indicator may provide information related to different welfare concerns, criteria are independent of each other and form a basic but complete list (Blokhuis et al., 2010).

Two broad categories of indicators can be used to assess animal welfare at farm level: animal--based and resource-based indicators (Main et al., 2003). The need to focus on animal-based emerged clearly from the EU Welfare Quality project (Blokhuis et al., 2010); however, few available indicators are centered directly on the animals (Johnsen et al., 2001), and they rarely target small ruminants. A recent review on the monitoring of on-farm welfare in small ruminants points out only a few animal-based candidate indicators and most of them deal with sheep (Caroprese et al., 2009). Resource-based indicators have been more frequently adopted in welfare assessment protocols, because measurements taken are usually quick and easy [e.g., the Animal Needs Index TGI 35L developed by Bartussek (1999) for several species]. Nevertheless, good management and environmental resources do not necessarily result in a high standard of welfare (Winckler, 2006). An animal-based approach seems more appropriate for measuring the actual welfare state of the animals. This represents a considerable change in perspective, a shift from a scheme that mainly measured environmental aspects (which may show a high variation from country to country due to different housing and management conditions) towards one that measures the way in which the animal itself responds to such an environment (EFSA, 2012). Furthermore, individuals with different genetic backgrounds (e.g., different breeds) may, in fact, respond differently to the same environment. Although specific examples for goats are not currently available, this has been observed in other ruminant species. For example, in dairy cattle, Mattiello et al. (2011) pointed out that individuals with different genetic backgrounds showed different levels of welfare under similar environmental conditions. This supports the decision to focus mainly on animal-based indicators rather than exclusively on resource-based ones.

The aim of this paper was to review promising animal-based indicators that could be used to set up a valid, reliable, feasible, and practical on-farm welfare assessment protocol for dairy goats, centered on the evaluation of lactating animals.

6.2 METHODOLOGY

This review is part of the Animal Welfare Indicators (AWIN) integrated 7FP project, funded by the European Commission, which is aimed at developing practical on-farm welfare assessment protocols for several species, including goats. Studies carried out for pinpointing animal-based indicators to be included in the protocols are still underway.

A review of the scientific literature to date was the starting point for identifying promising indicators. Databases (Web of Science, CAB Abstracts, PubMed and Scopus) were searched for English language studies addressing animal-based goat welfare indicators as of (and including) 1990. Key words such as “welfare”, “measure”, “indicator”, “assessment”, “disease”, “pain”, “human-animal relationship”, “body condition”, and “lameness” were used as major descriptors combined with “goat” or “small ruminant”. Most of the reviewed literature dealt with dairy goats; however, although lactating dairy goats are our main target, papers considering other productive categories (e.g., kids, dry goats) and goats farmed for different purposes, or even other species, were taken into account wherever they provided evidence to support the use of indicators that could be included in an on-farm welfare assessment protocol for lactating dairy goats.

In this review, we refer specifically to the most widespread management system for dairy goats in Europe and North America, which consists of intensive housing systems where goats are kept indoors, with occasional access to pasture on some farms. In these systems, dairy goats are usually housed on straw litter, receive a total mixed ration or forage (mainly hay) and concentrate feed once or twice per day, and are milked twice a day in a milking parlour. Kids are usually separated from their mothers after birth.

We excluded indicators that focus exclusively on resources and management, as well as animal-based indicators that require further laboratory analysis (e.g., metabolic profiling), may be time-consuming (e.g., observations performed by video-recording), or may require the use of specific instruments (e.g., stethoscope, thermometer, heart rate monitor, or automatic devices to record behaviour; Desnoyers, Béchet, Duvaux-Ponter, Morand-Fehr & Giger-Reverdin, 2009; Mononen et al., 2012). We also excluded indicators that could only be recorded at the slaughterhouse.

In addition, we did not include routinely collected herd data, such as milk production and composition or fertility indexes, although we acknowledge their potential importance in assessing animal welfare. These data are often already available in national databases, especially for dairy cows. Their potential value in estimating animal welfare is recognized, even if only as a pre-screening tool. In real terms, in dairy cows, such data would seem to indicate a high prevalence of herds with apparent welfare problems, which is not always the case and needs to be confirmed by on-farm assessment (de Vries et al., 2014).

All selected published indicators were classified according to the four principles and 12 criteria of Welfare Quality assessment protocols (Welfare Quality, 2009a, 2009b). Some indicators (e.g., body condition score and hair coat condition) seemed promising for providing information on more than one criterion. In those cases, validity is discussed in relation to each pertinent criterion.

Promising indicators are summarized in tables, including information on animal category, housing, sample size, validity, reliability and on-farm feasibility. The latter three attributes (see definitions in Table 2) for each potential indicator were previously discussed and agreed upon by a group of experts on goat welfare during a meeting of the AWIN project, held in Milan (Italy) in November 2011.

Table 2. Definitions of attributes (validity, reliability and feasibility) used to identify potential promising welfare indicators in adult dairy goats

	Definition		Reference
Validity	The relation between a variable and what it is supposed to measure or predict. Criterion-related validity picks one or more criteria or standards for evaluating a scale, such as a predictive or a concurrent measure.		Acock, 2008
	Predictive validity	Ability of an indicator to predict some later criterion, such as a state of disease, or hunger, or discomfort, and so on.	Kamphaus & Frick, 2005
	Concurrent validity	Significant correlation between an indicator and other measures to which it is theoretically related (i.e., gold standard).	Kamphaus & Frick, 2005
Reliability	The extent to which a measurement is repeatable and consistent.		Martin & Bateson, 2007
	Intraobserver reliability	The agreement between successive observations of the same individual or group by a single observer, based on statistical significance of correlations ($p < 0.05$) or to Kendall's Coefficient of Concordance (> 0.7). According to time between measurements, reliability may be classified in short-(1-7 days), medium-(1 week – 1 month), or long-term reliability (> 1 month).	
	Interobserver reliability	The agreement between different observers during a simultaneous observation, based on statistical significance of correlations ($p < 0.05$) or to Kendall's Coefficient of Concordance (> 0.7).	
On-farm feasibility	The practical chance of using the indicators during on-farm inspection. It considers several constraints.		Knierim & Winckler, 2009
	Time constraints	Our survey among the stakeholders pointed out that a maximum acceptable time for a welfare protocol on-farm should be less than 2 h/farm and less than 5 min/animal. Depending on the number of indicators to be collected, the maximum time that can be spent for each indicator may vary; however, we considered feasible the indicators with an estimated duration < 30 min for a pen and of <1 min/animal for individual assessments. Indicators should not require to be further processed after collection (e.g., for laboratory analysis).	

Table 2. (continuation)

Cost	The indicator should not be expensive (e.g., in terms of specific equipment, consumables or laboratory costs).
Actual possibility to perform on commercial farms and stakeholders' and farmers' acceptability	The indicator should not require more than one person to be collected, should not require to alter the farm routine (e.g., moving animals out of the pen, or altering feeding or milking time, etc.), should not require a specific location to be recorded, should not cause stress to the animals (e.g., isolation, fear), should not require individual identification of the animals, should be easily recorded on all animals or on a representative sample of animals.

As AWIN's aim is to have a high level of acceptability from the stakeholders in the development of protocols, stakeholder opinion was also taken into account in the drafting of promising indicators. Opinions were collected in different ways: a link to an online survey (translated into 5 languages) was available for 15 months on the AWIN project website and other institutional sites (e.g., Food and Agricultural Organization of the United Nations; two stakeholder meetings were organized (one in Italy and one in Portugal) and attendees were asked to complete a questionnaire. Thirty-eight individuals from different European and non-European countries (e.g., Italy, Spain, Portugal, United Kingdom, United States, Brazil, Mexico, and Australia) answered our online survey, whereas 21 out of 40 people surveyed in Italy, and 11 out of 21 people surveyed in Portugal responded to the questionnaire during the stakeholder meetings. Both the online survey and questionnaire involved veterinarians, farmers, technicians, and advisors. The results of these interviews are reported in the text, when appropriate.
Only the most promising animal-based indicators (in terms of validity, reliability and feasibility) will be explained and discussed further.

6.3 GOOD FEEDING PRINCIPLE

The good feeding principle considers criteria related to the absence of prolonged hunger and of prolonged thirst by ready access to adequate diet and fresh water so as to maintain full health and vigour. All the indicators for these criteria are presented in Table 3.

Table 3. Animal-based indicators for assessing good feeding, excluding physiological measurements in dairy goats

	Animal-based welfare indicator	Age class [1]	Sex [2]	Housing system [3]	Sample size	Validity [4]	Intraobserver reliability [4]	Interobserver reliability [4]	On-farm feasibility [4]	References
Absence of prolonged hunger	Body condition score	A	F	C	1,520	O	-	-	Y	Anzuino et al., 2010
		J	M	E	60	Y	L	-	Y	McGregor & Butler, 2008
	Queuing (at the feed barrier)	A	F	E	48	Y	S	-	N	Jørgensen et al., 2007
	Displacements (at the feed barrier)	A	F	E	48	Y	S	-	Y	Jørgensen et al., 2007
		K	F/M	E	30	Y	M	-	Y	Van et al., 2007
		A/J	F	E	70	Y	L	-	N	Aschwanden et al., 2009b
		K	.	E	13	-	-	-	N	Mazurek et al., 2005
	Hair coat condition	A	F	E	48	-	-	-	Y	Battini et al., 2013
	Anal soiling	A	F	C	1,520	-	-	-	Y	Anzuino et al., 2010
	Feeding (simultaneously)	A/J	F	E	96	Y	M	-	N	Aschwanden et al., 2009a
	Agonistic interactions (with physical contact)	A/J	F	E	96	Y	M	-	N	Aschwanden et al., 2009a
		A/J	F	E	70	N	L	-	N	Aschwanden et al., 2009b
		K	F/M	E	30	Y	M	-	N	Van et al., 2007
		A	F	E	48	Y	S	-	Y	Jørgensen et al., 2007
	Agonistic interactions (without physical contact)	A	F	E	48	N	S	-	Y	Jørgensen et al., 2007
		K	F/M	E	30	Y	M	-	Y	Van et al., 2007
	Feeding bouts (duration)	A/J	F	E	70	Y	L	-	N	Aschwanden et al.,2009b
	Feeding (duration)	A	F	E	48	Y	S	-	N	Jørgensen et al., 2007
		K	.	E	13	-	-	-	N	Mazurek et al., 2005
	First feeding-place change (latency)	A/J	F	E	96	Y	M	-	N	Aschwanden et al., 2009a
	Feeding-place changes (frequency)	A/J	F	E	96	Y	M	-	N	Aschwanden et al., 2009a
	Feeding at night (duration)	A/J	F	E	70	N	L	-	N	Aschwanden et al., 2009b
	Standing/walking	A	F	E	48	N	S	-	N	Jørgensen et al., 2007
	Lying	A	F	E	48	Y	S	-	N	Jørgensen et al., 2007

Table 3. (continuation)

Absence of prolonged thirst	Queuing (at the drinker, duration)	A	F	E	30	Y	S	-	N	Ehrlenbruch et al., 2010b
	Displacements (at the drinker, frequency)	A	F	E	30	Y	S	-	N	Ehrlenbruch et al., 2010b
	Drinking (duration)	A	F	E	30	Y	S	-	N	Ehrlenbruch et al., 2010b
		A	F	E	30	N	S	-	N	Ehrlenbruch et al., 2010b
		A	F/M	E	14	N	S/L	-	N	Ogebe et al., 1996
	Drinking (frequency)	A	F	E	30	Y	S	-	N	Ehrlenbruch et al., 2010b
		A	F/M	E	14	N	S/L	-	N	Ogebe et al., 1996

[1] Age class: A = adult (>6 months); J = juvenile (3-6 months); K = kid (<3 months).
[2] Sex: M =male; F = female; - = information not available.
[3] Housing system: C = commercial farm, E = experimental farm.
[4] Validity, intraobserver and interobserver reliability, on-farm feasibility: Y = tested and valid or reliable or feasible; N = tested but not valid or not reliable or not feasible; - = not tested; O = validity assessed in other species; S = short-term reliability; M = mid-term reliability; L = long-term reliability.

Absence of Prolonged Hunger

Changes in the nutritional status of the animals have a substantial impact on the animal's health, welfare, and hence on production. Body condition scoring (BCS) is a method for subjective assessment of the nutritional status of farm animals based on the estimation of their body fat. It is considered a valid welfare indicator in many species: cattle (Welfare Quality, 2009a, 2009b; Winckler et al., 2003), buffalo (De Rosa et al., 2009; Winckler et al., 2003), sheep (Phythian et al., 2011; Russel, Doney & Gunn, 1969), and goats (Anzuino et al., 2010; McGregor & Butler, 2008; Santucci et al., 1991). Its concurrent validity has been confirmed by Russel et al. (1969) and Santucci et al. (1991), who found that BCS is a good predictor of fat deposits and is highly correlated with carcass fat content. Santucci et al. (1991) also found a significant correlation between body fat content assessed by BCS and fertility, with goats in lower levels of BCS having delays in conception. McGregor and Butler (2008) observed a correlation between BCS and mortality, finding that in cold weather conditions, mortality increased rapidly when the BCS of goats was < 2.0 (5-point scale). A wide range of BCS systems have been developed and used for research purposes and practical monitoring on commercial farms. The main distinctions among systems are whether they are merely visual or require palpation (or both), and whether the animal is assessed as a whole or separate scores are given for different anatomical regions, which are then summarized or adjusted to give a whole animal score (as reviewed by Leach, Knierim & Whay, 2009). The assessment of all the reviewed welfare indicators always requires specific training; however, BCS assessment in goats is particularly challenging without specific training and previous experience, because these animals generally have important visceral and internal fat deposits rather than subcutaneous fat (McGregor & Butler, 2008). McGregor and Butler (2008) applied a BCS based on 5-point scale (from 1 to 5) identical to that described by Jefferies (1961) for sheep, whereas Santucci et al. (1991) proposed a 6-point scale (from 0 to 5). However, the purpose of including BCS in on-farm welfare assessment schemes is to identify animals that are either too thin or too fat; hence, the scoring system does not need to be extremely detailed. For example, in the welfare protocols for cattle set up by Welfare Quality (2009a), only three BCS levels were adopted. This reasoning was followed by Anzuino et al. (2010), who also divided goats into obviously thin and obviously fat animals, differentiating animals by visual appraisal; however, we have no evidence of the validity of this simplified approach. Until now, BCS on goats has been assessed only by palpation, because this seems the most reliable method. However, this procedure requires goats to be individually restrained for palpation, and this is not always feasible in a short time span, especially on large farms. The development of a reliable visual method to highlight animals in extreme nutritional conditions (i.e., too thin or too fat) from outside the pen would be useful to reduce the time required, thus increasing feasibility.

Behaviour and social interaction during feeding time can be good indicators to evaluate the absence of prolonged hunger. When given the opportunity (e.g., *ad libitum* feed distribution), goats may eat up to seven to 10 h/day (Ferreira et al., 2013): they generally feed during two long periods (four hours each; early and late in the daytime), with several small meals in between. On intensive dairy farms, goats usually receive restricted feed twice a day (Görgülü et al., 2008). This management procedure may alter nutritional condition (e.g., reduction in feed intake; Görgülü et al., 2008) and behavioural patterns (e.g., coping strategy; Görgülü et al., 2008; Jørgensen, Andersen & Bøe, 2007). Feed can be a limited resource because the amount of feed is restricted, or because the feed type and composition are not appropriate (e.g., the roughage:concentrate ratio is too low, therefore feed is consumed very quickly and competition is increased), or because feeding space is not accessible for all individuals in the group at the same time (Görgülü et al., 2008; Jørgensen et al., 2007). In a competitive environment, such as on intensive or semi-intensive dairy farms, these problems can be partly overcome by adopting a coping strategy that consists of consuming the feed at different times of the day to optimize access to the feed trough (Jørgensen et al., 2007; Shinde, Verma & Singh, 2004). However, under these competitive circumstances, low-ranking goats may have access to lower quality feed and may experience a negative emotional state similar to the frustration caused by the time spent queuing at the feed trough. Carbonaro, Friend, Dellmeier & Nuti (1992) have documented frustration related to food thwarting in dairy goats, confirming that frustration may elicit physiological alterations (e.g., an increase of norepinephrine) and behavioural reactions (e.g., pawing, head movements, rearing).

Queuing animals at the feeding rack may be a promising indicator, as this behaviour is exacerbated by the increasing number of goats per feeding place (Jørgensen et al., 2007), thus confirming its predictive validity. However, the use of 24-h video-recording adopted by Jørgensen et al. (2007) is not feasible for an on-farm welfare evaluation protocol; hence, different observation strategies need to be explored (e.g., direct data collection per segment/pen in a predetermined timeframe, as already adopted in a dairy cow protocol; Laister et al., 2009). Although goats usually prefer to adopt the queuing strategy to cope with reduced access to feed, displacements also proved to be a valid indicator in assessing absence of prolonged hunger in both adult female goats (Aschwanden, Gygax, Wechsler & Keil, 2009b; Jørgensen et al., 2007) and kids (Van, Mui & Ledin, 2007), as well as in other species, such as pigs and cattle (Nielsen, Lawrence & Whittemore, 1995; Olofsson, 1999). The number of displacements may increase not only in response to a competitive environment (Aschwanden et al., 2009b; Jørgensen et al., 2007; Van et al., 2007), but also in response to feed composition and hence feed preferences (e.g., hay *vs* silage, Jørgensen et al., 2007; jackfruit *vs* concentrate; Van et al., 2007).

Some researchers suggest that hair coat condition can be used as a first warning of a goat's nutritional status and health (Lengarite, Mbugua, Gachuiri & Kabuage, 2012; Sarkar et al., 2010; Smith & Sherman, 2009a; Veit, McCarthy, Friedericks, Cashin & Angert, 1992); this concept was widely accepted by farmers and technicians. Recent research seems to confirm that this indicator can be valid and practical for on-farm welfare assessment, as goats with rough or scurfy hair can be easily identified and present a very low BCS (Battini et al., 2013).

Anal soiling is another interesting indicator of good feeding, as it reflects problems with nutrition and digestion (Grove-White, 2004); namely, ruminal acidosis (Braun, Rihs & Schefer, 1992). The validity of this indicator has never been assessed in goats, but it is accepted and used in cattle (Leach, Knierim & Whay, 2009). It is considered a feasible indicator for goats (Anzuino et al., 2010); however, it is difficult to assess it in the pen and might be best recorded in the milking parlour.

Jørgensen et al. (2007) consider the video-recorded 24-h time budget of general activity behaviours - feeding, walking, standing and lying - in relation to hunger (Table 3). Their results confirm the validity of feeding time (that significantly decreased when the number of goats per feeding place increased) but not of the other behaviours considered. Furthermore, a prolonged observation time is required in order to obtain a reliable figure of the time budget (Martin & Bateson, 2007), which reduces its on-farm feasibility.

Feed intake (difference between weight of daily offered feed and weight of residues) is another valid indicator of the absence of prolonged hunger, as it can be reduced by insufficient space availability (Jørgensen et al., 2007). However, this indicator is not measured directly on the animals, is time consuming (in terms of time required to weight the residual roughage), and therefore not feasible for our purposs.

Absence of Prolonged Thirst

Available literature sources give scarce indication as to animal-based indicators for evaluating the absence of prolonged thirst. Ready access to fresh water is important to maintain full health and vigour; welfare can be compromised if animals cannot drink whenever they feel the need to, either because fresh water is not available or because of competition with other goats. Continuous and prolonged lack of access to fresh water may eventually lead to chronic dehydration, especially during the hot periods (Darcan, Cedden & Cankaya, 2007; Ogebe, Ogunmodede & McDowell, 1996).

As mentioned before, although goats may adopt coping strategies in competitive environments (e.g., intensive farms; Ehrlenbruch, Pollen, Andersen & Bøe, 2010; Jørgensen et al., 2007), which may lead to change in social behaviours, they mainly prefer drinking around feeding time (Rossi & Scharrer, 1992). This behaviour is generally socially facilitated (Forkman, 1996) and quite synchronized (Rook & Penning, 1991). Reduced possibility of simultaneous drinking can lead to decreased drinking time and hence lower water intake (Ogebe et al., 1996; Van et al., 2007). Queuing animals and displacements at the drinkers may be used to detect animals suffering from thirst. During a 2-h video-recording, Ehrlenbruch et al. (2010b) found increased agonistic behaviour and queuing when the ratio of nipples to goats was less than 1:15, a situation that may occasionally occur on commercial farms. As already stated, the use of cameras is not feasible for practical on-farm use, because it requires additional time to analyse the recorded information; therefore, other more feasible observation strategies should be identified to collect data regarding these indicators. A possible suggestion may be to concentrate direct observations on queuing animals and displacements at the drinkers within a short time after feeding (Rossi & Scharrer, 1992), making use of suitable observation strategies that need to be specifically validated for this aim. Other indicators used in previous studies to detect water deficiencies (e.g., respiratory rate, rectal temperature, body mass, daily outputs of urine and faeces, haematocrit values and plasma volume; Rahardja, Toleng & Lestari, 2011; Al-Ramamneh, Riek & Gerken, 2012) are not feasible for on-farm welfare evaluation, because they require further laboratory analysis or the use of specific instruments. However, they can be useful for validating other, more feasible, indicators.

6.4 GOOD HOUSING PRINCIPLE

This principle involves criteria related to farm structures and housing conditions. All the reviewed indicators for these welfare criteria are presented in Table 4.

Table 4. Animal-based indicators for assessing good housing, excluding physiological measurements, in dairy goats

	Animal-based welfare indicator	Age class [1]	Sex [2]	Housing system [3]	Sample size	Validity [4]	Intraobserver reliability [4]	Interobserver reliability [4]	On-farm feasibility [4]	References
Comfort around resting	Resting (in contact with wall)	A	F	E	24	Y	S	-	Y	Andersen & Bøe, 2007
		A	F	E	24	Y	S	-	Y	Ehrlenbruch et al., 2010a
		A	F	C	40	N	M	-	Y	Loretz et al., 2004
	Resting (in contact with other goats)	A	F	E	24	N	S	-	N	Andersen & Bøe, 2007
		A	F	C	40	N	M	-	N	Loretz et al., 2004
		A	F	E	24	N	S	-	N	Ehrlenbruch et al., 2010a
	Resting (synchronously)	A	F	E	24	N	S	-	N	Andersen & Bøe, 2007
	Resting (duration)	A/J	F	E	70	N	L	-	N	Aschwanden et al., 2009b
		A	F	C	40	Y	M	-	N	Loretz et al., 2004
	Resting (in the activity area)	A	F	E	24	N	S	-	N	Andersen & Bøe, 2007
	Cleanliness	A	F	C	1,520	O	-	-	Y	Anzuino et al., 2010
	Average distance (between lying animals)	A	F	C	40	N	M	-	N	Loretz et al., 2004
	Nosing on/exploring another goat (duration)	A	F	E	24	N	S	-	N	Andersen & Bøe, 2007
	Agonistic interactions (with physical contact)	A	F	E	24	N	S	-	N	Andersen & Bøe, 2007
		A	F	E	24	N	S	-	N	Ehrlenbruch et al., 2010a
	Agonistic interactions (without physical contact)	A	F	E	24	N	S	-	N	Andersen & Bøe, 2007
		A	F	E	24	N	S	-	N	Ehrlenbruch et al., 2010a
	Displacements	A	F	E	24	Y	S	-	N	Andersen & Bøe, 2007
		A	F	C	40	N	M	-	N	Loretz et al., 2004
		A	F	E	24	N	S	-	N	Ehrlenbruch et al., 2010a
	Time budget	A	F	E	24	N	S	-	N	Andersen & Bøe, 2007

Table 4. (continuation)

Thermal comfort	Panting score	A	F	E	30	Y	S	-	Y	Darcan et al., 2007
		A	F	C	1,207	Y	-	-	Y	Fioni, 2014
	Shivering score	A	F	C	1,207	Y	-	-	Y	Fioni, 2014
	Huddling	A	F	E	20	N	-	-	N	Bøe & Ehrlenbruch, 2013
	Lying (duration)	A	F	E	9	Y	L	-	N	Bøe et al., 2007
	Standing (duration)	A	F	E	9	N	L	-	N	Bøe et al., 2007
	Moving (duration)	A	F	E	9	Y	L	-	N	Bøe et al., 2007
	Feeding (duration)	A	F	E	9	Y	L	-	N	Bøe et al., 2007
		A	F/M	E	14	Y	S/L	-	N	Ogebe et al., 1996
		A	F	E	30	Y	S	-	N	Darcan et al., 2007
	Ruminating (duration)	A	F	E	30	Y	S	-	N	Darcan et al., 2007
	Drinking (duration)	A	F	E	30	Y	S	-	N	Darcan et al., 2007
		A	F/M	E	14	N	S/L	-	N	Ogebe et al., 1996
	Walking (duration)	A	F	E	30	Y	S	-	N	Darcan et al., 2007
	Resting (duration)	A	F	E	30	Y	S	-	N	Darcan et al., 2007
	Drinking (frequency)	A	F/M	E	14	N	S/L	-	N	Ogebe et al., 1996
	Rumination rate	A	F/M	E	14	Y	S/L	-	N	Ogebe et al., 1996
Ease of movement	Kneeling	A	F	C	116 pens	Y	-	-	Y	Anzuino et al., 2010
	Standing up score	A	F	E	35	Y	-	-	Y	Mazurek et al., 2007
	Leaving the feed barrier (duration)	A	F	E	55	Y	M	-	N	Nordmann et al., 2011

[1] Age class: A = adult (>6 months); J = juvenile (3-6 months); K = kid (<3 months).
[2] Sex: M =male; F = female.
[3] Housing system: C = commercial farm, E = experimental farm.
[4] Validity, intraobserver and interobserver reliability, on-farm feasibility: Y = tested and valid or reliable or feasible; N = tested but not valid or not reliable or not feasible; - = not tested; O = validity assessed in other species; S = short-term reliability; M = mid-term reliability; L = long-term reliability.

Comfort Around Resting

As defined in Welfare Quality (2009a), "animals should have comfort when they are resting". When goats have the possibility to choose, they prefer to rest against a wall rather than in the middle of the pen (Andersen & Bøe, 2007; Ehrlenbruch, Jørgensen, Andersen & Bøe, 2010), as already observed in other farm animals and summarized by Ehrlenbruch et al. (2010a; cattle: Stricklin, Graves & Wilson, 1979; sheep: Bøe, Berg & Andersen, 2006; Færevik, Andersen & Bøe, 2005; Marsden & Wood-Gush, 1986; fowl: Cornetto & Estevez, 200). This may be due to increased comfort or to an anti-predatory strategy, suggesting that the animals may feel safer close to a wall than in an open area. If space allowance is reduced, goats are forced to choose different areas, including those without walls, in which to lie down. Resting in contact with the wall showed a predictive validity in Andersen & Bøe (2007) and

Ehrlenbruch et al. (2010a) in a lying size area ranging from 0.5 to 1.0 m^2/goat, but not in Loretz, Wechsler, Hauser & Rüsch (2004), where a larger lying area was provided (1.0 to 2.0 m^2/goat). These findings suggest that, below a given individual space availability, goats cannot choose their preferred resting areas. This supports the hypothesis of using this indicator to detect comfort around resting. Feasibility needs to be improved, as data collection in the cited studies was performed through prolonged video recording, resulting in a method that is too time consuming. Direct observation needs to be further studied before being efficiently applied. Moreover, it is crucial to identify the right moment to observe animals. Generally, they tend to have synchronized resting patterns (Rook & Penning, 1991). However, the percentage of observations when all animals rest simultaneously may vary depending on the size of the lying area; for example, Andersen & Bøe, (2007) observed that the percentage of observations with all goats lying simultaneously ranged from 8.5% at 0.75 m^2/goat to 21.1% at 1.00 m^2/goat. This suggests that, when enough space is provided, goats tend to rest simultaneously and therefore the level of synchronization can give reliable information about comfort around resting. However, an observation time of 24-h at 10-min intervals, as used by Andersen & Bøe (2007), cannot be considered feasible for an on-farm protocol.

Goats dislike wet areas when resting. Cleanliness is already used as a valid welfare indicator in pigs (Hughes, 2001), poultry (Scott et al., 2007), and cattle (Andreasen & Forkman, 2012). Dairy goats are much cleaner than dairy cattle, because they generally have drier faecal matter and are usually housed on straw bedding year round. Manure management is much easier in goats than in cattle, but the way in which goats are handled and moved to the milking parlour, as well as the cleanliness and dryness of walkways, may significantly influence the goat's cleanliness. In dairy cattle, the main factors affecting the cleanliness of limbs are the cleaning frequency of barn alleyways, the ease in moving procedures, group density, and the number of times animals are moved (Hughes, 2001; Schreiner & Ruegg, 2003). Anzuino et al. (2010) used cleanliness as a possible welfare indicator; however, the best location to score it has yet to be identified. The milking parlour may be a good location to record cleanliness of the rear body region, but other locations should be identified, in order to gather information about the front body area, such as the sternum region over which goats lie. Observation of cleanliness may be very time consuming in large herds. In these cases, the development of a representative sampling strategy may be required.

Based on previous results by Loretz et al. (2004), Andersen and Bøe (2007) hypothesized that lying time would decrease when space availability was reduced; therefore, they used the 24-h time budget (time spent lying, standing, and moving in the resting area and in the activity area) as an indicator of comfort around resting. However, their results do not support the validity of time budget in relation to comfort around resting. Furthermore, as already mentioned, on-farm feasibility for time budget assessment is low because of the long observation time required (Martin & Bateson, 2007).

Thermal Comfort

It is known that ruminants have a broad thermal comfort zone and a high degree of thermal tolerance (Sejian & Srivastava, 2010). Scarce information about the limits of thermal comfort, allowing goats to maintain a near-constant body temperature of approximately 39°C, is available in literature. Toussaint (1997) suggested that adequate temperatures for goats kept indoors range from 6°C to 27°C (optimum from 10° to 18°C), with relative humidity from 60 to 80%. Beyond the limits of thermal comfort, behavioural and physiological changes may occur that reduce or increase the heat loss (Bøe & Ehrlenbruch, 2013; Darcan et al., 2007; McGregor, 2002; Mount, 1979).

Inadequate temperatures, high humidity, and wind and rain are the main factors affecting thermal comfort in goats (Bøe & Ehrlenbruch, 2013; McGregor, 2002). Observation of respiratory rate can provide reliable and practical information for estimating the severity of heat stress in farm animals (Silanikove, 2000). A panting score has already been used for cattle (Gaughan, 2003) and it showed a predictive validity in goats (Darcan et al., 2007; Fioni, 2014). The score is assigned based on the visual observation of behaviour, using a 5-point (Darcan et al., 2007) or a simplified 3-point (Fioni, 2014) scale. The 3-point score system proved sufficient to highlight a condition of thermal discomfort, such as in presence of severe heat stress situations, where goats showed signs ranging from accelerated respiration with mouth closed (score 1) to panting with open mouth and excessive salivation (score 2).

Although goats are frequently described as rustic or highly adaptable animals, research supports the fact that goats are negatively impacted by low temperatures (Bøe & Ehrlenbruch, 2013; McGregor, 2002), to the extent that shivering sets when they are exposed to low critical temperatures (Fioni, 2014; Mount, 1979). A feasible 3-point scale scoring system was developed and validated by Fioni (2014), who recorded signs of bristling hair on the back (score 1) and shivering with arched posture (score 2) related to severe cold stress.

Huddling is widely used by pigs to reduce heat loss at low temperatures (Andersen, Bøe & Hove, 2000), but this strategy is not described in studies of natural populations of goats (Shackleton & Shank, 1984), although some farmers report anecdotal evidence of this behaviour. Bøe and Ehrlenbruch (2013) recorded the lying behaviour (lying alone *vs* huddling) in five pre-defined weather conditions. Even though weather had a significant on huddling behaviour, there was no clear evidence of its relation to decreasing temperatures.

Some physiological and blood parameters have been identified as valid indicators. Water intake, rumination rate, rectal temperature, pulse and respiration rate (Darcan et al., 2007;

Ogebe et al., 1996), skin temperature (Darcan et al., 2007), glucose, total cholesterol, urea, and cortisol (Sejian & Srivastava, 2010) increase during heat stress situations. In contrast, total protein and aldosterone decrease under these conditions (Sejian & Srivastava, 2010). These indicators are not feasible, but they could be useful in future studies for validating indicators that are more practical for on-farm welfare assessment.

Some authors consider the time budget of general activity behaviours (e.g., feeding, lying, ruminating, drinking) as being related to thermal comfort ((Bøe, Andersen, Buisson, Simensen & Jeksrud, 2007; Darcan et al., 2007); Table 4), because the effect of low or high environmental temperatures may affect behavioural patterns. Results from these authors confirmed that activity rhythms increase at low temperature and decrease at high temperature. In spite of their validity, the measurement of such behaviours is not recommended for on-farm use, as feasibility is hampered by the length of time required to compile the time budget.

Ease of Movement

By ease of movement, we mean the freedom of the animals to explore their surroundings without injuring themselves; in other words, “animals should have enough space to be able to move around freely” (Welfare Quality, 2009a), without risk of injury, at an adequate density. The animals have to be in sound enough condition to be able to walk, lie down and stand up.

Anzuino et al. (2010) suggest kneeling (at the trough) is a promising welfare indicator to assess discomfort due to inappropriate housing conditions. Kneeling behaviour consists of goats dropping on their front knees. Anzuino et al. (2010) report that out of 24 British farms, 79.2% showed goats kneeling (at the trough), but no correlation between lame and kneeling goats (at the trough) was found. This confirms that a badly designed feeding trough (e.g., floor level bunk) may force both lame and sound goats to assume this position to reach the feed. The on-farm feasibility is presumably high for kneeling (at the trough) as this behaviour is easily observed.

A standing up-score has been used by Mazurek, Marie & Desor (2007) to detect problems in transition movement. The authors describe goats that stand up “lifting their back first” as animals with standing-up abnormalities. However, this behaviour is commonly described as the normal standing up transition movement for cattle (Albright & Arawe, 1997), and farmers and goat experts also consider it as the normal transition movement for goats; therefore, the standing-up score definition used by Mazurek et al. (2007) needs to be redefined. The main concern related to the feasibility of the standing up score is that the observer has to catch the precise instant when the transition movement starts. Therefore, observation of this indicator is closely linked to a specific moment, which lowers its on-farm feasibility.

Nordmann et al. (2011) used the duration of leaving the feed barrier to detect problems related to housing structure. This behaviour was described as "the duration taken by the goats to leave the feed barriers. It began when a goat started to move its head with the intention to leave the feed barrier and finished just when the whole head of the goat (including nose) had left the feed barrier." Although this indicator has been validated, the feed barrier design may influence the behaviour (Nordmann et al., 2011), and thus we consider it unfeasible in commercial farming conditions, as many goats are likely to leave the feed barrier at the same time, making the simultaneous direct observation of many animals very difficult.

6.5 GOOD HEALTH PRINCIPLE

Indicators related to health are the most quoted in our review because of the widespread availability of studies on this topic. All the reviewed indicators for this principle are presented in Table 5.

Table 5. Animal-based indicators for assessing good health, excluding physiological measurements, in dairy goats

	Animal-based welfare indicator	Age class [1]	Sex [2]	Housing system [3]	Sample size	Validity [4]	Intraobserver reliability [4]	Interobserver reliability [4]	On-farm feasibility [4]	References
Absence of injuries	Lameness	A	F	E	170	Y	S	-	Y	Christodoulopoulos, 2009
		K/J/A	F/M	C	100/76/308	O	-	-	Y	Eze, 2002
		A	F	C	1,520	O	-	-	Y	Anzuino et al., 2010
		A	F	E	40	O	-	-	Y	Mazurek et al., 2007
	Claw overgrowth	A	F	C	1,520	O	-	-	Y	Anzuino et al., 2010
	Lesions and swellings	A	F	C	1,520	O	-	-	Y	Anzuino et al., 2010
	Teats and udder abnormalities	A	F	C	1,520	O	-	-	Y	Anzuino et al., 2010
	Teats and udder conformation trait	A	F	C	1,520	O	-	-	Y	Anzuino et al., 2010
	Standing up score	A	F	E	35	Y	-	-	Y	Mazurek et al., 2007
Absence of disease	Body condition score	J	M	E	60	Y	L	-	Y	McGregor & Butler, 2008
		A	F	C	1,520	O	-	-	Y	Anzuino et al., 2010
		A	F	E	60	Y	S	-	Y	Laporte-Broux et al., 2011
		A	F	C	149	Y	-	-	Y	Santucci et al., 1991

Table 5. (continuation)

	Presence of abscesses	A	F	C	16	-	-	-	Y	Mantova, 2012
	Hair coat condition	A	F	E	48	Y	-	-	Y	Battini et al., 2013
	Discharges (nasal, ocular, vulvar)	A	F	C	1,520	-	-	-	Y	Anzuino et al., 2010
	Cleanliness	A	F	C	1,520	O	-	-	Y	Anzuino et al., 2010
	Obviously sick/dull goats	A	F	C	116 pens	-	-	-	N	Anzuino et al., 2010
	Pruritus	A	F	C	116 pens	-	-	-	N	Anzuino et al., 2010
	Coughing	A	F	C	116 pens	-	-	-	N	Anzuino et al., 2010
	Dyspnea	A	F	C	116 pens	-	-	-	N	Anzuino et al., 2010
	Hesitate/refuse movement	A	F	E	108	Y	-	-	N	Mazurek et al., 2005
		A	F	E	108	Y	-	-	N	Mazurek et al., 2007
	Time budget	A	F	E	28	N	S	-	N	Laporte-Broux et al., 2011
	Agonistic interactions (with physical contact)	A	F	E	60	N	S	-	N	Laporte-Broux et al., 2011
	Agonistic interactions (without physical contact)	A	F	E	60	N	S	-	N	Laporte-Broux et al., 2011
	Positive interactions	A	F	E	60	Y	S	-	N	Laporte-Broux et al., 2011
Absence of pain induced by management procedures	Ear tear	A	F	C	1,338	-	-	-	Y	Anzuino et al., 2010

[1]Age class: A = adult (>6 months); J = juvenile (3-6 months); K = kid (<3 months).
[2] Sex: M =male; F = female.
[3]Housing system: C = commercial farm, E = experimental farm.
[4]Validity, intraobserver and interobserver reliability, on-farm feasibility: Y = tested and valid or reliable or feasible; N = tested but not valid or not reliable or not feasible; - = not tested; O = validity assessed in other species; S = short-term reliability; M = mid-term reliability; L = long-term reliability.

Absence of Injuries

According to this criterion, animals should be free from physical problems that may affect their health; for example, skin damage and locomotion disorders (Welfare Quality, 2009a).

Lameness is taken into consideration in several papers, as it is an important indicator of pain (O'Callaghan, Cripps, Downham & Murray, 2003) and may lower productivity in dairy goats by reducing milk yield (Christodoulopoulos, 2009; Hill et al., 1997) and fertility (Christodoulopoulos, 2009; Eze, 2002; Hill et al., 1997), as well as contributing to pregnancy toxaemia (Lima et al., 2012b) and neonatal diseases (Eze, 2002), and hence premature culling (Hill et al., 1997). The main causes of lameness in intensively kept dairy goats are claw overgrowth with or without deformation and diseases affecting the limb joints, such as caprine arthritis and encephalitis (CAE) and caprine contagious agalactia (Bergonier, Berthelot & Poumarat, 1997; Hill et al., 1997; Smith & Sherman, 2009a; Winter, 2011). Lameness ranges from 9.1% to 24% in commercial farms (Christodoulopoulos, 2009; Eze, 2002; Hill et al., 1997; Mazurek et al., 2007). This large variation in prevalence of lameness can be due to various resources (e.g., access to pasture, indoor environment) and management practices (e.g., infrequent claw trimming), and especially to the fact that the authors assessed lameness with different methods and scoring systems, and therefore the definition of clinical lameness in goats is unclear. Lameness scores have already been validated in some species (e.g., cattle: Thomsen, Munksgaard & Tøgersen, 2008; sheep: Winter, 2008), but there are no well-developed, established gait scoring systems in goats, so lameness is generally scored using different point scales (Anzuino et al., 2010; Hill et al., 1997; Mazurek et al., 2007) or by classifying goats as lame or sound (Christodoulopoulos, 2009). Hill et al. (1997) and Christodoulopoulos (2009) both showed evidence of a strong association between lameness and some of the foot lesions identified in their studies (hoof separation, abscess of the sole, footrot, interdigital dermatitis). However, none of these studies assessed reliability when using these scoring systems, and hence, further studies are needed.

The location in which lameness assessment is performed varies: animals can be scored at the exit of the milking parlour (Anzuino et al., 2010) or in the group pen (Anzuino et al., 2010; Christodoulopoulos, 2009; Mazurek et al., 2007). According to Anzuino et al. (2010), the prevalence of lameness estimated while the goats are in their pens is usually much lower than that observed when the goats exit the parlour. This may be because the observer can see the goats better when they are exiting the parlour or because the goats' locomotion is better when they walk on soft straw surfaces in their pens than when they walk on hard surfaces. This finding is important, because it suggests that assessing lameness in goats while they are housed in their pens may underestimate the severity and prevalence of the problem. However, if lameness is going to be assessed when goats are entering or exiting the milking parlour, a reliable sampling strategy should be planned, as the prevalence of lameness may depend on milking order. For example, in cows, a higher prevalence of lame animals has been observed in the last groups entering the milking parlour, compared with the first groups (Main et al., 2010). Furthermore, this not only highlights the importance of standardizing the location and type of flooring to perform lameness assessment in goat farms, but also the procedure for evaluating animals (e.g., how many strides) to obtain comparable results across different farms.

The sensitivity of the scoring system may vary across species. In goats, a maximum of four levels has been adopted by Anzuino et al. (2010), whereas in cattle scales up to five (e.g., Hoffman, Moore, Vanegas & Wenz, 2014; Ito, Chapinal, Weary & von Keyserlingk, 2014; Kougioumtzis, Valergakis, Oikonomou, Arsenos & Banos, 2014; Thomsen et al., 2008) or six levels (e.g., Bell et al., 2009) have been used.

To make the scoring system more effective without reducing the on-farm feasibility, the drafting of a continuous lameness scale for individual goats would represent a significant advance in this field, like the modified visual analogue scale (VAS) recently established for cattle (Tuyttens et al., 2009). This modified scale has the advantage of enhancing sensitivity and, at the same time, producing continuous data that can be subjected to algebraic operations needed for aggregating the welfare indicators assessed in a protocol into a welfare index (Tuyttens et al., 2009). The reliability of this scoring system needs to be tested in goats.

Lameness is often correlated with claw overgrowth, which is a major problem in commercial dairy goat farms (Christodoulopoulos, 2009; Hill et al., 1997; Winter, 2011). In the UK, Anzuino et al. (2010) report that out of 1,520 sample animals, 79.8% had overgrown claws, and the problem was present at different levels of severity in all the farms surveyed. This high prevalence is probably due to a lack of hoof wear when animals are housed on straw bedding, or to poor management, such as an insufficient frequency of foot trimming (Anzuino et al., 2010; Eze, 2002; Hill et al., 1997). In dairy goats, claw overgrowth can be scored from moderate to severe, and a significant correlation with lameness has been demonstrated (Anzuino et al., 2010; Christodoulopoulos, 2009; Hill et al., 1997). The assessment of this indicator may be performed in the milking parlour and, if supported by an effective sampling strategy (similar to that reported for lameness; Main et al., 2010), can be feasible on-farm.

The presence of lesions (including skin damages, swelling, and hair losses) is another possible indicator of poor health. Anzuino et al. (2010) observed that body and neck skin lesions are quite frequent (19.9% and 14.2% out of 1,520 goats, respectively), but they consist mainly of hair loss, whereas most of the lower limb skin lesions (6.2%) consist of both skin damage and hair loss. Such lesions may not be painful but can still be important indicators of welfare, as they may reflect structure deficiencies (e.g., physical obstructions to normal behaviour) or may arise from trauma (e.g., hornless goats housed with horned goats; Aschwanden, Gygax, Wechsler & Keil, 2008) or ectoparasites (Smith & Sherman, 2009a).

In dairy animals, other possible animal-based indicators are lesions and abnormalities of teats and udders (Anzuino et al., 2010). In the same set of goats on UK farms mentioned above (Anzuino et al., 2010), teat and udder abnormalities in dairy goats were found with respective prevalences of 7.6% and 33.8%, including lesions, wounds, inflammations, and accessory teats. Teat and udder lesions can affect both welfare and production in dairy goats (Contreras et al., 2007; Leitner, Silanikove & Merin, 2008; Mavrogianni et al., 2011; Perrin, Mallereau, Lenfant & Baudry, 1997), but little published information describes the welfare significance and aetiology of different lesions (Menzies & Ramanoon, 2001). Certain aspects of dairy goat farming, such as rapid milking rates, large herd size, high milk production, number of stockpersons, and minimal hygiene routine at milking, as well as some specific goat behaviour, such as teat biting or self-sucking, may contribute to the development of teat and udder lesions (Anzuino et al., 2010; Martínez-de la Puente et al., 2011; Stelwagen & Knight, 1997; Torres, Castro, Hernández-Castellano, Argüello & Capote, 2013). Evaluating teat and udder conformation traits may also be important for goat welfare. Research on dairy goats showed a relationship between pendulous udders and mastitis (Ameh, Addo, Adekeye & Gyang, 1993; Deinhofer & Pernthaner, 1995), partly due to the increased risk of injury to the udder and teats when the distance between the teat ends and the floor is small. Udder asymmetry has been associated with chronic intramammary infection (e.g., CAE, contagious agalactiae, retroviral hard udder) causing induration and atrophy of one half (Alawa, Ngele & Ogwu, 2000; Krieg & Peterhans, 1990; Paterna et al., 2014). Ameh et al. (1993) and Ameh & Tari (1999) found teat injuries to be associated with mastitis. This evidence supports the hypothesis that udder conformation is a risk factor for disease that may compromise dairy goat welfare and that a chronic change remains even after an udder has recovered from infection or injury (Klaas, Enevoldsen, Vaarst & Houe, 2004; Krieg & Peterhans, 1990; Smith & Sherman, 2009a). Asymmetry and pendulous udders were recorded with prevalences of 22.0% and 5.3%, respectively, out of 1,520 observed goats in UK farms (Anzuino et al., 2010) and 5.7% and 10.2%, respectively, out of 423 goats in African breeds (Amao et al., 2003).

Teat and udder lesions, abnormalities, and conformation traits can easily be recorded by visual assessment from a short distance. The milking parlour is probably the best location to assess these indicators.

Absence of Disease

According to Welfare Quality (2009a), animals should be free from disease. Many indicators related to disease have never been validated in goat studies, but have been validated in cattle (as reviewed by Canali, Whay & Leach, 2009).

Body condition score is a valid and feasible indicator not only of the absence from prolonged hunger (as described and discussed for the good feeding principle), but also of the absence of disease. It is generally accepted that this indicator is important for identifying chronically ill goats; for example, BCS is decreased in the case of chronic contagious diseases, such as caseous lymphadenitis, paratuberculosis, or CAE, gastro-intestinal parasitism, painful conditions (arthritis, footrot, laminitis), or in animals that have dental problems (Mantova, 2012; Smith & Sherman, 2009a). In contrast, very thin or obese pregnant dairy goats risk pregnancy toxaemia (Brozos, Mavrogianni & Fthenakis, 2011; Lima, Pascoal & Stilwell, 2012a; Schlumbohm & Harmeyer, 2008), and BCS can therefore be successfully adopted to prevent the occurrence of this disease in periparturient goats (Laporte-Broux et al., 2011).

The presence of external abscesses is a common sign of pathology in goats. In most cases, external abscesses are located in lymph node areas (mainly mandibular, prescapular, prefemoral, and supramammary lymph nodes; Smith & Sherman, 2009a). This external sign is often associated with caseous lymphadenitis caused by *Corynebacterium pseutuberculosis*. This disease can also affect internal lymph nodes or organs, such as lungs, liver, or kidneys (Baird & Fontaine, 2007), but this form cannot be recorded by external examination and is not very common in goats (Smith & Sherman, 2009a). The presence of abscesses maybe a valid and feasible animal-based indicator for on-farm welfare assessment, reflecting a general poor condition of the animal, as shown by reduced feeding time and low BCS recorded in goats with external abscesses (Mantova, 2012).

Hair coat condition can also be regarded as an interesting indicator to gather information not only on a goat's nutritional status (as already described and discussed for the good feeding principle), but also on its health status. In fact, recent research showed a higher prevalence of abnormal lung sounds (probably related to chronic respiratory disease) and a general condition of poor health in goats with rough and scurfy hair (Battini, 2013; Battini et al., 2013).

Kneeling has been already discussed according to the "ease of movement" criterion. Although the behaviour is the same as that already described, Anzuino et al. (2010) distinguish between animals seen in kneeling posture at the trough and in the pen, as this behaviour may assume different meanings. Kneeling (in the pen) consists of goats standing or walking on their front knees, without being involved in exploratory or feeding behaviour. Such goats are generally not able to stand up and frequently reach the milking parlour or the feeding rack in this abnormal posture. The origin of this behaviour is still unclear: the significant correlation recorded by Anzuino et al. (2010) between the prevalence of goats kneeling (in the pen) and of severely lame goats recorded when exiting the milking parlour suggests that kneeling behaviour may be related to painful limb ailments. For example, kneeling is a common clinical sign in CAE and is frequent in infected farms (Adams, Klevjer-Anderson, Carlson, McGuire & Gorham, 1983; Smith & Sherman, 2009a). The prevalence of farms with goats kneeling (in the pen) recorded by Anzuino et al. (2010) in the UK was extremely high (75% out of 24 farms), supporting the need for further investigation. The on-farm feasibility for kneeling (in the pen) as an indicator is high, as this behaviour is not related to a specific moment and kneeling goats are quite easy to observe in the pen area.

Goats presenting discharges were recorded with low frequency on UK farms, where 5% of goats showed vulvar (mainly haemorrhagic), 0.6% nasal, and 6% ocular discharges (Anzuino et al., 2010; see Table 5 for sample size). Discharges have already been included in cattle welfare assessment protocols (Canali et al., 2009). Their feasibility is accepted, but the best location to observe them needs to be identified, as the milking parlour only allows for the observation of vulvar discharges, whereas observations at the feeding rack only allow for the recording of those that are nasal and ocular.

We have already dealt with cleanliness as an indicator related to comfort around resting, and the feasibility of this indicator has already been discussed. In dairy cattle, udder cleanliness has been used as an indicator to assess the risk of mastitis (Hughes, 2001; Reneau et al., 2005; Schreiner & Ruegg, 2003), and is therefore considered related to health in this species. Studies are needed to understand if this indicator is valid as an assessment of health status in goats.

In accordance with the findings of Whay, Main, Green & Webster (2003) in dairy cows in the UK (1.6% of cows out of 20% of sampled animals in 53 herds), Anzuino et al. (2010) report a low herd prevalence of “obviously dull/sick goats” (one or two goats per herd), but the authors do not provide an objective description of this indicator. Smith & Sherman (2009a) use the terms “dull and listless” to define the behaviour of goats with diseases such as chronic enterotoxaemia, pregnancy toxaemia, rumen impactation, and gastro-intestinal parasitism. A more accurate and precise description of this term may be useful to facilitate the identification of these goats in the herd and increase the reliability of the indicator. Farmers and technicians report that sick goats try to isolate themselves from the group, and they generally stand immobile, sometimes facing the wall or other parts of the housing structure. Although deriving from anecdotal information, this description is clear and could form the basis for an accurate definition of “obviously dull/sick goats”. The presence of this behavioural phenotype might be a promising indicator for pinpointing animals in poor health in a herd-living, gregarious species such as goats (Miranda-de la Lama & Mattiello, 2010). Therefore, the presence of isolated individuals may be interesting as an early-warning indicator of health problems, but it needs to be well defined and tested for validity and on-farm feasibility.

Laporte-Broux et al. (2011) considered different general activity behaviours (e.g., standing, feeding), aggressive interactions (e.g., threat, chase), and positive interactions (e.g., grooming) that may be influenced and altered by the presence of diseases, such as induced pregnancy toxaemia. The validity of some of these behaviours has been assessed, as goats with pregnancy toxaemia due to restricted diet increase feed searching and related active behaviours (e.g., walking, exploratory behaviours) but decrease positive interactions. The validity of behaviour for predicting diseases has already been confirmed in other ruminant species, such as cattle, where the time spent feeding and the frequency of visits to the feeder were significantly reduced by severe metritis (Huzzey, Veira, Weary & von

Keyserlingk, 2007) or ketosis (Goldhawk, Chapinal, Veira, Weary & von Keyserlingk, 2009). The feasibility of these indicators remains an issue if the time budget needs to be recorded.
In dairy animals, somatic cell count (SCC) is an important milk characteristic possibly related to animal welfare. Research in dairy goats suggests that this indicator is related to intramammary infections, as SCC increases in presence of clinical or subclinical mastitis (Koop, Nielen & van Werven, 2012; Paape et al., 2007), although the latter is asymptomatic and is not reported to affect goat welfare (Jimenez-Granado, Sanchez-Rodriguez, Arce & Rodriguez-Estevez, 2014). Somatic cell count is routinely collected herd data on bulk milk; therefore, it is available in most farms. However, it lacks sensitivity, because it can be affected by several factors, such as parity, lactation stage, and milk yield (Koop et al., 2013), that may increase SCC even in the absence of udder infection. To our knowledge, no appropriate statistical models are available to correct for all named factors in goat bulk milk so far.

Absence of Pain Induced by Management Procedures

This welfare criterion is related to the statement that "animals should not suffer pain induced by inappropriate management, handling, slaughter, or surgical procedures (e.g., castration, dehorning)" (Welfare Quality, 2009a).
One common management practice that may have a negative impact on the welfare of adult dairy goats is ear tagging. If tags are not correctly placed through the center of the ear, inflammation or ear tear (Anzuino et al., 2010; Smith & Sherman, 2009a) may occur, causing moderate pain. The prevalence of ear tear recorded by Anzuino et al. (2010) was 6.2% of goats (out of 1,520 goats). To our knowledge, these are the only data available on the damage associated with ear tagging in goats. In-depth studies on this potential indicator are needed, as it seems a promising inclusion for on-farm welfare protocol.
Routine trimming is a required management procedure, essential to maintain normal hoof structure (Smith & Sherman, 2009a). This procedure should be carried out at least twice a year due to the fact that on intensive farms, the limited movement of goats may alter the normal growth of the hoof, leading to claw overgrowth. Welfare problems associated with overgrown claws are widely explained in "absence of injuries" criterion. However, if trimming is not correctly performed, it may cause pain or expose inner tissue to trauma and infection (Nagy & Pugh, 2012).
Other procedures that may have a negative effect on goat welfare are improper disbudding (incomplete burn of the horn bud through the skin) and dehorning. Presence of horns or trace of horns in adult goats due to previous improper disbudding when they were kids may be a welfare issue. Scurs can press against the head or eye, causing lesions and pain (Smith & Sherman, 2009a) and they have been associated with adverse *sequelae*, such as sinusitis, brain abscesses or tetanus (Plummer & Schleining, 2013; Smith & Sherman, 2009a). Farmers also report that scurs may get caught in fences and pen partitions. Furthermore, horned or partially horned goats may

adopt agonistic behaviour, causing lesions and possible social stress to other goats (Waiblinger et al., 2011). For this reason, mixed groups of horned and hornless goats should be carefully managed by farmers, thus good practices are required (e.g., group stability, space allowance). Signs that adult animals have been dehorned or have traces of horns due to improper disbudding are clearly visible on-farm, especially when goats are at the feed trough; thus, improper disbudding may be a suitable indicator to be adopted on farm.

6.6 APPROPRIATE BEHAVIOUR PRINCIPLE

The "appropriate behaviour" principle is related to the opportunity that goats have to express social and species-specific behaviours and also to their relationships with humans. All indicators for the welfare criteria included in this principle are presented in Table 6.

Expression of Social Behaviours

In gregarious animals such as goats, that are usually housed in group pens, social behaviour is very important, and the possibility to "express a normal, non-harmful, social behaviour (e.g., allo-grooming)" (Welfare Quality, 2009a) should always be guaranteed.

Social behaviour can be altered by management procedures (e.g., regrouping; Miranda-de la Lama & Mattiello, 2010) or housing conditions (e.g., stocking density: Van et al., 2007; lack of resources suh as feed: Jørgensen et al., 2007; Laporte-Broux et al., 2011; Loretz et al., 2004; water: Ehrlenbruch, Pollen, et al., 2010; lying space: Andersen & Bøe, 2007; Loretz et al., 2004).

Agonistic behaviour has two main aims: resource achievement and dominance establishment (Shackleton & Shank, 1984), but the increase of these interactions is considered as a negative expression of social behaviour in cattle (Laister et al., 2009). Compared with other female ungulates, goats are reported to have a significantly higher rate of aggressive interactions (Fournier & Festa-Bianchet, 1995). The validity of agonistic behaviour as an indicator has been confirmed in many papers. In fact, the increase of agonistic behaviour is commonly related to decreased size of the resting area (Andersen & Bøe, 2007; Aschwanden et al., 2009b; Van et al., 2007), to regrouping (Andersen et al., 2008; Fernández, Alvarez & Zarco, 2007), or to insufficient resource availability (e.g., scarce number of feeding or water places: Ehrlenbruch et al., 2010b). Fernandez et al. (2007) even found a decrease in milk production in goats subjected to regrouping: it is known that regrouping may lead to an enhancement of aggressive interaction, thus increasing stress. Stress plays a fundamental role in reducing milk yield, as reported in cows (Shamay, Shapiro, Barash, Bruckental & Silanikove, 2000; Varner & Johnson, 1983), therefore, a concurrent validity is confirmed for agonistic interaction. Feasibility is questionable; observation usually requires a prolonged period of

Table 6. Animal-based indicators for assessing “appropriate behaviour”, excluding physiological measurements, in dairy goats

	Animal-based welfare indicator	Age class [1]	Sex [2]	Housing system [3]	Sample size	Validity [4]	Intraobserver reliability [4]	Interobserver reliability [4]	On-farm feasibility [4]	References
Expression of social behaviours	Agonistic interactions (with physical contact)	A/J	F	E	96	Y	M	-	N	Aschwanden et al., 2009a
		A	F	E	40	-	-	-	-	Mazurek et al., 2005
		A	F	E	35	-	-	-	-	Mazurek et al., 2007
		A	F	E	55	N	M	-	N	Nordmann et al., 2011
		K	F/M	E	30	Y	M	-	N	Van et al., 2007
		A	F	E	24	N	S	-	N	Ehrlenbruch et al., 2010a
		A	F	E	60	N	S	-	N	Laporte-Broux et al., 2011
		A	F	E	55	N	M	-	N	Nordmann et al., 2011
		A	F	E	15	Y	M	-	N	Fernandez et al., 2007
		A	F	E	32	Y	S	-	N	Andersen et al., 2008
		A	F	E	24	N	S	-	N	Andersen & Bøe, 2007
	Agonistic interactions (without physical contact)	A	F	E	40	-	-	-	-	Mazurek et al., 2005
		A/J	F	E	70	N	L	-	N	Aschwanden et al., 2009b
		A	F	E	55	N	M	-	N	Nordmann et al., 2011
		A/J	F	E	96	Y	M	-	N	Aschwanden et al., 2009a
		K	F/M	E	30	Y	M	-	N	Van et al., 2007
		A	F	E	24	N	S	-	N	Andersen and Bøe, 2007
		A	F	E	60	N	S	-	N	Laporte-Broux et al., 2011
		A	F	E	15	Y	M	-	N	Fernandez et al., 2007
		A	F	E	32	Y	S	-	N	Andersen et al., 2008
		A	F	E	24	N	S	-	N	Ehrlenbruch et al., 2010a
		A	F	E	35	-	-	-	-	Mazurek et al., 2007

Table 6. (continuation)

Displacements	A	F	E	40	-	-	-	-	Mazurek et al., 2005
	A	F	E	35	-	-	-	-	Mazurek et al., 2007
	A	F	E	32	Y	S	-	Y	Andersen et al., 2008
	A	F	E	24	Y	S	-	Y	Andersen & Bøe, 2007
	A	F	C	40	N	M	-	Y	Loretz et al., 2004
	A	F	E	24	N	S	-	Y	Ehrlenbruch et al., 2010a
	A/J	F	E	70	Y	L	-	Y	Aschwanden et al., 2009b
	A/J	F	E	30	Y	S	-	Y	Ehrlenbruch et al., 2010b
	K	F/M	E	30	Y	M	-	Y	Van et al., 2007
	A	F	E	15	Y	M	-	Y	Fernandez et al., 2007
	A	F	E	55	Y	M	-	Y	Nordmann et al., 2011
	A	F	E	40	-	-	-	-	Mazurek et al., 2005
	A	F	E	35	-	-	-	-	Mazurek et al., 2007
Queuing (at the feed barrier)	A	F	E	48	Y	S	-	N	Jørgensen et al., 2007
Queuing (at the drinker, duration)	A	F	E	30	-	S	-	N	Ehrlenbruch et al., 2010b
Vocalizations	J	F	E	48	Y	S	-	N	Siebert et al., 2011
	A	F	E	40	-	-	-	-	Mazurek et al., 2005
	A	F	E	40	-	-	-	-	Mazurek et al., 2007
Time budget	A	F	E	40	-	-	-	N	Mazurek et al., 2007
	J	F	E	48	Y	S	-	N	Siebert et al., 2011
	A	F	E	32	Y	S	-	N	Andersen et al., 2008
	A	F	E	24	N	S	-	N	Andersen & Bøe, 2007
Rearing (frequency)	J	F	E	48	Y	S	-	N	Siebert et al., 2011
Jumping (frequency)	J	F	E	48	N	S	-	N	Siebert et al., 2011

time to obtain reliable information (Nordmann et al., 2011). Among agonistic behaviours (e.g., threats, frontal clashes, chases, butts, displacements), the most feasible is displacement, as already described in this review for "absence of prolonged hunger and prolonged thirst" criteria (Table 3).

Queuing at the feed barrier has been already described for assessing the "absence of prolonged hunger" criterion, as well as the effect of queuing on nutritional deficiencies and frustration in dairy goats. However, queuing at the feed barrier may also be used to evaluate the quality of social interactions, as confirmed by Jørgensen et al. (2007), who found that low-status goats spent more time queuing than high-status goats, demonstrating that competitive environments are a welfare issue for subordinate animals rather than for dominants. The validity of queuing at the drinker as an indicator of social behaviour has not been tested (Ehrlenbruch et al., 2010b), but it is assumed that results similar to queuing at the feed barrier may occur when the ratio of nipples to goats is inadequate, as found for the "absence of prolonged thirst" criterion.

Vocalizations may be an interesting indicator of welfare in farm animals (e.g., pigs and cattle; Manteuffel, Puppe & Schön, 2004), and some authors report the use of vocalizations to assess the quality of social behaviour in goats (Mazurek, Marie & Desor, 2005; Mazurek et al., 2007; Siebert, Langbein, Schön, Tuchscherer & Puppe, 2011). However, specific studies need to be conducted to gather more information about the motivation of goats to vocalize. For example, increased vocal responses in situations of social isolation can be interpreted either as an adaptive and active attempt to communicate with companions or as a sign of distress and fear (Siebert et al., 2011). Moreover, the mere verbal description of sounds is incomplete and misleading as it lacks loudness, duration, and frequency traits (Manteuffel et al., 2004); therefore, specific instruments for recording (e.g., digital signal processing; Siebert et al., 2011) and analysing sounds (e.g., digital sonograms bio-acoustical analyses; Manteuffel et al., 2004) are required and, thus, the feasibility of this indicator is compromised.

Expression of Other Behaviours

Other behaviours include the possibility "to express species-specific natural behaviours such as foraging" (Welfare Quality, 2009a). If animals cannot meet their natural needs and cannot perform their entire behavioural repertoire, they may experience a negative affective state. However, no feasible indicator of absence of natural behaviours has been identified so far. Therefore, in this section we will only consider as indicators of negative affective state the presence of behaviours that are not species-specific under natural conditions or that are performed with unnatural frequency. Excessive scratching or rubbing (pruritus) and abnormal oral behaviour can be considered as promising indicators of a poor welfare condition (Mason & Rushen, 2008). Several studies are available to support the observation of these behaviours as welfare indicators in many species (e.g., pigs: Brunberg, Wallenbeck & Keeling, 2011; cattle: Mattiello et al., 2002; Kamboj, Vishwakarma & Singh, 2007). In goats, Anzuino et al. (2010) reported that 91.7% out of 24 UK farms had goats affected by pruritus, and the validity of this indicator seems to be supported by the positive correlation with the prevalence of body skin lesions.

Behaviours such as oral manipulation of inert objects are potential welfare issues requiring further investigation. Lack of roughage and of concentrate feed was associated with abnormal oral behaviour in goat kids (Mattiello, Villa & Cioccarelli, 2008). According to Anzuino et al. (2010), abnormal oral behaviours in goats are mainly directed at the bars (found in 91.7% out of 24 farms) or walls of pen structures (83.3% of farms). Self-suckling and inter-suckling are also abnormal oral behaviours common among certain goat breeds (e.g., Murciana) and farms (Griffioen, 2012). As reported by Martínez-de la Puente et al. (2011) in the Majorera breed, self-suckling has been shown to be reduced when extra fibber is added to the ordinary diet (composed of corn, soy, dehydrated lucerne, dehydrated beetroot, lucerne hay, vitamin-mineral supplement), and distributed twice a day. The authors of this study suggest that by having *ad libitum* access to wheat straw, goats ruminate more and thus satisfy their oral behavioural needs. The main issue related to the recording of these behaviours is the fact that direct observation requires not only a long observation period, but also an accurate choice of observation time (i.e., just before feeding time). In addition, abnormal behaviours can be difficult to record in large groups (Anzuino et al., 2010), especially in the absence of an adequate sampling strategy.

Good Human-Animal Relationship

This criterion is based on the assumption that "Animals should be handled well in all situations, i.e. handlers should promote good human-animal relationships" (Welfare Quality, 2009a).

Many published studies provide proof of the effect of stockmanship on animal production (e.g., Hemsworth, 2003; Hemsworth & Barnett, 1991; Lensink, Raussi, Boivin, Pyykkönen & Veissier, 2001) and welfare (e.g., de Passillé & Rushen, 2005; Rousing & Waiblinger, 2004). Nevertheless, the effect of the quality of human–animal relationships on productive traits and welfare has rarely been investigated in goats. The human–animal relationship tests enable us to gain information about the level of fear as a consequence of the quality and frequency of the previous human–animal interactions (Waiblinger et al., 2006). The promising tests identified so far belong to two main categories: reaction to a stationary person and reaction to a moving person. The latency to approach a stationary person test, already used in cattle by several authors (Breuer, Hemsworth, Barnett, Matthews & Coleman, 2000; Hemsworth, Coleman, Barnett & Borg, 2000; Jago, Krohn & Matthews, 1999; Lensink et al., 2000), was applied by Jackson & Hackett (2007) to evaluate the positive effect of gentle handling treatment in goats. The research showed that animals approached a person more quickly if they had already experienced positive contact with humans (Jackson & Hackett, 2007). The avoidance distance test to a moving person developed for cattle (Windschnurer, Schmied, Boivin & Waiblinger, 2009) was validated in goats by Mattiello et al. (2010), who confirmed its validity to detect different avoidance reactions of goats depending on farm size and, consequently, on management practices. Behavioural tests for assessing human-animal relationship seem to be valid, feasible, and reliable in several species (sheep: Napolitano, De Rosa, Girolami, Scavone & Braghieri, 2011; beef heifers: Mazurek et al., 2011; buffalos: De Rosa, Napolitano, Grasso, Pacelli & Bordi, 2005; dairy cows: Rousing & Waiblinger, 2004). However, they require time and training, and might induce stress in the animals. For example, Muri et al., (2013) reported strong avoidance behaviour and goats flocking around when the observer tried to perform the avoidance distance test. The development of less intrusive and time consuming, but still valid, indicators would be useful. Farmers have suggested other possible indicators, such as alert calls, but no scientific evidence is available to support their use in assessing the quality of human-animal relationship in goats.

Positive Emotional State

This last criterion is very difficult to assess, and it is based on the concept that "Negative emotions such as fear, distress, frustration or apathy should be avoided whereas positive emotions such as security or contentment should be promoted" (Welfare Quality, 2009a).

Most welfare indicators focus on negative aspects, whereas positive emotional state is not taken into consideration very often. Play behaviour is one of the few positive indicators in the existing literature on farm animals (calves: Jensen & Kyhn, 2000; pigs: Reimert, Bolhuis, Kemp & Rodenburg, 2013); however, this is seldom expressed by adult ruminants and it is too time consuming to record (Napolitano, Knierim, Grasso & De Rosa, 2009). New indicators need to be developed to detect positive emotional states. Currently, one promising approach is the qualitative behaviour assessment, which consists of a whole-animal approach, in which the observer can integrate perceived behavioural details and signals to judge an animal's behavioural expression, using qualitative descriptors (e.g., content, relaxed, anxious) that reflect the animal's emotional state (Wemelsfelder, 2007). Studies on many species, such as horses (Minero, Tosi, Canali & Wemelsfelder, 2009), pigs (Rutherford, Donald, Lawrence & Wemelsfelder, 2012; Wemelsfelder, Millard, De Rosa & Napolitano, 2009), and sheep (Phythian, Michalopoulou, Duncan & Wemelsfelder, 2013), have shown that data generated from such observations is reliable and repeatable, and it correlates with animal behaviour observed by traditional quantitative methods. So far, no information has been made available on the use of qualitative behaviour assessment in goats. Research on this topic is worth pursuing.

6.7 SUMMARY AND CONCLUSIONS

This review aimed to highlight goat welfare indicators that may be suitable for on-farm welfare assessment. Validity, reliability, and feasibility were the major attributes taken into account. According to these attributes, few indicators can be considered promising and all require further adjustment or research. Furthermore, not all the principles and criteria addressed in this review are covered by the indicators available to date.

As can be expected, indicators related to injuries and diseases are the most studied, whereas those related to other principles need further investigation. Health indicators were attributed more attention due to their economic relevance. Similarly, many indicators have been found to evaluate "absence of prolonged hunger", whereas few animal-based indicators have been developed for on-farm assessment of "absence of prolonged thirst", possibly because they are easily replaced by resource-based parameters (e.g., number of drinkers/animal and water flow). The most promising indicators are summarized in Table 7. In line with the present state of the art, no indicator can be included in an on-farm welfare assessment protocol as it stands. Each one of the potential indicators needs to be adjusted or further refined in terms of validity, reliability, or feasibility. For example, interobserver reliability has seldom been assessed and further studies are needed: concordance among observers is essential when setting up a reliable data collection. Easy-to-perform observations or simplified point scales are recommended for obtaining high interobserver agreement but, in any case, rigorous training is required. In contrast, short-term intraobserver reliability has frequently been assessed, and some studies also consider mid- and long-term reliability. This issue is very important, as the consistency of indicators over time is crucial in welfare assessment protocols that need to be used at any time.

From a methodological point of view, most of the studies carried out so far have been performed under experimental conditions, so the feasibility in commercial farms of the indicators adopted in these studies needs to be carefully evaluated. Our list of promising indicators took this issue into account, highlighting the need for further studies for on-farm feasibility.

Another issue related to feasibility is that the recording of behaviour is generally time consuming. Most of the behavioural indicators reviewed in this paper were collected by video-recording, not by direct observation. This makes them unsuitable for inclusion in a practical on-farm welfare assessment protocol, unless we find alternative ways to collect the data. This is why we have continually emphasized the idea of applying an effective sampling strategy to record data on-farm.

It is important to underline that some indicators, such as BCS, hair coat condition, and queuing, can be applied to different criteria. These indicators warrant particular attention, as they could be very useful in order to save time, by providing information about several welfare aspects.

Existing welfare assessment protocols usually highlight negative aspects, and most of the reviewed indicators are actually focused on welfare problems, such as presence of lesions, lameness, diseases, abnormal behaviour, or poor body conditions. However, positive indicators deserve further attention and their inclusion should be fostered, as they may play a key role in the communication of animal welfare to the stakeholders.

The aim of this review was to act as a starting point in the identification of animal-based indicators in dairy goats. Further studies are needed in order to bridge the gaps highlighted by the present review, to confirm the already existing results, and to develop a complete and effective on-farm welfare assessment protocol for dairy goats. Such a protocol may then be used as a basis for developing specific protocols for other categories, such as kids and bucks, and for different production systems (e.g., meat and fibber) and extensive husbandry that may have specific welfare problems not addressed in this review.

Table 7. Promising indicators identified as suitable for inclusion in on-farm welfare assessment protocol, classified according to Welfare Quality principles and criteria

Principle	Criteria	Promising indicators
Good feeding	1-Absence of prolonged hunger	BCS Queuing at the feed barrier Hair coat condition
	2-Absence of prolonged thirst	Queuing at the drinker Displacements at the drinker
Good housing	3-Comfort around resting	Resting in contact with the wall
	4-Thermal comfort	Shivering score Panting score
	5-Ease of movement	Kneeling (at the trough)
Good health	6-Absence of injuries	Lameness Claw overgrowth Lesions and swellings Teats and udder abnormalities and conformation traits
	7-Absence of disease	BCS Abscesses Hair coat condition Kneeling (in the pen) Vulvar, ocular and nasal discharges
	8-Absence of pain induced by management procedures	Ear tear Improper disbudding and dehorning
Appropriate behaviour	9-Expression of social behaviours	Queuing at the feed barrier and at the drinker Displacements at the feed barrier and at the drinker
	10-Expression of other behaviours	-
	11-Good human-animal relationship	Avoidance distance test Latency to first contact
	12-Positive emotional state	QBA

CHAPTER 7

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

Body condition score has been widely used as a welfare indicator. However its inclusion in on-farm welfare protocols presents some constraints. Chapter 7 addresses the difficulty of goat restraint by presenting a visual body condition scoring system that was developed by identifying representative images of three body condition categories: very thin, normal and very fat goats. The scoring system, assessed for validity and reliability, is a practical tool for a quick first step assessment of fat reserves changes that have an impact on animal's welfare, health and production.

This chapter has been accepted for publication as: Vieira, A., Brandão, S., Monteiro, A., Ajuda, I., & Stilwell, G. (2015). Development and validation of a visual body condition scoring system for dairy goats with picture-based training. *Journal of Dairy Science*, 98(9), 6597-6608. http://dx.doi.org/10.3168/jds.2015-9428

A. Vieira contribution started conducting the field work for the data acquisition, and the digital image analysis. The author also reviewed the available literature that led to the methodological options undertaken during development of the visual body condition scoring system, and leaded the inter-observer reliability study analysing its results. Finally, the author actively discussed all the results and drafted the manuscript.

Body condition scoring (BCS) is the most widespread 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 restraining for body palpation.

In this study, AWIN (Animal Welfare Indicators project) 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 three categories: very thin, normal and very fat, and was built from datasets with photographs of animals scored by a commonly used six-point scoring system that requires palpation in two anatomical regions. The AWIN scoring system development required three steps: i) identification and validation of a body region of interest; ii) sketching the region from photographs; iii) creation of training material. The scoring system 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 dataset from 171 goats. To account for variability in animal sizes and camera positions, 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 two field reliability studies. In the first test, no training was given to four observers, while in the second, training using the threshold images was delivered to the same observers. In the first experiment, interobserver agreement was substantial, showing the visual scoring system is clear and unambiguous. Moreover, after training, results improved, reaching almost perfect agreement in the very fat category. Our results demonstrate that the visual body condition scoring system is not only a practical tool for BCS in dairy goats, but also shows potential to become fully automated, which would enhance its use in welfare assessment schemes and farm management.

7.1 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 substantial impact on the animal's health, welfare and production (Caroprese et al., 2009; Roche et al., 2009). There are different methods to monitor changes in body fat reserves but the body condition scoring (BCS) is the most widespread method across species, being considered simple and repeatable (Ferguson, Azzaro & Licitra, 2006).

For cattle, sheep and goats, the scoring systems most commonly used to assess BCS are numerical rating scales (NRS) with five (Edmonson, Lean, Weaver, Farver & Webster, 1989; Ferguson, Galligan & Thomsen, 1994), six (Harwood, 2006; Hervieu & Morand-Fehr, 1999; Lowman, Scott & Somerville, 1976; Russel et al., 1969) or even eight-point scales (Wildman et al., 1982). 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 considerable training and experience are needed from observers. Moreover, all these methods are subjective as they rely on the observer's assessment, generally based on visual appraisal and/or palpation of anatomic locations that may demand animal restraining, which is time consuming and not easily achieved in large farms (Edmonson et al., 1989; Halachmi, Polak, Roberts & Klopcic, 2008).

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

Body condition scoring is considered a valid, reliable and feasible welfare indicator with high potential to be included in on-farm welfare protocols (Botreau et al., 2009; Winckler et al., 2003). 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 & Winckler, 2009).

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

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, 2009b), by limiting the scoring systems to two or three levels: e.g. very thin, normal and very fat. Because our main objective was to assure feasibility when using body condition scoring 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 times, 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 (Bewley et al., 2008; Ferguson et al., 2006; Negretti, Bianconi, Bartocci, Terramoccia & Verna, 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.

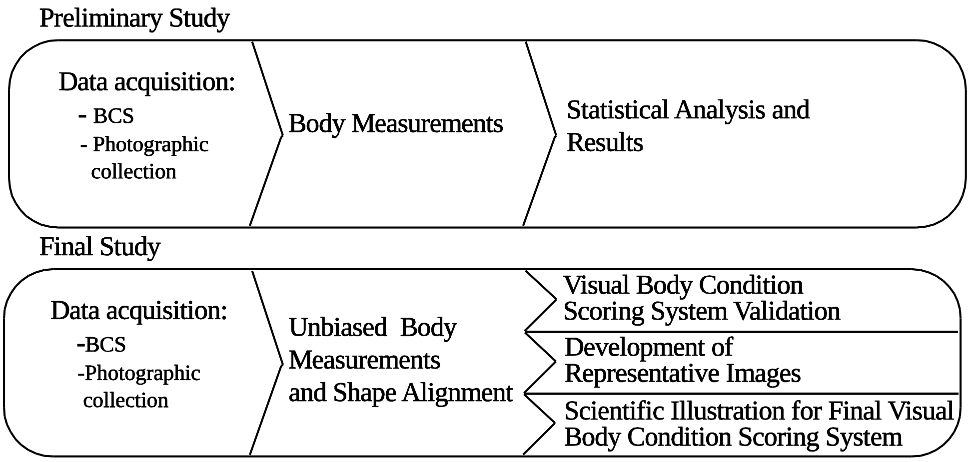
Thus, visual scoring systems should convey the appearance of a representative animal from each category. The definition of a representative for a given category, for the purpose of classifying subsequent sets of new images, is a common research topic in computer vision. Approaches either focus on the statistical distribution of features (Tu & Yuille, 2004), e.g., identifying mean/median images, or on the identification of threshold features, e.g. images that are on the transition between two different categories (Burges, 1998). We believe assessors, after getting 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, there are three objectives for this study. 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. Finally, to assess the inter-observer reliability under field conditions.

7.2 MATERIALS AND METHODS

The general flow for the development of a visual body condition scoring system is presented in Figure 2. The study was organized in two 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 analysed 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).

Figure 2. Visual body condition scoring system development outline



Preliminary study for scoring system development

a) Data acquisition

We used 32 goats [20 Saanen, five Alpine and seven crossbreed (Saanen x Alpine)] from an intensive dairy goat farm to select the goat's body region that would give us the best representation of the BCS levels. The goats were housed in pens with straw litter, with occasional access to exercise exterior areas. Diet was composed of a total mixed ration distributed twice per day, and goats were milked twice a day in an automatic milking parlour (average of 2.0 litters per goat per day). Kids were separated from their mothers after birth.

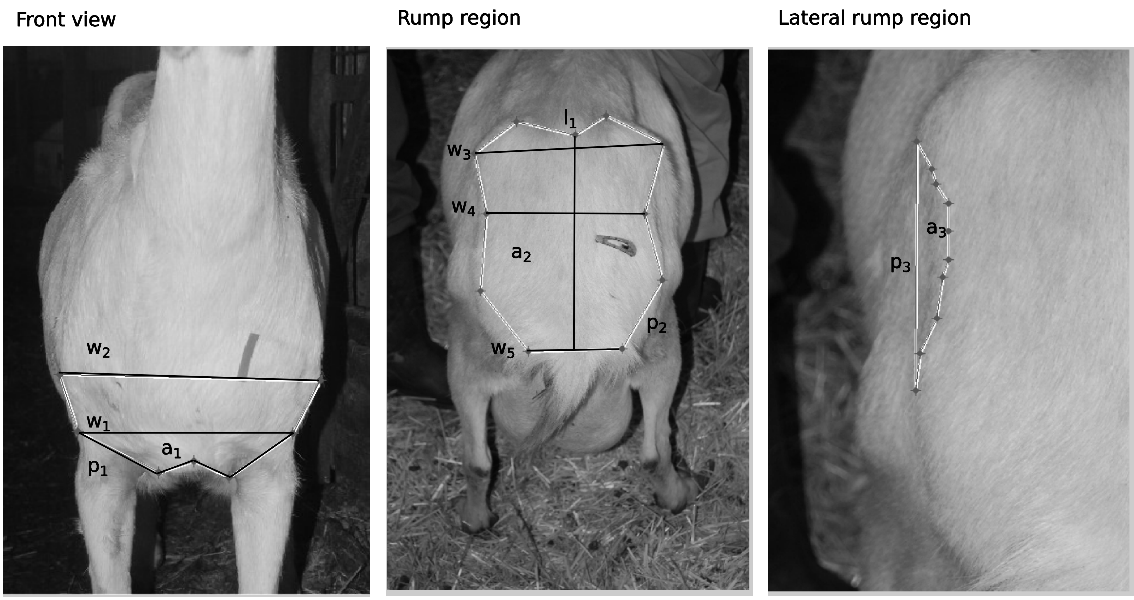
Two experienced assessors independently scored the 32 goats using the BCS method developed by Hervieu and Morand-Fehr (1999). If there was any discrepancy, the final score was decided by consensus. Only adult goats were scored, independently of milk yield and days in milk. The BCS median was 3 with a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

Using a camera (Nikon D60) with automatic exposure adjustment and focus, we took photographs from 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 & Morand-Fehr, 1999). Environmental conditions (e.g. light and background) and distance 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.

b) Body measurements

For image analysis we performed a total of four and eight measurements in millimetres of the front and rump regions, respectively (Figure 3). 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/).

Figure 3. Measurements and their relation with body features. Figure on the left (front view) shows the 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). Figure on the center (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), rump region perimeter (p2). Figure 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 that is limited by the line that unites the hipbone and the ipsilateral thurl; the lateral virtual perimeter (p3) corresponds to the line surrounding a3.



c) 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 the body measurements as the independent variables. Front and rump view images were analysed separately.

The rump region view stepwise regression had a higher adjusted R^2 representing the region whose variables better explain the relationship between the BCS and the measurements performed (Table 8). The lateral virtual area (a3) showed the highest significance (Table 8).

Table 8. Summary of the stepwise linear regression model applied to the front and rump view measurements (in millimetres).

Body measurements[1]	Coefficients				Adjusted R^2	Residual Standard error (degrees of freedom)
	Estimate	Standard error	T value	Significance		
Front view					0.39	0.81 (29)
Intercept	4.136	1.397	2.961	0.00606 **		
a1	7.387 $x10^{-7}$	2.313 $x10^{-7}$	3.194	0.00337 **		
p1	-1.454 $x10^{-5}$	7.812 $x10^{-6}$	-1.861	0.07297 .		
Rump view					0.88	0.36 (28)
(Intercept)	5.485	2.204	2.489	0.0190 *		
a2	6.554 $x10^{-5}$	2.810 $x10^{-5}$	2.333	0.0271 *		
p2	-6.640 $x10^{-3}$	4.106 $x10^{-3}$	-1.617	0.1171		
a3	6.489 $x10^{-4}$	1.018 $x10^{-4}$	6.373	0.0000 ***		

[1] The BCS value assigned by the Hervieu and Morand-Fehr (1999) method is the dependent variable and the body measurements are the independent variables. The front and rump view are compared by means of their adjusted R2. The view that has a greater adjusted R2 represents the region whose variables better explain the relationship between the BCS and the measurements performed.
[2] a1 = 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

Final study for scoring system development

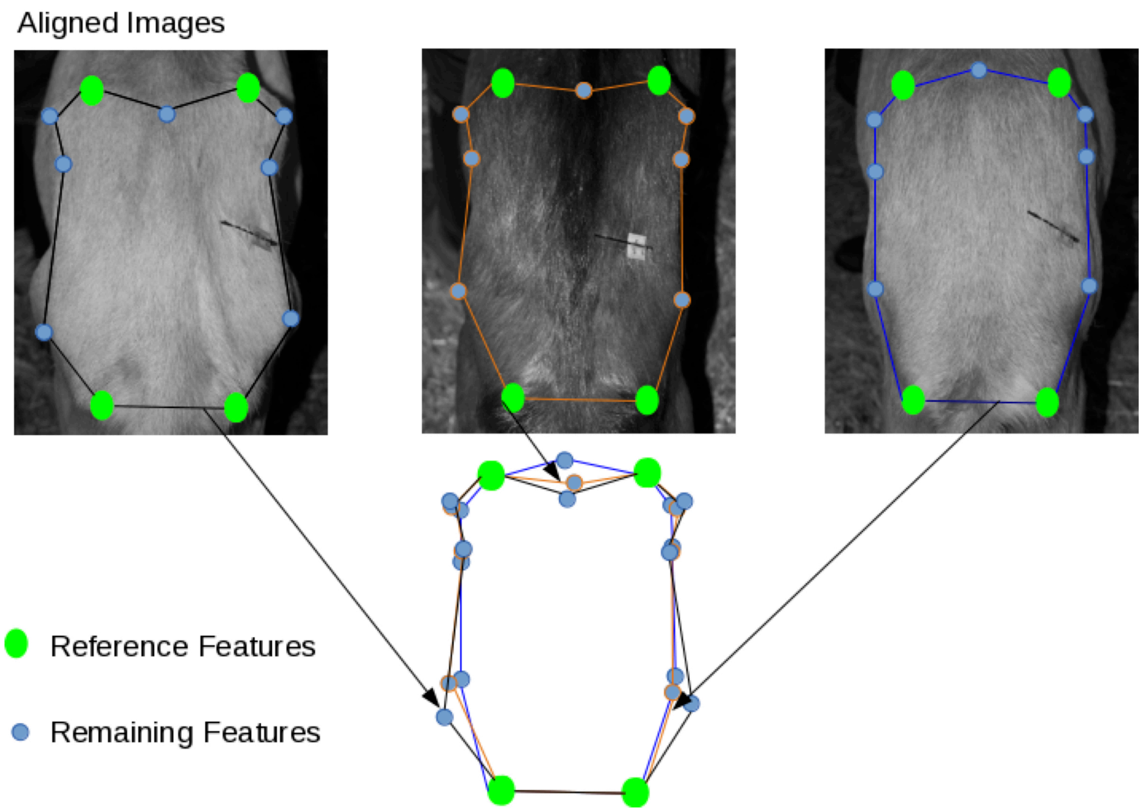
a) Data acquisition

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 four different dairy farms and included the following breeds: Saanen (73), Alpine (24), Murciana (41) and crossbreed Saanen x Alpine (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 took photographs and used the rump region measurements. We followed the same image collection procedures.

b) 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 fixes a posture and a scale for the animal, so that the hip and pin bones (the four reference features) lay in the same position across photographs, as illustrated in Figure 4. The corrected body measurements were computed using the aligned features.

Figure 4. Scheme showing the shape alignment process of different body size and camera pose using the four reference features in the animal rump



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 four reference features, hip and pin bones, laid 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 is in that plane, which is a valid assumption for an observer far from the rump, we can apply the same transformation to the whole rump, including the remaining features. To determine the transformation between image and template, we first extract the rump feature positions, i.e., their pixel position in terms of its column u and line v in the image. To allow a technician to recover each feature pixel coordinates, we developed a Matlab application for manual feature labelling.

Following the recommendations of Azzaro et al. (2011) two technicians were used in the labelling phase, to ensure the final dataset was not biased by subjectivity.

Labelling consisted of two parts. In the first, we annotated 11 points in the rump region, as illustrated in Figure 3 (rump region). In the second part, we annotated the lateral virtual area (a3) and the lateral virtual perimeter (p3) as illustrated in Figure 3 (lateral rump region).

From the labelling process, each feature f in each image becomes 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 four reference features in the original image, transforms them to the coordinates of the four 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, i.e., an image where the reference features are aligned with the template, we used, T_p to align images using the Matlab *function imtransform*.

c) Visual body condition scoring system validation

We evaluated the relationship between the 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.

While the independent variables may not present a normal distribution, which will be the case with the order two and three monomials, the linear regression is still the estimator with the minimum variance.

d) Development of representative images

We grouped all the animals into BCS categories: 0 - Very thin (BCS≤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 falls 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 body condition score $\hat{b}$, i.e., for each animal on a given category x, we computed the estimated body condition score 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, the animal a_c at the center of a category, and therefore its best representative, is the animal with an estimated body score closest to $\hat{b}_c$. To avoid a sampling bias, we further selected 10 more animals with a similar visual distribution features', i.e., 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 one. We did so by looking at the distribution of each pixel value over the whole set of images. For example, consider the pixel in the sketch image with coordinates (u, v). To estimate the colour of each pixel, $c_{sketch}(u,)$, we first retrieved the colour of the pixel with coordinates $(u,)$ over all the 10 images, $s(u,) = \{c_1(u,v), c_2(u,v), \dots, c_{10}(u,v)\}$, and then defined $s(u,v)$. To minimize the impact of different breeds' colour variation, we first converted all the colours to a grey scale, using Matlab *function rgb2gray*, and ensured that the distribution of colours was similar, using Matlab *function histeq*.

e) 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 applying the idea of threshold images. We defined the threshold as the interval of overlapping estimated body measurements, between two consecutive categories. The animals that were close to the center of the interval were used as threshold representatives. We determined if an animal was in the threshold interval if its estimated body condition score was higher/lower than those of a consecutive category. For example, consider an animal a_i in the very thin category with an estimated body condition score $\hat{b}_i$. Consider also that the minimum estimated body condition score 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.

Inter-observer reliability study

The goats used in the inter-observer reliability were kept in the same husbandry and management conditions as described above. Each individual goat was observed simultaneously by four observers. Observers 1 and 3 were veterinarians with 20 and five years of experience in clinical and research work with small ruminant, using BCS routinely in their studies in pregnancy toxaemia. Observers 2 and 4 were MSc veterinary students involved in research with dairy goats for their master thesis. Two observers (observers 1 and 3) had previous experience in BCS with the Hervieu and Morand-Fehr (1999) method, but the other two (observers 2 and 4) did not. The standing goats were assessed from behind, making sure there was a good visualization of the rump region. The observers were positioned caudally and slightly dorsally to each animal. The assessment of each goat took approximately one minute. The observers were not allowed to discuss their scorings during the observations.

Considering an expected inter-observer 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 a 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; two observers were considered for the calculation of the sample size needed for the reliability study. This study included two experiments.

a) Experiment 1

The day of field testing took place in December 2013. Inter-observer 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 (26), Alpine (7) and crossbreed (Saanen x Alpine) (12).

b) Experiment 2

The experiment was performed during one day of field testing in February 2014. This second part was identical to the first one with the exception that the observers had a training period 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 the 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). 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 (25), Alpine (9) and crossbreed (Saanen x Alpine) (15).

Agreement (proportion of exact agreement and disagreement, by one and two 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 is calculated by dividing the number of agreements by the total number of agreements and disagreements (Martin & Bateson, 2007). A value around 75% is generally considered a good agreement (Burn & Weir, 2011).

Reliability measures the ability of observers to differentiate among the different levels of a scoring system (Kottner et al., 2011; Streiner & Norman, 2008). 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), being 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 are 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 & Norman, 2008).

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

7.3 RESULTS AND DISCUSSION

Preliminary study for scoring system development

In our preliminary study the "rump region view" stepwise regression had a higher R^2 representing the region whose variables better explained the relationship between the BCS and the measurements performed. These results are in line with studies in dairy cows (Ferguson et al., 2006) and buffalos (Negretti et al., 2008) that 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.

So it was demonstrated that there is a strong relationship between dairy goats' rump body measurements and body condition score. 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, study in which we formally validate the relationship.

Final study for scoring system development

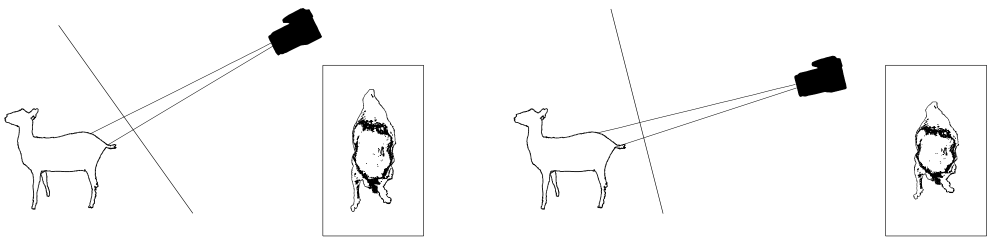
The main objective of including BCS in on-farm welfare assessment schemes is to identify animals at risk of nutritional deficiency or excess, corresponding to very thin animals and very fat animals, that are predisposed to metabolic diseases and poor welfare; hence the method does not need to be very detailed (Winckler et al., 2003). In a study in dairy cows, although the mean differed between observers, assessors were consistent in their identification of the 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 & Sherman, 2009b). 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 & Butler (2008) observed a correlation between BCS and mortality, finding that in cold weather conditions mortality increased rapidly when goats BCS was < 2.0, with no mortality of animal with BCS ≥ 2.5. For these reasons, we established the cut-off ≤ 2 for the very thin goats. On the opposite side of the scoring system we considered very fat goats as the ones scored > 3.5, as BCS above this value corresponded to a higher risk for the development of pregnancy toxaemia, and other health problems (Smith & Sherman, 2009b).

In the final study, the animals' BCS variability increased considerably so as to represent the reality in many dairy farms. We also increased the number of images and included animals in the BCS extremes, which is not very common in the majority of farm animal body condition research (Bewley et al., 2008). We thus had animals from different breeds and with a large spectrum of sizes and body condition scores. To avoid biasing the result due to changes in body size, that may not be a consequence of the body condition score, 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), as we had distortion from changes in the camera orientation with respect to rump plane. As depicted in Figure 5, the camera orientation introduces different scale factors over the horizontal and the vertical axis 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, contrary 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.

Figure 5. Scheme representing the distortion introduced from changes in the camera orientation with respect to rump plane



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 features (Lowe, 2004).
Using the corrected features in the extended dataset we inferred that the unbiased body measurements and the BCS are related by a third order polynomial, demonstrated by the highest adjusted R^2 (Table 9).

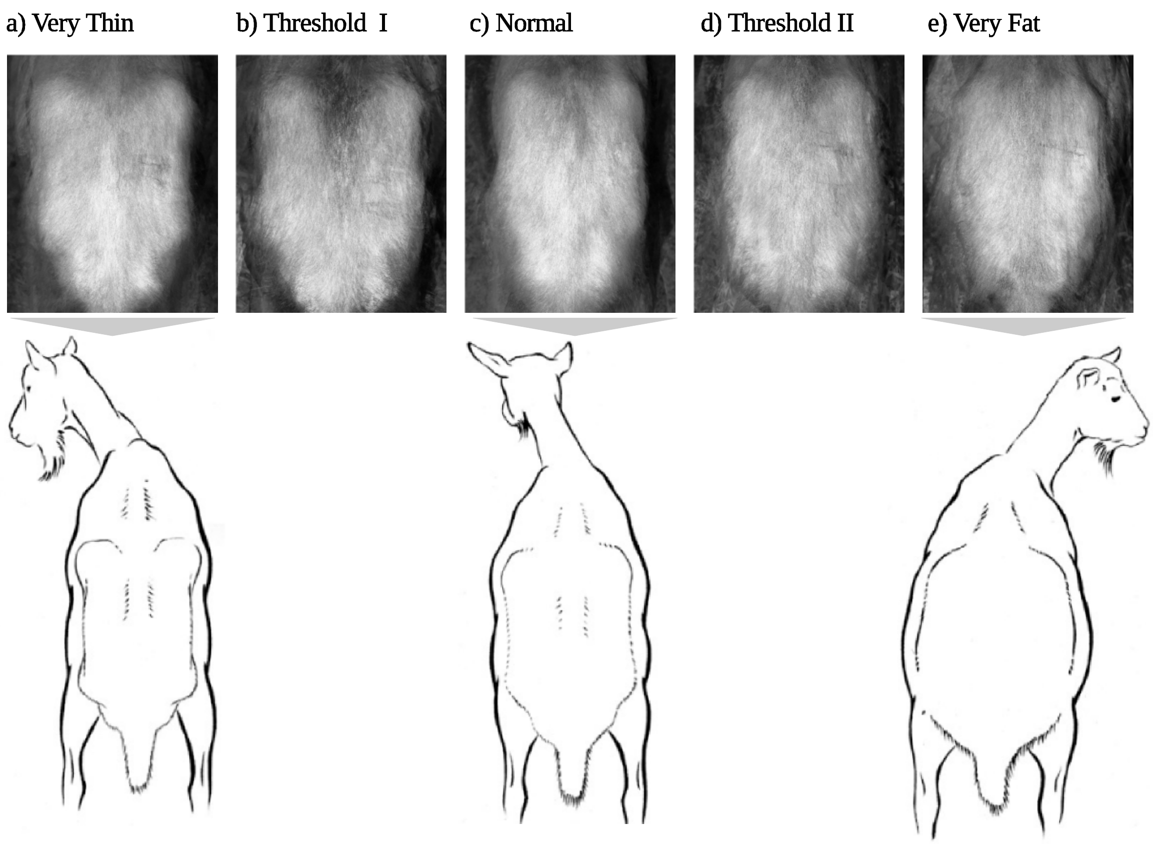
Table 9. Summary of the adjusted R^2 and residual standard error (degrees of freedom) for the three analysed models where we evaluated the relationship between the BCS and the unbiased body measurements using a stepwise linear regression. In the first model we explore a linear relation (First order polynomial), we then explore relations above the linear, computing the second (Second order polynomial), and third order monomial (Third order polynomial) of the measurements. Higher adjusted R^2 is used as the criterion for determining the best model that explains the relationship between the BCS and the unbiased body measurements.

Model	Adjusted R^2	Residual Standard error (degrees of freedom)
First order polynomial	0.70	0.46 (164)
Second order polynomial	0.73	0.44 (160)
Third order polynomial	0.74	0.43 (160)

We could now use the third order relation to parameterize our animals using the estimated body score $\hat{b}$. The parameterization defines an order over the whole set of animals, as if we had mapped the set of all the animals into a line. This line includes information on both the body condition score and the 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 to be the one closest to the $\hat{b}$ distribution mean.
When selecting a group of images close to the representative, and as the rump shape presents a large variability, we need to ensure that we choose a set of images where the rump of the animals presents an adequate overlap. We thus chose a set of animals whose appearance was similar to that of the representative. Having showed the relationship between the rump features and the anatomical characteristics we wished to highlight, we could use the rump features to define a similarity between animals instead of a more common texture or shape descriptors (Belongie, Malik & Puzicha, 2002). We finally got to the representative sketches and the scientific illustration for each category (Figure 6). The estimated body condition scores for the illustrations correspond to 1.75, 2.48 and 3.70 for the very thin, normal and

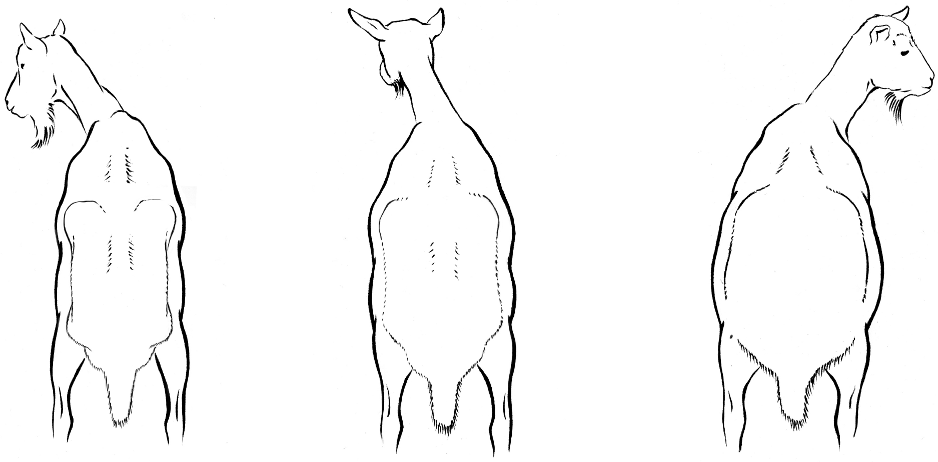
very fat respectively. Also, we obtained an estimated body condition score of 2.10 and 3.14 for the respective threshold image sketches.

Figure 6. Representative images for the BCS categories (very thin, normal and very fat) and the respective scientific illustration. 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.



We found that the scientific illustrations increased clarity, reproducibility and repeatability. For the final scoring system (Figure 7) we also included some verbal descriptors to each level of assessment. This is frequently used with visual descriptors (Bana e Costa & Beinat, 2005) to aid assessors making their decision.

Figure 7. Final visual body condition scoring

	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.

Training development for observers and Inter-observer reliability study

A benefit for 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 6) greatly benefited the observers training, as can be seen by the results of the inter-observer reliability study (Table 10).

Table 10. Inter-observer agreement in the evaluation of the visual body condition scoring system before and after training of the observers, respectively.

	Before training (n = 45)				After training (n = 49)			
Observers	Exact agreement (% / n)	One point difference* (% / n)	$kappa$ (95% CI)	$kappa_{weighted}$ (95% CI)	Exact agreement (% / n)	One point difference (% / n)	$kappa$ (95% CI)	$kappa_{weighted}$ (95% CI)
Observers 1-3[1]	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 2-4[2]	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)

CI, Confidence interval
[1] Experienced observers
[2] Non-experienced observers

Before training, the percentage of exact agreement was lower for the two experienced observers (67%), probably due to biases of previous experience with other BCS methods. In contrast, the non-experienced observers had a very good level of agreement (82%) (Table 10). After training, exact agreement increased in both experienced and non--experienced observers, reaching very similar levels (86 and 82%, respectively) (Table 10). Moreover, before training, the pair of experienced observers had an 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% the agreement on, respectively, the normal and very fat goats (data not shown). The non--experienced observers reached an 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, had already very high percentages of exact agreement on the very thin goats, for this reason this category seems to be self-evident. However, the training improved the agreement in both pairs of observers when scoring the very fat and the normal goats, showing the importance of the training especially on the threshold images, to clearly identify and distinguish the normal from the 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 higher increase of the k when compared with the increase of the k_w, means that more scores became absolutely equal after the training.

The non-experienced observers did not show any improvement in their k and k_w coefficients after training. The training of the non-experienced observers seemed to have only a marginal effect on their k and k_w values, however, their values were already considered almost perfect before the training, and so there was not much space for improvement. This can be an indication that the visual scoring system was by itself clear and unambiguous, and that the non-experienced observers were very open to new information and interested in the new method, while the experienced observers had to be more “convinced” on 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 inter-observer reliability. In the context of welfare assessment where the training of different observers with different levels of experience is the most likely situation, the development of training material that can be made available online is valuable.

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

7.4 CONCLUSIONS

This study provides a practical tool for BCS in dairy goats by reducing subjectivity. The AWIN visual body condition scoring system for dairy goats was assessed for validity and repeatability, having been demonstrated that it is a simple, repeatable and expeditious compared to conventional methods. It is also less stressful to the goats, and can be used even by non-experienced observers with substantial levels of inter-observer reliability. It should be said that the AWIN scoring system does not aim to replace the current more detailed scores (especially considering the normal category) that are valuable for effective production monitoring, but aims 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. Additionally, with this study we propose a systematic method for the creation of visual scoring systems with the potential to be applied in other contexts, and introduce the concept of threshold images to be used in the development of training programs.

CHAPTER 8

Design and test of a web-survey for collecting observer's ratings on lameness scoring data

Chapter 8 describes how we designed, implemented and ran a web-survey with the objective of collecting observer's ratings on goat lameness scoring data. Being an innovative method for data collection in animal lameness scoring, we give information on the web-survey outreach, and detail some insights from the work developed.

This chapter has been submitted for publication as: Vieira, A., Oliveira, M.D., Nunes, T. & Stilwell, G. *Design and test of a web-survey for collecting observer's ratings on lameness scoring data*. Manuscript submitted for publication.

A. Vieira conceived the idea for paper, designed the web-survey and coordinated its implementation with the informatics company. The author compared the web-survey with other published methods of data collection and critically wrote this technical note.

A considerable amount of applied research studies make use of scoring systems that are based on observer's ratings collected during farm visits. New methods for collecting this type of data have recently emerged, for instance using videos or photographs that are presented and assessed during workshops or meetings. In this paper we present and discuss the use of web-surveys as an alternative method for data collection, using goat lameness scoring as an example. This paper describes how we designed, implemented and ran a web-survey, and were able to outreach a very high and diverse number of respondents. We explain how the web-survey based on an innovative technology-based design was implemented, and discuss how the survey results can enhance current investigation and open new research lines. We believe that this study contributes to the understanding of the role of web-surveys in animal clinical and welfare research, and presents important information for researchers and practitioners that need to collect and analyze observer's ratings.

8.1 INTRODUCTION

Over the years lameness scoring research has been based on live assessment by an observer (Brenninkmeyer et al., 2007; Channon, Walker, Pfau, Sheldon & Wilson, 2009; Hill et al., 1997; Winckler & Willen, 2001), postal surveys (Wassink, Grogono-Thomas, Moore & Green, 2003), or meetings in which multiple observers scored videos (Fuller, Bladon, Driver & Barr, 2006; Kaler et al., 2009; Van Nuffel, Sprenger, Tuyttens & Maertens, 2009) or a combination of videos and photographs (Foddai, Green, Mason & Kaler, 2012). Assessments in these studies have mostly been done within face-to-face meetings, generally workshop sessions, but the availability of new technologies allows for the development of alternative formats.

Web-surveys are data collection tools being used successfully in several fields of research, for instance in the psychology field to develop scales of social competence (Yager, 2012), or in the marketing research area to assess consumer's behaviour (van der Heijden, 2003; Horn et al., 2005).

The main purpose of our study was to test whether web-surveys could be used to collect lameness assessments of people working in animal production, health or welfare, with the ultimate objective of collecting a large number of observer's responses and analyzing the grounds for developing an alternative lameness scoring system. At a second level, and specifically, we aimed to analyze the scope for building a modified visual analogue scale (VAS) for dairy goats, as already developed for other species (Nalon et al., 2014; Tuyttens et al., 2009). A modified VAS is a continuous scale tagged with thresholds that divide its continuum into segments representing a numerical rating scale descriptor; therefore, modified VAS retain both the advantages of ordinal and continuous scales (Nalon et al., 2014) and may be more reliable than commonly used numerical rating scales. Our web-survey was designed to identify specific lameness signs to be included in the mentioned segments, and the position of the thresholds along the scale. Accordingly, some of the design options presented in this technical note – e.g., the scoring of three lameness signs using VAS – are associated with the specific aim of the above mentioned study, and hence discussed elsewhere.

To our knowledge this is the first published study in which a web-survey is used with the objective of assessing a health indicator in the veterinary field. We herein detail how the web--survey was developed and applied from a researcher viewpoint, so that others can replicate the experience and benefit from our experience. We further compare our web-survey with other published methods and lay-outs, and describe which kind of research will benefit from this data collection system, using goat lameness as an example.

8.2 DEVELOPMENT AND APPLICATION OF THE WEB-SURVEY

Web-survey preparatory stage

The videos collected for this study were taken between December 2012 and March 2013 in two commercial dairy goat farms in Portugal, with an average milk yield of 2.1 liters per goat per day. The breeds included were Saanen, Alpine and crossbreed. Image collection was done during daily farm routines, mainly as goats exited the milking parlour, making sure that the goats were walking on an accustomed hard, levelled, non-slippery surface where at least four complete walking movements could be assessed. All filming took place within three hours after morning or afternoon milking. The goats were filmed from the side using a fixed camera (Sony handycam HDR-CX190) at 50 frames per second with a 1440 x 1080 pixel resolution. In total we got 450 minutes of observation time. The videos were analyzed by one researcher who produced a total of 115 clips of approximately 10 seconds. Goats were filmed independently of their lameness condition and in a way not to disturb their natural behaviour.

In April 2013, the 115 clips were scored by three experts. Considering that there is no universally established scoring system for lameness scoring in goats, we chose the one developed by Anzuino et al. (2010) because the type of farms and animals were similar in this study. This scale defines score 0 as absence of lameness, score 1 as slight lameness, score 2 as moderate lameness and score 3 as severe lameness. The criteria for expert's recruitment were: clinical experience with goats, and behavioural observation competence with goats and other ruminants. The experts did not contact one another during the scoring process. A total of 82 video clips were approved with the following distribution per lameness score: 0 (n=21 videos), 1 (n=23), 2 (n=15) and 3 (n=23). The remaining 33 video clips were excluded due to direct suggestion of the experts, or because each expert assigned it a different lameness level. This previous scoring was essential in the context of our study to compare the survey respondent's scores with an overall lameness scoring, this being a strategy followed in other studies (Nalon et al., 2014).

Web-survey platform development and testing

Web-surveys are part of Web 2.0 services (such as social networking or micro-blogging), and a sound, easy and cost-effective tool for formal scientific investigation (Buchanan & Hvizdak, 2009). The web-survey used in this study was developed in a *WordPress* platform. No sensitive data such as names of respondents or IP addresses were collected. Although measures were taken to keep the anonymity of each respondent, there was the option to leave an e-mail contact for future communication of results.

The web-survey consisted of three parts. In the first part there were some guidelines explaining the scope, giving a definition of the VAS and an explanation of what we intended respondents to assess (Figure 8). In the second part the respondents were asked about a) age and gender, b) level of education (primary, secondary or higher), c) occupation (farmers and stockpersons, animal scientists, veterinarians, researchers in animal behaviour and students), d) country of residence and e) experience in scoring goat lameness (inexperienced, little experience, experienced or very experienced) (Figure 9).

In the third part of the survey, respondents were asked to assess lameness of animals shown on the videos. Each individual survey was composed of nine videos, randomly selected from the total pool. The order in which each video appeared to the respondent was randomized. Each respondent could watch the videos as many times as desired (Figure 10).

Figure 8. Print-screen of the web-survey presenting the guidelines page

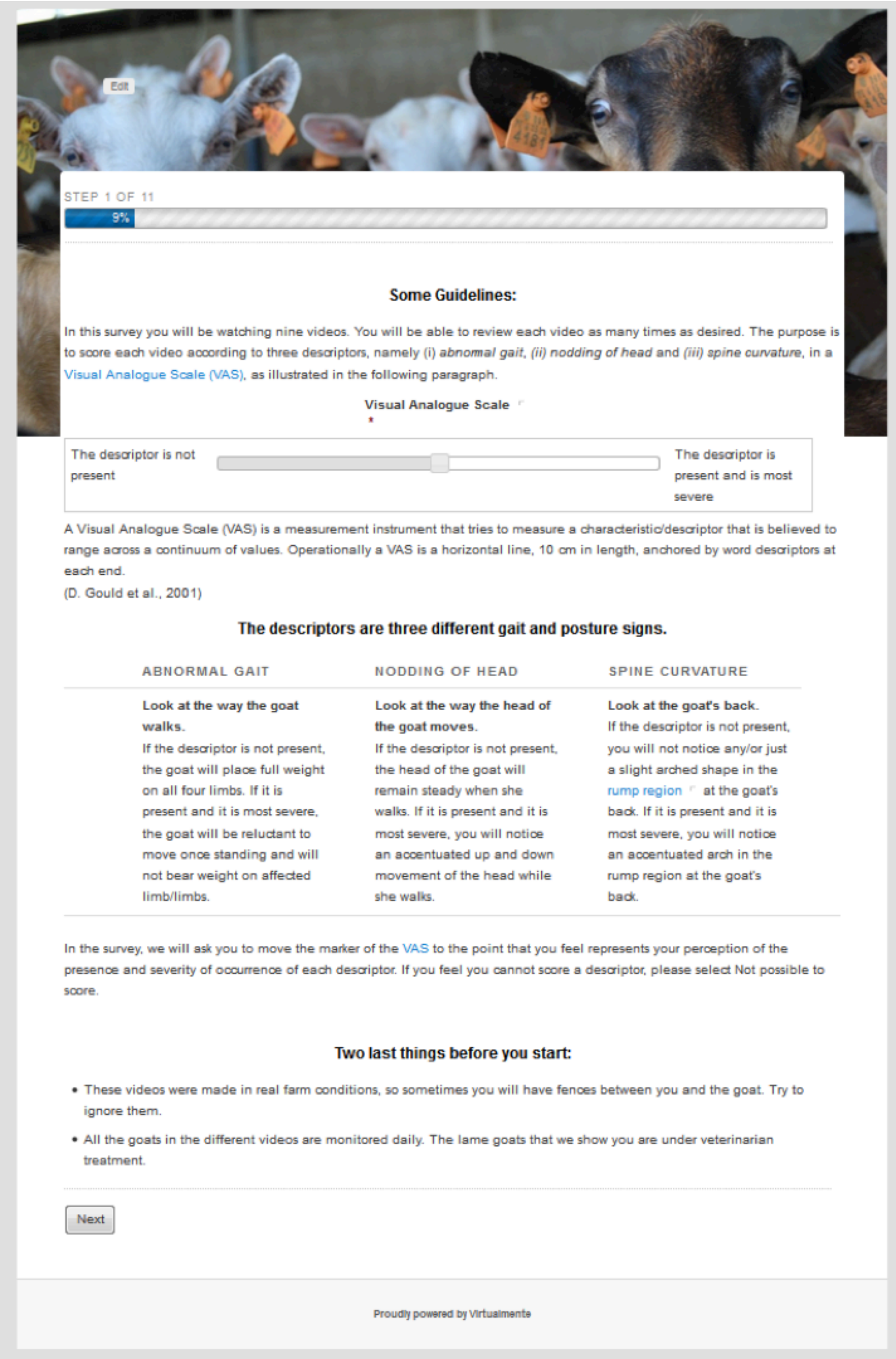


Figure 9. Print-screen of the web-survey presenting the section for the respondent's characteristics collection

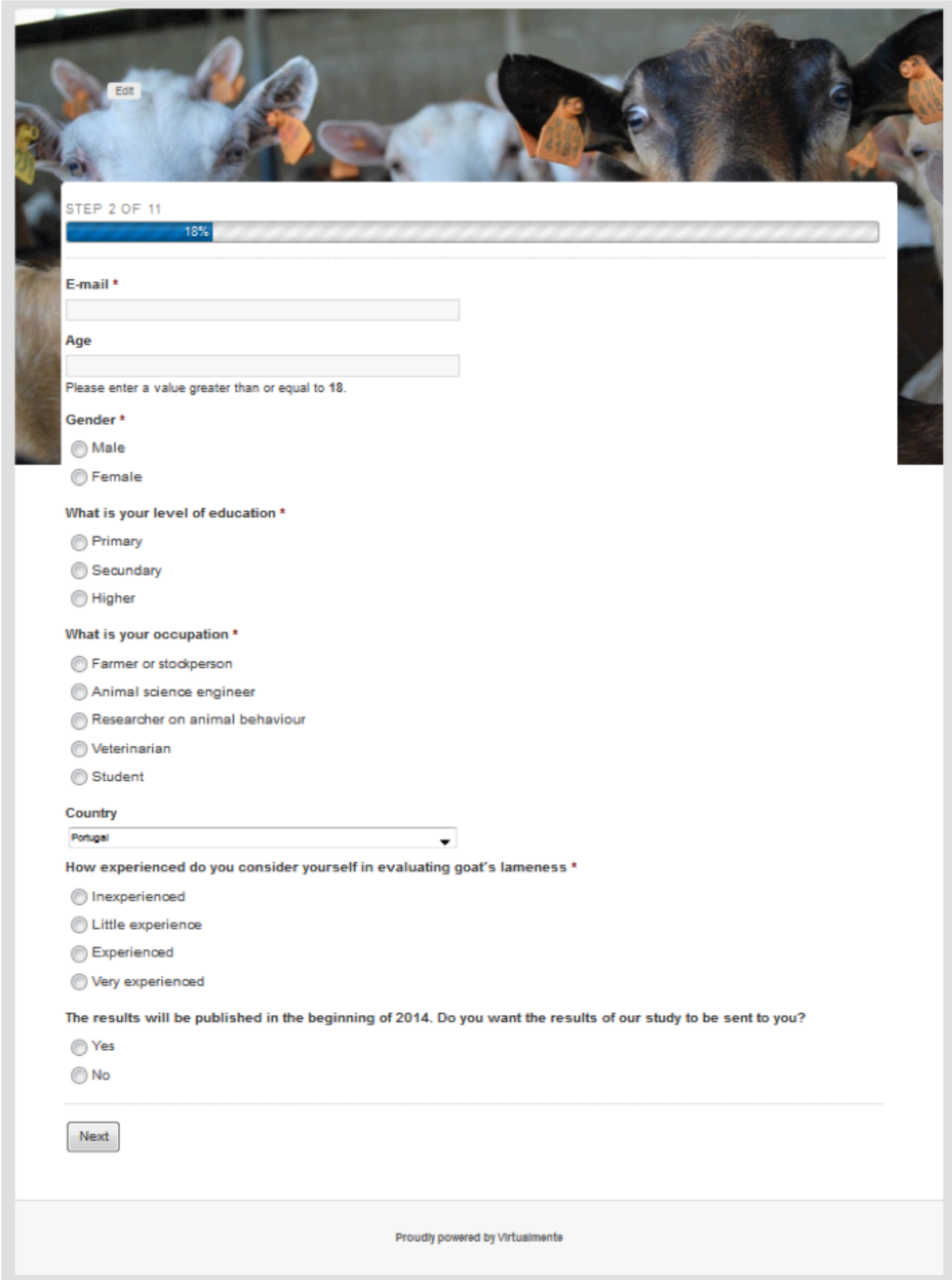
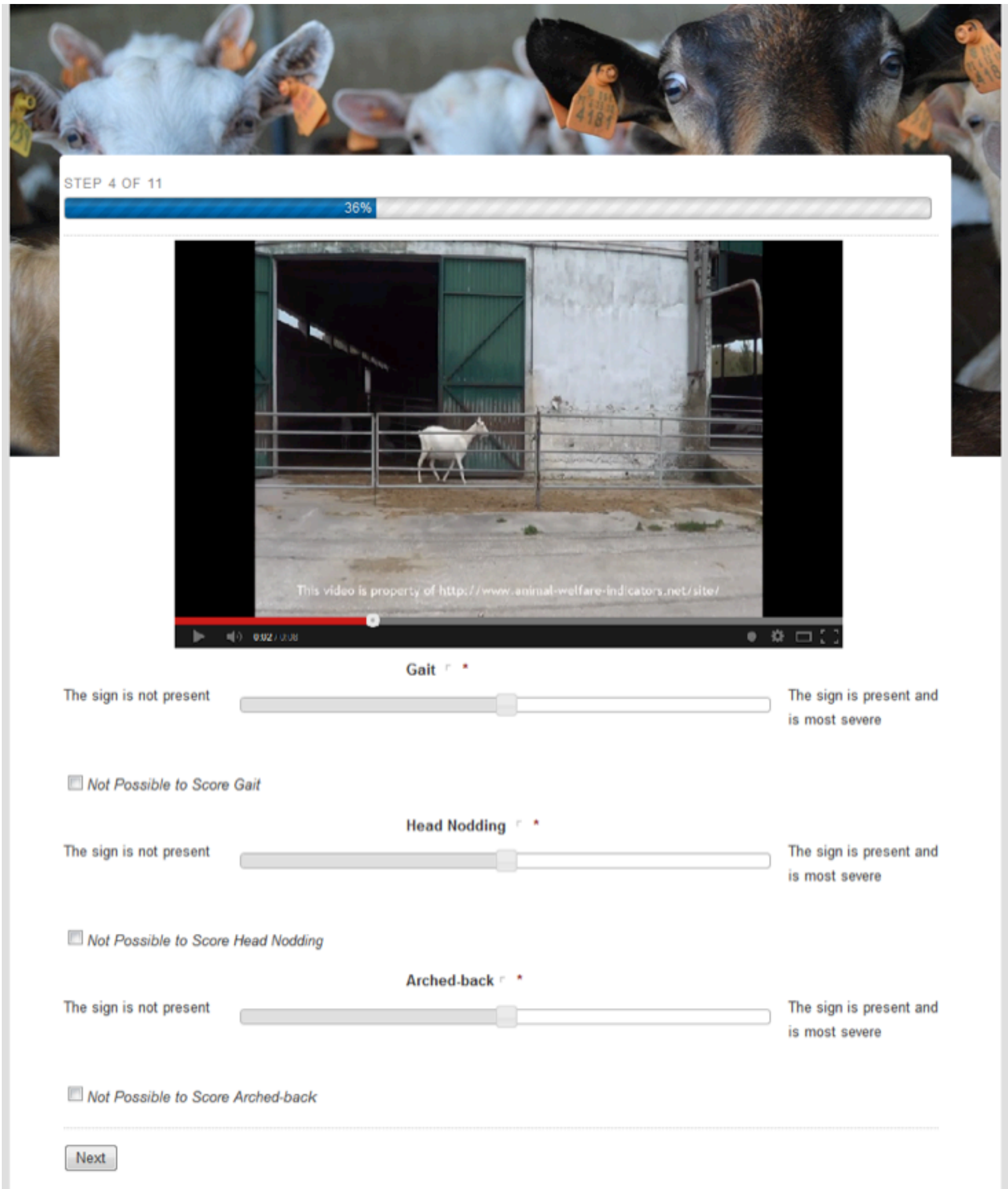


Figure 10. Print-screen of the web-survey presenting a video and three individual VASs



Considering that our study aims at collecting a large number of observer's responses and analyzing the grounds for developing an alternative lameness scoring system, the respondents were asked to score three different lameness signs separately, rather than using the overall scale previously used by the experts, th Anzuino et al. (2010). The first sign to be assessed was "gait" as this is the one used in all lameness scales for goats (Anzuino et al., 2010; Hill et al., 1997; Mazurek et al., 2007). Additionally, we selected two other signs – head nodding and arched-back – that address specific lameness details.

Respondents scored each of the three signs in an individual VAS. The VAS used in the web--survey was a 100 mm line, anchored on its extreme limits by “The sign is not present” and “The sign is present and is most severe”, with these verbal descriptors helping the respondents to understand the intermediate scores. Respondents were asked to move the marker along the continuum and their score was automatically measured from the zero to the respondent’s mark and stored online. The data was expressed to the nearest 0.5mm.
Next to each VAS cursor, there were help buttons with the definition of the sign, as presented in the guidelines. Whenever unable to score the sign, the respondent could select the option “not possible to score”.
Before the web-survey went online, we conducted a two week testing of the platform. The link to the web-survey was sent to 15 eligible respondents. All the process from the link access to the way the web-survey developer collected and sent the data was analyzed. These 15 participations were not considered for the final results of the survey, but gave valuable input in the form of suggestions.

Recruitment of respondents and collection of data

The web-survey went online on the first week of May and ended on the last week of October 2013. Respondents were recruited through announcements to the main international farm animal associations, universities and research institutions, working with small ruminants, mainly goats, or farm animal behaviour. The survey was also announced in internet platforms, such as Facebook and LinkedIn.
Following Buchanan & Hvizdak (2009) ethical advice, the link to the web-survey always provided an explanation of the purpose of the study and how the data was going to be used and ensuring anonymity of information, as if to seek an informed consent for data use.
As already mentioned, the respondents’ ratings were automatically measured and stored online in a mySQL file, and then stored with Microsoft Excel (Microsoft Office 2010).

8.3 WEB-SURVEY OUTREACH

Our web-survey had 2,312 views that resulted in 600 participations, and hence a response rate of 26%. By “views” we mean people who visited the website and started the web-survey but it also includes those that did not finish it.
Of the total, 367 respondents were female and 203 were male, with ages ranging from 18 to 75 years (Mean=36; SD=12). We had respondents from 35 different countries in all five continents – Portugal (261), United Kingdom (54), United States (40), Brazil (29), Canada (22), Italy (19), Germany (15), Austria (14), Norway (12), Switzerland (12), Greece (11), and Spain (11), with all the remaining countries having less than 10 respondents. Respondents’ education, occupation and experience were also very diverse (Table 11).

Table 11. Characteristics (education, occupation and experience) of the final sample of respondents (n=570) to the web-survey on goat lameness assessment.

Education	
Higher	547 (96%)
Secondary	21 (3.7%)
Primary	2 (0.3%)
Occupation	
Veterinarian	270 (47%)
Researcher in animal behaviour	133 (23%)
Animal science professionals	50 (9%)
Student	89 (16%)
Farmer and stockperson	28 (5%)
Experience	
Inexperienced	217 (38%)
Little experience	249 (44%)
Experienced	85 (15%)
Very experienced	19 (3%)

8.4 INSIGHTS FROM DESIGNING AND IMPLEMENTING THE WEB-SURVEY

The main strengths of the web-survey were the number and diversity of the respondents. Our survey collected a total of 570 true respondents (i.e., those that changed the default level of at least one individual VAS), whereas other studies ranged from three (Foddai et al., 2012; Fuller et al., 2006; Kaler et al., 2009) to 209 (Kaler & Green, 2008) respondents. Approximately 47% of the respondents were veterinarians, 23% were researchers in animal behaviour, 16% were students, 9% were animal scientists, and 5% were farmers. It is not frequent for studies to be able to congregate such a large variety of backgrounds. For example, the 40 observers in Tuyttens et al. (2009) and Van Nuffel et al. (2009) studies were all participating in an Animal Welfare Conference, Nalon et al. (2014) used 100 undergraduate veterinary students from the same University and Kaler & Green (2008) used only farmers, and Kaler et al. (2009) only researchers.

The same applies to experience. In our study around 38% of the respondents said to be inexperienced in goat lameness assessment, but 62% considered themselves to have some degree of experience in assessing goat lameness; among these, 29% deemed to be experienced or very experienced. Nalon et al. (2014), Tuyttens et al. (2009), and Van Nuffel et al. (2009) also collected participations from individuals with different levels of experience but Foddai et al. (2012), Fuller et al. (2006) and Kaler et al. (2009) only used experienced observers and Garner, Falcone, Wakenell, Martin & Mench (2002) only naïve observers. In short, our web-survey proved to be appropriate for analyzing the impact of background and experience on, in this case, lameness ratings or on the use and acceptability of different lameness scales.

One of the reasons for the small number and respondents' homogeneity in some of the previously mentioned studies, is the logistic cost associated with data collection. A key advantage of using a web-survey is that the researchers do not have to organize meetings to collect the observers' perception, and this might be especially important when the studies have more than one session planned or there are time constraints associated. Another advantage in having online participation is that respondents have the opportunity of answering at their own pace, without being influenced or pressed by the interviewer or other respondents in the room.

Using a web-survey also carries important technologic advantages in terms of experimental design and data transcription. In our study, each respondents' score was automatically collected and stored online. Additionally, if the respondent felt he or she could not perform the scoring, there was the option to select "not possible to score". In general, other studies have used paper with information having to be later transcript into a database. The studies generally do not discuss this, but the process may entail an enormous amount of work. Moreover, Rhodes, Bowie & Hergenrather (2003) estimated that automatic data is associated with less entry mistakes, and with 20 to 80% saving in data collection costs.

The web-survey platform employed several technological tools (e.g. help buttons) with the aim of providing an attractive and easy-going interface. Additionally, the fact that every time a different respondent started the survey a different selection of randomized videos was scored, allowing for different sets of clinical cases to be used in the survey, instead of a reduced selection or a limited combination of videos, which is a very important feature in clinical context.

Through the characteristics and strengths mentioned above, a web-survey can potentiate on--going investigation or open new research lines. For example, online available data can be helpful in studies where researchers want to assess intra-observer agreement. Researchers can either insert the same video twice in the same scoring session, as we did in our experimental design, or invite the observers to repeat their scoring online a few days/weeks later. In live surveys, due to logistic constrains, the observations sets are usually separated only by a few hours with observers probably remembering previous scorings. This is particularly evident in most on-farm studies (Thomsen et al., 2008; Welsh et al., 1993), with only one or two hours elapsing between observations.

Another process in animal sciences that may benefit from the use of web-surveys is the evaluation of the effect of training in lameness scoring. Most studies include training sessions that are usually short and completed just before the scoring sessions. For example, Tuyttens et al. (2009) discusses how short training before the scoring sessions can influence inter-observer reliability. Performing or assessing the training online could overcome some of these drawbacks. Additionally, by establishing a fixed time for video or photographs visualization, or by comparing videos in different settings, it is possible to simulate on-farm limitations when scoring lameness, thus improving training programs.

Another potential advantage is to effectively identify true respondents by investigating different levels of competence. To be able to fulfil this objective, we can test for inconsistencies (depicting a conflict between the decision-maker assessment and a fixed model), a concept used in decision analysis studies (González-Pachón et al., 2014). When the fixed model is a numerical rating scale we can assess the order of the answers, ascertaining for ordinal consistency (Keeney, 1976). When the fixed model is a continuous scale it is important also to assess the intensity of the assessment, and therefore we can apply different cardinal consistency levels (Keeney, 1976). This way of analyzing lameness data can be an alternative to inter-observer reliability, allowing for individual variation to be considered.

Because of the web-survey enormous outreach potential, even apparent low response rates are usually associated with high participations. For example, our response rate (26%) is lower than that obtained in the postal survey by Wassink et al. (2003) that achieved a 53% response rate. However, our figure corresponds to 570 respondents, a number far superior to the 209 in the study on sheep footrot.

Some drawbacks associated with this data collection method might be pointed out. For instance, the recruitment strategy used does not ensure a true random sampling of respondents, but as Rhodes et al. (2003) discuss in their work, it is always a challenge to make generalizations between a sample and the target population. In fact, nearly all lameness assessment studies are based on small groups of observers in workshops or conferences or in expert opinions in which extrapolations should assume some margin for mistake.

Additionally, the cost of the development and implementation of the web-survey can also be considered a weakness for some projects. To estimate the cost of a web-survey several items have to be taken into account (these values report to private companies operating in Portugal): the web-survey platform development (500-1,000€, depending on the design needs, requisites in technological tools and interface design), and the technical support and data transcription (20-50€, per six months). Additional costs with the web-hosting (10-15€, per month), and web-site domain (10-20€, per month) should be considered.

The time needed to prepare the web-survey can range from two weeks to one and a half month. This, associated with the time the web-survey is online, can also be considered a downside considering the timeline of some projects. However, it is important to acknowledge that we do not know the costs and time associated with data collection methods used in other published studies.

The potential use of videos or photographs of animals in poor welfare conditions by people with purposes other than the participation in the web-survey should always be taken into account. Two strategies can be implemented: never publish videos or photos that were not authorized by the animals' legal caretaker and make an initial identification/registration to the web-survey mandatory. However, this can send to the participant a message of lack of anonymity which is not desirable in any study.

In conclusion, it was demonstrated that a web-survey can be a very valuable tool for data collection in veterinary or animal production research, such as lameness in dairy goats. It has the potential to reach a large number of respondents of varied background and experience, as well as providing ready to use data.

CHAPTER 9

Making the case for developing alternative lameness scoring systems for dairy goats

Chapter 9 presents the grounds for developing a modified VAS to assess lameness in goats, and discusses the level of measurement that can be used on dairy goat on-farm welfare assessment.

This chapter has been accepted for publication as: Vieira, A., Oliveira, M.D., Nunes, T. & Stilwell, G. (in press) *Making the case for developing alternative lameness scoring systems for dairy goats*. (Accepted at Applied Animal Behaviour Science)

A. Vieira conceived the idea for paper and designed the experiments. The author also collected and processed the data in cooperation with the informatics company, analysed and discussed the results, finally writing the manuscript.

Lameness is a behavioural indicator of pain that negatively affects dairy ruminants' health and welfare. Lameness is generally assessed by subjective methods, based on the observation of the animal's behaviour, using numerical rating scales (NRSs) – the most common scoring system – and visual analogue scales (VASs).

Distinct drawbacks have been associated with both types of scales. NRSs have been associated with a reduced sensitivity to capture variations in lower levels of lameness that may adversely impact animals' welfare assessments. VAS is considered too subjective and associated with low user-acceptance. Recent literature on health scales has been focusing on the development of modified VASs that define equal ranges along the scales' continuum, with thresholds representing a NRS descriptor. Although good results have been reported in using these modified VASs for lameness scoring, the literature recognizes that it is paramount to test whether existing NRS descriptors are equal spaced in the VAS continuum, as well as research the extent to which lameness intensity varies for different lameness and posture signs used to define NRS descriptors. The answers to these questions are vital for the development of new modified VASs to assess lameness in goats.

Aiming to address these questions we collected and analyzed lameness scorings using individual VASs to score three lameness and posture signs (gait, head nodding and arched--back). Lameness scorings were performed through a video-based web-survey. We collected a total of 570 valid participations from respondents with different occupations and experience. Because of expected differences in the respondents' ability to assess lameness, we analysed answers by levels of cardinal consistency. Our results showed: 1) respondents' difficulties in recognizing and discriminating across some NRS descriptors; 2) these difficulties varied with the lameness severity and with the lameness sign; 3) gait, the basis for NRS lameness descriptors in goats, was not scored evenly spaced along the continuum of the VAS; 4) similar results were found for the head nodding and arched-back signs.

In conclusion we suggest that the exact location of the thresholds along the continuum of the VAS should be reassessed, and the inclusion of different lameness and posture signs in scales should receive further attention before new modified VASs are developed. Moreover, the use of NRS in lameness scoring should only consider their ordinal measurement properties, therefore giving space for developing, validating and using alternative lameness scoring methods in farm animals that allow for higher measurement levels.

9.1 INTRODUCTION

Lameness is an important behavioural indicator of pain caused by claw or limb injury or disease. It has been shown to negatively affect dairy ruminants' feed intake and milk yield (Christodoulopoulos, 2009; Flower & Weary, 2009; Green, Hedges, Schukken, Blowey & Packington, 2002; Palmer, Law & O'Connell, 2012) and fertility (Eze, 2002; Flower & Weary, 2009; Hernandez, Garbarino, Shearer, Risco & Thatcher, 2005). It has also been shown that lameness affects individual (Blackie, Amory, Bleach & Scaife, 2011; Juarez, Robinson, DePeters & Price, 2003) and social behaviour (Galindo, Broom & Jackson, 2000) of dairy animals. Consequently, lameness is regarded as one of the most serious health and welfare problems in dairy ruminants (Flower & Weary, 2009; Webster, 2001).

Lameness can be assessed by objective (based on the use of equipment that collects kinetic and kinematic data) or subjective (based on the observers' ratings using different scoring systems) methods, although the latter are more generally used (Flower & Weary, 2009; Meagher, 2009).

Within subjective methods, it is very important to consider the effects of both the observer and the scoring system (Flower & Weary, 2009). Different scoring systems and scales have been used for different animal species. A recent review on locomotion scoring systems in dairy cows identified 25 different scoring systems of which 22 were numerical rating scales (NRSs), and three were visual analogue scales (VASs) (Schlageter-Tello et al., 2014). For sheep we found three NRSs (Kaler et al., 2009; Ley, Livingston & Waterman, 1989;

Welsh et al., 1993) and one VAS, developed by Welsh et al. (1993). For goats we found four NRSs (Anzuino et al., 2010; Christodoulopoulos, 2009; Hill et al., 1997; Mazurek et al., 2007) but no continuous lameness scale, namely VAS.

NRSs are explicit grading methods in which each individual is scored accordingly to different lameness descriptions that correspond to a whole number (Gaynor & Muir, 2008). This suggests that the different descriptors represent an equal increase or decrease in lameness intensity, which may not be true (Gaynor & Muir, 2008). NRSs are artificial constructs as lameness can be seen as varying in a continuous trait; hence, when we only allow the observers to make scorings based on a limited number of descriptors, there is reduced sensitivity on animals' welfare assessment and a loss of valuable information (Nalon et al., 2014; Streiner & Norman, 2008).

Therefore the VAS, as a continuous scale, may be considered a better alternative for lameness scoring. The VAS has the advantage of not imposing a choice for limited and closed categories, being possible to score a change on the VAS even if a change between categories would not occur, and hence being more sensible to small variations in signs (Averbuch & Katzper, 2004; Paul-Dauphin et al., 1999; Welsh et al., 1993). Nonetheless, VASs have not made their way into lameness scoring as they are generally viewed as being too subjective, with low user-acceptance, and difficult to use in farm conditions (Engel et al., 2003; Kaler et al., 2009).

Current research in scale development has recently been focusing on the development of modified VASs to assess different health indicators, for example, pain in humans (Averbuch & Katzper, 2004) and lameness in dairy cows (Tuyttens et al., 2009) and in sows (Nalon et al., 2014). According to these studies, the modified VAS holds some of the NRS strengths, as it provides extra help by placing thresholds, functioning as additional anchors or cues, along the scale to guide observers in their scorings, increasing perception, helping the observers to make consistent choices and hence increasing inter-observer reliability while keeping a higher resolution and lower error probability.

These examples of modified VASs split the continuum into equal ranges, placing (underlying) NRS descriptors as thresholds to these ranges for which a text box composed by different lameness and posture signs is added. However, it is imperative to assess if the distribution of the NRS descriptors in the VAS is in fact evenly spread.

In this study we assess the grounds for developing a modified VAS to assess lameness in goats by investigating the distribution of different lameness and posture signs along the scales' continuum. We did so by adopting a two-stage approach: first by observing the respondents ability to recognize signs' intensity increase or decrease, which are needed to discriminate the underlying lameness descriptors; and then, by analysing how different gait and posture signs are scored individually in a VAS.

9.2 MATERIALS AND METHODS

Web-survey and data collection

The web-survey consisted of three parts. In the first part guidelines explaining the scope and how respondents should conduct the scoring were presented. In the second part the respondents were asked about (1) their age and gender, (2) level of education (primary, secondary or higher), (3) occupation (farmers and stockpersons, animal scientists, veterinarians, researchers in animal behaviour or students), (4) country of residence, and (5) experience in scoring goat lameness (inexperienced, little experience, experienced or very experienced). In the third part of the survey, the respondents were asked to answer a questionnaire. Each questionnaire was composed of nine videos, randomly selected from a pool of 82 videos. The order in which each video appeared was randomized and each respondent could watch the videos as many times as desired. The videos were previously scored by three experts using a commonly used four-descriptor NRS, the Anzuino et al., (2010), which was developed for similar type of animals and farms. This scale considers one level for normal gait and therefore for absence of lameness (descriptor 0), and three consecutive increasing levels of lameness (descriptors 1, 2 and 3). This scoring system was used as the basis to compare the respondent's scores with an overall lameness scoring. As in lameness scoring no gold standard is available, we used the expert's scoring consensus to establish the overall lameness status, thereby adopting a similar strategy to other studies that have looked into the validity and reliability of lameness scoring scales (Engel et al., 2003; Nalon et al., 2014; Tuyttens et al., 2009; Van Nuffel et al., 2009).

In each survey the respondents were asked to score three different lameness signs separately, rather than using the scale previously validated by the experts, the Anzuino et al. (2010) lameness scoring system. The first sign to be scored was "gait" as this is the one used in all lameness scales for goats (Anzuino et al., 2010; Hill et al., 1997; Mazurek et al., 2007). Additionally, we selected two other signs – head nodding and arched-back – that address specific lameness posture signs and are commonly used in other species. Each respondent scored the three signs in an individual VAS, which consists on a nonverbal scale that records lameness in a continuous way. This scale was first developed for use in pain assessment in humans (Scott & Huskisson, 1976) and since then has been used to measure a variety of subjective phenomena in the behavioural and social sciences and is considered of potential value for the measurement of different clinical conditions (Wewers & Lowe, 1990). The VAS used in the web-survey was a 100 mm line, anchored on its extreme limits by: "The sign is not present" (corresponding to a 0 scoring) and "The sign is present and is most severe" (anchored in 100). Each respondent was asked to move the marker along the continuum and his/her score was automatically measured from the zero to the respondent's mark and stored online. The data was expressed to the nearest 0.5mm. Whenever unable to score the sign, the respondent could select the option "not possible to score". Further details of the web-survey development and implementation are available elsewhere (Vieira et al., submitted). The data collected from May to October 2013 was exported to Microsoft Excel (Microsoft Office 2010), and analysed using the R statistical language (R Core Team, 2013).

Design for consistency testing

To investigate whether the respondents were able to recognize the change in signs' intensity, that is needed to discriminate different descriptors, we defined four competence levels. These levels are based upon four cardinal consistency levels, in line with consistency concepts from the decision analysis literature (Keeney, 1976; Keeney & Kirkwood, 1975) – see Table 12. Descriptors 0 to 3 presented in Table 12 depict the descriptors in the Anzuino et al., (2010) lameness scoring system. Overall answers were included in dataset A. The first cardinal consistency level resulted in dataset B, composed by all respondents that were able to properly distinguish between non-lame (descriptor 0) and severely lame (descriptor 3) animals – this is in line with previous studies that pointed out that severely lame animals are almost always identified by the observers (Brenninkmeyer et al., 2007; Kaler et al., 2009; March et al., 2007; Christoph Winckler & Willen, 2001).

Table 12. Description of the datasets that correspond to the different levels of cardinal consistency being investigated. The different levels were built in line with consistency concepts from the decision analysis literature and depict state of the art literature in observers' lameness scoring inconsistencies.

Dataset	Cardinal consistency levels
A	All respondents. No exclusion criteria applied.
B	All respondents capable of distinguishing between non-lame (descriptor 0) and severely lame animals (descriptor 3).
C	Respondents able to distinguish non-lame and slightly lame (descriptors 0 and 1 considered together) from moderate and severely lame animals (descriptors 2 and 3 considered together).
D	Respondents able to distinguish non-lame (descriptors 0 and 1 considered together), moderate and severely lame animals.
E	All consistent. Respondents able to distinguish across all descriptors of lameness.

Note: The lameness descriptors mentioned are the ones described in Anzuino et al. (2010) lameness scoring system.

The second cardinal consistency level – analysed through dataset C – contained all respondents that were able to distinguish non-lame (descriptor 0), and slightly lame (descriptor 1) animals, from obviously lame ones [i.e., those with evident compromised motility, therefore moderate (descriptor 2) and severely lame (descriptor 3)]. This level is in line with studies describing that even with low levels of training and experience, individuals are able to distinguish between these groups of animals (Brenninkmeyer et al., 2007; March et al., 2007).

As available evidence points out for difficulties in scoring and discriminating the lower levels of lameness (Brenninkmeyer et al., 2007; Kaler et al., 2009; Nalon et al., 2014; Winckler & Willen, 2001), a third dataset – D – was also analysed for individuals who were able to distinguish all descriptors, with the exception of descriptors 0 and 1.

The last cardinal consistency analysis included respondents that were able to distinguish all descriptors so being fully consistent. This fourth group was designed dataset E.

To assess differences in the respondents' consistency, we applied these cardinal consistency levels (Table 12) to the three lameness signs separately (Table 14) and for the four datasets (note that respondents were included in all datasets for which they met the selection/inclusion criteria presented in Table 12).

Statistical design to test for evenly spaced descriptors

To test if the lameness descriptors were evenly spaced in all the datasets analysed and considering the three signs separately, we compared two alternative regression models and compared their explanatory power by means of their adjusted R^2 and of their Akaike information criterion (AIC). In regression analysis, as fit increases with the number of explanatory variables included in the regression, we used the adjusted R^2 to deal with this fact (Wooldrigde, 2008). Additionally, we used the AIC as a measure of the relative quality of statistical models such that minimum AIC values captures a better fit (Sakamoto, Ishiguro & Kitagawa, 1986). The higher adjusted R^2 and the lower AIC were used as the criteria for deciding upon the best model that captures the distribution of descriptors along the VAS. The two models were:

1) a linear regression model that assumes that descriptors are evenly spaced, i.e., that defines Y (the VAS score), by means of $\alpha + \beta x$, where x represents the lameness descriptor accordingly to the experts; x can take the values of 0, 1, 2 and 3. Hence, lameness descriptor 0 will be converted in α, lameness descriptor 1 will be $\alpha + \beta$, lameness descriptor 2 will be $\alpha + \beta * 2$, and lameness descriptor 3 will be $\alpha + \beta * 3$. This model assumes a univariable linear relationship where Y (the VAS score) is linearly related to the explanatory variable, the ordinal score accordingly to the experts;
2) a multiple linear regression that does not impose that descriptors are evenly spaced, i.e., that defines Y (the VAS score), by means of $\alpha' + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, where x_i is a variable equal to i when the observation that corresponds to the video was scored with the descriptor according to the experts, and zero otherwise. It means lameness descriptor 0 will be converted in α', lameness descriptor 1 will be $\alpha' + \beta_1 * 1$, lameness descriptor 2 will be $\alpha' + \beta_2 * 2$, and lameness descriptor 3 will be $\alpha' + \beta_3 * 3$.

If the multiple regression model presented the higher adjusted R^2 (this case being expected if lameness descriptors were not evenly spaced), the next step was designed to test if the partial regressions coefficients β_1, β_2 and β_3 were statistically equal. In order to do this, pairs of coefficients were statistically compared (e.g., $\beta_2 - \beta_1 = 0 <=> \beta_2 = \beta_1$, as the null hypothesis of the T test (Wooldrigde, 2008)).

9.3 RESULTS

Respondents' characteristics

We first conducted an analysis of the 600 participations so as to eliminate respondents that failed to complete the survey. This led to the elimination of 30 respondents that did not change the default level of at least one individual VAS. Of the final 570 accepted respondents 367 were female and 203 were male, with ages ranging from 18 to 75 years (Mean=36; SD=12). The characteristics of the 570 accepted respondents (from 35 countries, with 45% from Portugal) are shown in Table 13.

Table 13. Overview of the characteristics (gender, education, occupation and experience) of the final sample of respondents. We present the total number (and the percentage of the total number of respondents) of each characteristic in assessment.

	Individual characteristics	Number of respondents N=570
Gender	Female	367 (64%)
	Male	203 (36%)
Education	Higher	547 (96%)
	Secondary	21 (3.7%)
	Primary	2 (0.3%)
Occupation	Veterinarian	270 (47%)
	Researcher in animal behaviour	133 (23%)
	Animal science professionals	50 (9%)
	Student	89 (16%)
	Farmer and stockperson	28 (5%)
Experience	Inexperienced	217 (38%)
	Little experience	249 (44%)
	Experienced	85 (15%)
	Very experienced	19 (3%)

Answers' cardinal consistency

The data presented in Table 14 show the number of respondents that integrate each dataset after applying the levels of cardinal consistency (Table 12) to the three lameness signs separately. As the cardinal consistency levels get more demanding fewer respondents respect the consistency conditions, in all the lameness signs.

For the three signs it was shown that only 10% of observers were inconsistent when distinguishing between the non-lame (descriptor 0) and the most severe (descriptor 3) cases, and less than 41% were fully consistent when considering the four descriptors in assessment. The major drop in consistency occurs when it is asked to distinguish between non-lame and slightly lame animals. Considering the gait example, the percentage goes from 71% to 41%, while in the case of head nodding it goes from 57% to 32%. Between datasets C and D (that differ in the fact that in dataset D makes the distinction between moderate cases), the percentages of respondents' inclusion is not very different – again for the gait example, the percentage drops from 76% to 71%. In all the datasets, higher levels of cardinal consistency were observed for gait, when compared with the other two signs. The arched--back sign was the one that presented higher levels of inconsistency in all the analysed datasets, with the percentages of inclusion of respondents ranging from 90% (dataset B) to 25% (dataset D).

Table 14. Datasets (identified from B to E with the number of respondents that respect different levels of cardinal consistency considered, for each lameness sign (and as a percentage of the total number of respondents included in dataset A)

	B	C	D	E
Gait	558 (98%)	436 (76%)	405 (71%)	236 (41%)
Head nodding	531 (93%)	387 (68%)	326 (57%)	180 (32%)
Arched-back	513 (90%)	308 (54%)	245 (43%)	145 (25%)

Testing for evenly spaced descriptors

The regression models were applied to each combination of lameness sign and dataset (using the answers of the individuals identified in Table 14) and statistically confirmed a non-even distribution of the four underlying lameness descriptors along the VAS continuum. The multiple linear regression model showed to be more suitable for this data when compared with the linear regression model (as captured by a higher adjusted R^2 and lower AIC). In Table 15 we present the coefficients and results for the two regression models applied to the "All consistent" dataset and for the gait sign – we selected this case as an example. Similar results in adjusted R^2 and AIC were obtained for other datasets.

We found statistically significant differences in the pairwise comparisons of the partial regression coefficients for almost all signs and datasets (for the multiple regression), re-enforcing the previous suggestion of unequal distance between lameness consecutive descriptors of existing NRS. The only exception was in the dataset C (where we compare respondents able to distinguish the non-lame and slightly lame group (descriptors 0 and 1 considered together) from the moderate and severely lame animals (descriptors 2 and 3 considered together) for arched-back, in which the difference between the coefficients associated with lameness descriptors 1 and 2 was not found to be statistically significant.

Table 15. Summary of the linear and multiple linear regression models applied to the "All consistent" dataset and for the gait sign. The linear model implies a linear relation between the VAS score given by the respondents and the experts' ordinal score; hence it fits data where we can observe evenly spaced (consecutive) lameness descriptors. The multiple linear model assumes there is a simultaneous, and different, effect of the different lameness descriptors on the VAS score; for this reason it allows for an uneven distribution of the lameness descriptors. Higher adjusted R^2 and the lower AIC are used as the criteria for determining the best model that fits the distribution of descriptors along the VAS.

Model	Coefficients				Adjusted R^2	Residual Standard error (degrees of freedom)	AIC
	Estimate	Standard error	T value	Significance			
LINEAR REGRESSION MODEL							
Expert's scoring lameness descriptor 0 (α)	6.74	0.86	7.80	0	0.58	23.56 (2087)	19132.34
Expert's scoring lameness descriptor 1, 2 or 3 (β)	24.96	0.46	54.00	0			
MULTIPLE LINEAR REGRESSION							
Expert's scoring lameness descriptor 0 (α')	12.52	1.01	12.42	0	0.60	22.94 (2085)	19023.44
Expert's scoring lameness descriptor 1 (β_1)	12.67	1.42	8.92	0			
Expert's scoring lameness descriptor 1(β_2)	20.05	0.71	28.26	0			
Expert's scoring lameness descriptor 3 (β_3)	24.69	0.47	51.98	0			

9.4 DISCUSSION

Our web-survey respondents probably represent usual users of lameness scales with the majority having at least some experience in assessing goat lameness.

This novel data collection method allowed us to assemble a large dataset enabling us to answer the questions proposed in the beginning of the study and to establish the grounds for a modified VAS for lameness in goats.

To answer if respondents could recognize intensity changes in signs, we explored consistency through a cardinal consistency assessment design. This strategy was adopted as we expected different levels of competence in lameness scoring among our sample of respondents. In decision analysis studies, inconsistency represents a conflict between the decision-maker assessment and a fixed model (González-Pachón et al., 2014). This is also in line with the definition of inconsistency in inter-observer reliability lameness studies: to verify whether an observer's rating matches the NRS model, ascertaining for ordinal consistency. Similarly to ordinal consistency, cardinal consistency allows not only for assessing the order of the answers, but also the intensity of the assessment (Keeney, 1976).

To verify respondents' consistency, respondents' scores were analysed separately per lameness sign and organized into four datasets as previously explained. It is important to remember that the same respondent could be integrated into more than one dataset. In this analysis, we were not interested in the lameness scoring of respondents who only had a specific level of consistency, but wanted to compare lameness scoring results by increasing levels of consistency.

Our results showed that respondents had variable difficulties in recognizing the increase and decrease in intensity of the three lameness signs. These difficulties were first recognized among different levels of competence in lameness scoring. A very high proportion of respondents was able to differentiate the severely lame from the true non-lame goats (dataset B) in the three lameness signs observed, which is in line with results from previous studies (Brenninkmeyer et al., 2007; Kaler et al., 2009; March et al., 2007; Winckler & Willen, 2001). However, when analysing gait, only less than half distinguished the true-non lame descriptor from the slightly lame descriptor (dataset E), showing in goats what was already studied in other species – most difficulties in lameness scoring were in the lower side of the scales (Brenninkmeyer et al., 2007; Kaler et al., 2009; Nalon et al., 2014; Winckler & Willen, 2001).

Comparing datasets C and D, which differed in the inclusion of respondents that could differentiate moderate cases of lameness from the severe, the number of respondents was not very different, especially considering the gait sign. This result contradicts the recommendation to aggregate descriptors given by some authors (Brenninkmeyer et al., 2007; Channon et al., 2009), showing that for moderate cases this can entail a loss of important information in welfare assessment by failing to detect moderate cases before they become severely lame.

The number of respondents that fit into the head nodding and especially the arched-back datasets is consistently lower in all the cardinal consistency levels, when compared to gait, displaying more difficulties in scoring these signs. For this reason, the inclusion of different gait and posture signs in lameness descriptors should be done with care. Our results apply to lameness scoring in goats, but we believe that this question should receive further attention in future studies in other species, for which the inclusion of these lameness posture signs in NRS descriptors is very common. The cardinal consistency levels assessment was an effective way of analysing lameness data, excluding respondents that were not true-respondents, but allowing for individual variation to be considered.

To test if consecutive descriptors were evenly spaced, we compared two regression models: a linear regression model and a multiple linear regression model. The first model implies a linear relation between the VAS score given by the respondents and the experts' ordinal score; hence it fits data where we can observe evenly spaced (consecutive) lameness descriptors. The second model assumes that different lameness descriptors have distinct effects on the VAS score; for this reason it allows for an uneven distribution of the lameness descriptors. The multiple linear regression model was the one that better fitted all the combinations of datasets and lameness signs. Therefore, it suggests that respondents did not perceive these lameness descriptors as being evenly spaced. This result was observed in all lameness signs.

Three studies have examined the issue of evenly spaced lameness descriptors – two with dairy cows (Engel et al., 2003; Thomsen et al., 2008) and one with sheep (Welsh et al., 1993). Engel et al. (2003) found lameness descriptors to be equally distant; however it should be pointed out that this study did not include videos with the most severe lameness cases. Welsh et al. (1993) and Thomsen et al. (2008) concluded that lameness descriptors were not equally distant, but also included very few animal scorings in the higher descriptors. The three studies used a limited number (maximum 10) of observers with experience in lameness scoring (this is not stated in Engel et al. (2003) but assumed within the text), and all used a five-point NRS. The data collection was performed on-farm in Thomsen et al. (2008) and Welsh et al. (1993), and during five live sessions using video analysis in Engel et al. (2003). Our results are in line with those of Thomsen et al. (2008) and Welsh et al. (1993). An added value of our study is the very large number of respondents, enabled by the use of a web-survey, and the fact that we explore different lameness attributes separately.

Our results suggest that further research should be conducted to assess the exact location of thresholds along the continuum of the VAS, as the lameness descriptors cannot be treated as evenly spaced and furthermore justify the inclusion of different lameness signs for the labelling of such scales. Therefore, the results from our study can be seen as the basis for the development of new modified VAS for lameness scoring in goats, but are also relevant for the developers of modified VAS in other species.

Another additional contribution of our study regards the assessment of the proper measurement level in NRS. In clinical and behavioural research there is often a need for mean and variances calculation when comparing outcomes for different groups submitted to different treatments. In those cases, the use of an interval or ratio level of measurement is paramount. An interval or ratio level [with ratio scales applying if it is possible to determine a meaningful zero in the scale, the scale assumes a ratio level of measurement (Streiner & Norman, 2008)] of measurement allows for mode, median, arithmetic mean and variance to be calculated (Stevens, 1946), while in ordinal scales only a limited number of calculations can be performed because the distance between descriptors is unknown (Merbitz & Morris, 1989; Streiner & Norman, 2008). As our results show that the intervals between consecutive NRS descriptors are not equal and can change with the lameness sign considered, we recommend that only an ordinal level of measurement should be considered in NRS descriptors. This highlights the need for the development a new lameness scale that can encompass higher levels of measurement assessment, like interval or ratio scales.

9.5 CONCLUSIONS

Our web-survey showed that respondents had variable difficulties in discriminating lameness descriptors, and that these difficulties changes with the lameness sign. Furthermore, the scoring of different lameness signs was not evenly spaced along the VAS continuum, meaning that the respondent's ratings did not reflect an equal increase or decrease in lameness intensity between consecutive NRS descriptors. Our results set a new approach for the development of modified VAS, showing that the exact location of the thresholds along the VAS continuum has to be carefully assessed, and that the diagnostic value of each lameness sign should also the further investigated. Additionally, attention should be given to the proper level of measurement that NRSs convey. Our study focused on dairy goats, but we believe that our results apply to other species, especially ruminants for which lameness descriptors are frequently based on the three signs studied.

CHAPTER 10

On-farm welfare assessment of dairy goat farms using animal-based indicators: The example of 30 commercial Portuguese farms

Chapter 10 presents the results of the prevalence of the indicators in the prototype on-farm welfare assessment protocol prototype developed within the AWIN project.

This chapter has been submitted for publication as: Can, E., Vieira, A., Battini, M., Mattiello, A. & Stilwell, G. *On-farm welfare assessment of dairy goat farms using animal-based indicators: The example of 30 commercial Portuguese farms*. Manuscript submitted for publication.

Part of this manuscript was also published as the following Master thesis that the author co- supervised: Can, E.M.A.V. (2015). Welfare assessment in Portuguese dairy goat farms: On-farm overall feasibility of an international prototype. Veterinary Medicine Master Thesis. Lisboa: Faculdade de Medicina Veterinária - Universidade de Lisboa

A. Vieira designed the study, participated in the data collection, and advised on data analysis. Finally, the author actively discussed all the results and drafted the manuscript.

On-farm welfare assessment is of growing importance in the livestock production sector. In order to correctly address this issue in the goat dairy sector there is the need to identify the main welfare problems that can be found in farms across countries. By the application of a welfare assessment prototype protocol, using animal-based indicators developed by the AWIN project, this study aimed to have an insight on the main welfare problems affecting dairy goats (*Capra hircus*) in Portugal.
Initially, thirty Portuguese dairy goat farms, organized in three size categories, were assessed from January to March 2014. Pen-level observations were carried out on 2715 goats and detailed individual observations were performed on 1172 of these animals. The main areas of concern were claw overgrowth, queuing at feeding, very fat animals, poor hair coat condition, and improper disbudding. The animal-based indicators' prevalence showed that these welfare issues have a tendency to be affected by farm size, with larger farms heading higher concerns, although these differences were not statistically significant at the 0.05-level.

Subsequently, to investigate consistency over time, 10 of the 30 farms were revisited by the same assessor after four months, ensuring that no major changes in management routines or housing conditions had been made during this interval. Analysing the variation in the indicators' prevalence, an overall consistency of results was evident.

This study contributes to the evaluation of the most prevalent welfare problems affecting different dairy goat farming realities in Portugal. As far as the authors are aware, there is no previous research in general overview of farmed dairy goats' welfare according to farm size.

10.1 INTRODUCTION

For the past 20 years a growing interest in goat milk and goat milk products took place all over the world, with the most organized programs for selection, processing and commercialization of goat milk being situated in Europe (Dubeuf et al., 2004; Morgan et al., 2003). Furthermore, dairy goat farming has proved to be of paramount importance to the economies of the Mediterranean countries (Pirisi et al., 2007), where goat milk is traditionally consumed raw or as handmade cheese (Boyazoglu et al., 2005).

Although more accurate statistics referring specifically to dairy goat herd sizes are not available, total goat population in the Mediterranean basis, namely in countries like Portugal, Italy, France, Spain and Greece is approximately 0.4, 0.9, 1.3, 2.6 and 4.3 million head, respectively (Food and Agriculture Organization of the United Nations Statistics Division, 2015), presenting an average herd size of 13 (Tiberio & Diniz, 2014), 50 (Vaccari et al., 2009), 72 (Vaccari et al., 2009), 86 (de Rancourt et al., 2006), and 39 (Hadjigeorgiou, 2011) animals, correspondingly. The intensification of agriculture has lead to the predominant use of exotic dairy breeds such as Murciano-Granadina, Saanen, Alpine and Malagaña (Bruno-de-Sousa et al., 2011). However, more statistics are needed to determine the future perspectives regarding the dairy goat populations and their productivity (Aziz, 2010).

In face of an increasing production dimension and intensification and the concurrent consumers' ethical concerns, tools for farm animal welfare assessment are urgently needed. In this study the focus was set on intensively bred adult dairy goats, as this system is becoming increasingly popular in Europe and the threats to goats' welfare are potentially severe although still largely unknown (European Commission, 2011).

Originally, on-farm welfare assessment focused on the evaluation of resources provided to the animal (e.g., Bartussek, 1999; Bracke et al., 2002). More recently, the interest in measuring animal welfare through direct measures has been increasing, as these measures seem more appropriate for evaluating the actual welfare state of the animals. Therefore using animal-based indicators is now predominant following the approach suggested by EFSA (2012) for welfare evaluation. A recent review on animal-based indicators for welfare assessment of dairy goats identified a considerable number of indicators that can be used for such purpose, however it also highlighted that the majority of these indicators need further research in order to be integrated in welfare assessment schemes, particularly on commercial farms (Battini et al., 2014). Anzuino et al. (2010) have also identified the need for studies aiming to assess prevalence of different dairy goat health and welfare indicators that would give a general overview of the welfare status of the species in different countries and farming conditions. However, we only identified two studies that provided a general overview of welfare and that helped identifying the main welfare problems affecting dairy goat farms: one was carried out in 24 dairy goat commercial farms in the United Kingdom (Anzuino et al., 2010), and the other in 30 dairy goat commercial farms in Norway (Muri et al., 2013).

Besides analysing on-farm prevalences, it is essential to evaluate the consistency of welfare assessments over time, and only a few studies have addressed the quality criteria of these measurements (e.g., Plesch et al., 2010; Temple et al., 2013). As stated by Capdeville and Veissier (2001) and Winckler et al. (2003), the consistency of results over time ensures that the results obtained are representative of the longer-term farm situation and not sensitive to small changes in environmental or animals' internal conditions. The aim of this study is to survey the main welfare problems affecting dairy goats on 30 Portuguese farms, by using the animal-based indicators included in a welfare assessment prototype protocol developed by the AWIN project for dairy goats in intensive husbandry systems. Additionally a preliminary assessment of welfare indicators' variation over time was carried out.

10.2 MATERIALS AND METHODS

Farm sample and data collection

This study was conducted in 30 Portuguese intensive dairy goat farms, from January to March 2014. The farms were sampled from the total national dairy goat farms under intensive production system (n=269), in which breeds as Murciano-Granadina, Saanen and Serrana are predominant, according to *Direcção-Geral de Alimentação e Veterinária* (DGAV; personal communication, 2015). Three farm size categories were created: small size farms (50-100 goats; n=75), medium size farms (101-500 goats; n=114) and large size farms (>501 goats=16). From the 269 national farms a convenience sample of 10 farms from each category was drawn. Farm managers were contacted before the farm visits to discuss the visit's objectives, timetable and methods. It was also verified that the day of the visit was a regular day on farm, to avoid events (e.g., veterinary visits) that would disrupt the normal functioning of the routine. Lastly, security and biosecurity issues were discussed to assure that all farm rules were followed.

All farms kept the animals indoors on concrete, soil or grit covered with straw as bedding material. In 23 farms there was also an outdoor grazing or exterior pen, where the goats had the opportunity to exercise. Diet was composed of total mixed ration distributed twice per

day, and goats were milked twice a day. Kids were separated from their mothers after birth. The number of dairy goats on each farm ranged from 50 to 2000 animals, with a mean (±SD) of 292 (±410) goats. The average herd size in the small, medium and large group farms, was 79 (±17), 309 (±74) and 834 (±451) dairy goats, respectively. Overall, detailed individual observations were carried out on 1172 dairy goats and pen-level observations were performed on 2715 animals. The average number of animals in the assessed pens was 113 ± 84 dairy goats and sample size ranged from 30 to 55 animals. On the days of assessment, environment temperatures ranged from 7ºC to 25ºC and relative humidity from 43% to 93%.
In January 2014, a total of six assessors were trained before the farm visits were initiated. The training consisted of a week period of classroom presentations and exercises, followed by practical field assessments. The assessment of each farm was carried out by two of these trained assessors and all data were collected on the same day.

On-farm assessment of animal-based indicators

Twenty-four animal-based indicators, classified in accordance with the four principles and 12 criteria developed by Welfare Quality (Botreau et al., 2007) were assessed. Descriptive criteria used to assess each animal-based indicator are presented in Table 16 (pen-level observations) and Table 17 (individual assessment). When existent, research studies related to the indicators are referred, but most of the information regarding the indicators used may be found in the AWIN welfare assessment protocol for goats (AWIN, 2015).
Pen-level observations began with the recording of the number of goats improperly disbudded, queuing at feeding/drinking place , with poor hair coat condition, oblivion, with signs of thermal stress (either shivering or panting) and kneeling at the feeding rack. Immediately after this assessment, Qualitative Behaviour Assessment (QBA) was conducted. Subsequently, the assessor entered the pen, and human-animal relationship (HAR) tests (latency to the first contact and avoidance distance (AD) tests) were carried out. Finally, observations of individual animals were performed adopting the sampling strategy developed by Welfare Quality for dairy cows (Welfare Quality, 2009), and individual animal-based indicators (e.g., BCS, cleanliness, overgrown claws; see Table 17) were collected. All animal-based indicators included in individual assessment were recorded on the same animals, with both sides (left and right) being considered, and were scored using a binary assessment system (present or absent), except for BCS and knee calluses. After the individual observations, the group was again assessed to identify severe lameness and kneeling.

Table 16. Animal-based indicators used to assess the welfare of dairy goats in the 30 Portuguese farms – Pen-level observations.

Pen-level observations			
Animal-based indicator		Description	Data collection
Improper disbudding (Ajuda & Stilwell, 2014; AWIN, 2015)	Score 1	Presence of residual horns (scurs) on the head of adult goats that have been disbudded when kids.	The number of goats presenting these conditions is recorded.
Queuing (AWIN, 2015)	At feeding	A goat is queuing if it is standing within 0.5 m behind another goat that is feeding/ drinking, with the head oriented towards the feed barrier/ water place.	
	At drinking		
Hair coat condition (AWIN, 2015; Battini et al., 2015)	Poor hair coat	The hair coat is matted, rough, scurfy, uneven, shaggy hair coat, frequently longer than normal.	
Oblivion (AWIN, 2015)		An oblivious goat seems physically or/and mental isolated comparing to the rest of the group, frequently facing the wall or other parts of the housing structure, sometimes with ears down.	
Shivering score (Battini et al., 2015a)	Score 1	The hair is bristling on the back; the goat has a thick coat.	
	Score 2	The goat is shivering and may take a posture with arched back and head down.	
Panting Score (Battini et al., 2015a)	Score 1	Elevated respiration: from slightly to moderate panting with mouth closed.	
	Score 2	Panting: from heavy to severe open-mouthed panting.	
Kneeling at the feeding rack/ in the pen (AWIN, 2015)		Goats with the front legs flexed, and the rear up compared to the other goats, at the feeding rack or in the pen.	
Severe lameness (Ajuda et al., 2014; AWIN, 2015)	Score 1	Severely lame goats.	

Table 16. (continuation)

Qualitative Behaviour Assessment (QBA) (AWIN, 2015)	QBA relies on the ability of humans to integrate perceived details of behaviour, posture, and context into descriptions of an animal's style of behaving, or "body language", using descriptors such as relaxed, content, frustrated or curious, which have an emotional connotation, and provide information that is directly relevant to animal welfare and may be a useful addition to information obtained from quantitative indicators.	Goats are observed from outside the pen and the assessment is conducted on the whole pen, by selecting the suitable observation points and, consequently, the timing of the observations, which may last from 10 to 20 minutes. At the end of the observation period, a list of 13 qualitative descriptors (e.g., relaxed, content, curious) that reflect the animal's emotional state is rated using the visual analogue scale (VAS).
Latency to the first contact test (AWIN, 2015)	These group assessment tests are applied to evaluate the quality of human-goat relationship: gently handled animals are more inclined to approach people, and consequently suffer less stress related to handling procedures.	Recording of the latency in seconds of the first goat that enters in contact with the test person, who stand immobile into the pen (max. 300 sec).
Avoidance distance test (Mattiello et al., 2010)		Assessment of the number of goats that can be contacted ('contact') or that accept to be gently stroked ('acceptance').

Table 17. Animal-based indicators used to assess the welfare of dairy goats in the 30 Portuguese farms - Individual assessment.

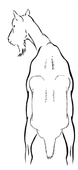
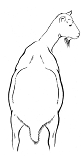
Individual Assessment			
Animal-based indicator		Description	Data collection
Body Condition Score (BCS) (AWIN, 2015; Vieira et al., in press)	Very thin		Binary assessment system (Presence/ Absence).
	Very Fat		
Udder asymmetry (Ajuda & Stilwell, 2014)	Score 1	Asymmetric udder: one half is at least 25% longer than the other (excluding the teats).	
Cleanliness	Score 1	Wet hair, separate or continuous plaques of dirt on the hindquarters, lower legs and udder.	

Table 17. (continuation)

Lesions	Score 1	Presence of lesions/swellings (skin damage with/without hair loss) on the hindquarters, lower legs body, neck and head.
Knee calluses	Score 1	Skin damage with/without hair loss and reddened skin on the knee (carpus), but no enlargement of any joint.
	Score 2	Skin damage with hair loss on the knee (carpus) and enlargement of at least one joint; a thick callus is present over one or both knees.
Abscesses (AWIN, 2015)	Score 1	Presence of abscesses.
Overgrown claws (Ajuda et al., 2014; AWIN, 2015)	Score 1	With severe claw overgrowth.
Discharges (AWIN, 2015)	Score 1	Presence of discharge from eyes, nostrils or vulva.
Faecal soiling (AWIN, 2015)	Score 1	Presence of faeces below the tail head.

Consistency over time of animal-based indicators

To investigate the indicators' consistency over time (COT) 10 of the 30 farms were revisited in July 2014 (three small farms, three medium farms and four large farms), by the same assessors who executed the first assessments. All these farm revisits followed the methods described above. An average of 3.7 ± 1.0 months (SD≈30 days) passed between the two visits and no significant alterations, in management and housing conditions, were implemented during this period. The number of adult dairy goats on each farm ranged from 46 to 2000 animals, with a mean (±SD) of 512 (±613) adult dairy goats. Group assessment was accomplished in 1116 animals, and individual observations were made in 494 adult dairy goats. The mean number of animals in the evaluated pen was 153 ± 95 animals, and the individual animals sampled varied from 32 to 61. In the course of these days, temperatures ranged from 15ºC to 26ºC and relative humidity from 42% to 86%.

Data management and statistical analysis

Data were entered, compiled and statistically analysed using SPSS v22 (IBM® SPSS® Statistics, NY, USA). To perform an overall analysis, the prevalence of each indicator was calculated at farm category level and categorised into "most prevalent" and "less prevalent", according to a 5% warning threshold established, as defined by Welfare Quality (2009). For most of the indicators, prevalence was expressed as the proportion of animals/farms affected on the total of animals/farms assessed. Regarding particular indicators, queuing at feeding and drinking indicators were also recorded as the proportion of animals queuing/farms presenting queuing animals on the total of animals/farms assessed. Similarly, latency to the first contact and avoidance distance tests were expressed in seconds and as the proportion of contacts and/or acceptances, respectively.

For the purpose of QBA data analysis, for each of the descriptors, the distance from minimum to where the assessor ticked the VAS scale was measured in mm. The 13 QBA descriptors' values were used as variables and submitted to Principal Component Analysis (PCA) using a correlation matrix with no rotation. Two Principal Components (PCs) were extracted. A loading plot was produced in order to explore the relationships among variables on the first two PCs and a score plot was also generated in order visualize the position of each farm (classed according to farm size) on the first two PCs.

A Chi-square test of independence was used to examine the relation between all animal--based indicators' results per farm and farm categories, at the 0.05-level.

To preliminarily evaluate the consistency over time of the indicators, a Wilcoxon signed rank test was performed to verify whether the prevalences obtained during the two visits were significantly different at the 0.05-level, as performed by Temple et al. (2013) in pig farms. Also a brief analysis of the variation in indicators' prevalence between visits was made.

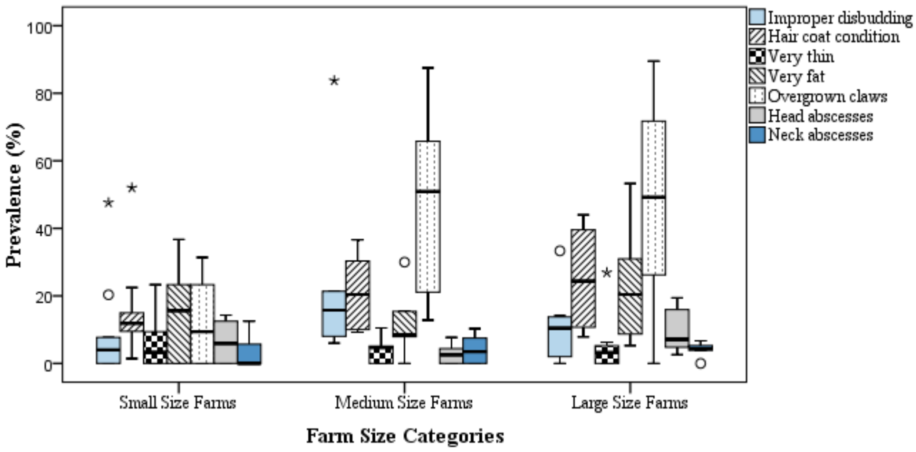
10.3 RESULTS

As previously mentioned, a warning threshold analysis was initially performed considering the most and less prevalent indicators. Subsequently, behavioural indicators were analysed. Finally, the different indicators' assessment was considered for their consistency over time.

Warning threshold analysis

A preliminary analysis of the collected data showed that there was no statistical difference for lesions, abscesses and cleanliness scored between the right and left side of the animal. Therefore, the prevalence of 'Abscesses', 'Lesions' and 'Cleanliness' presented are always from the animal's left side. Some indicators were very common, exceeding the warning threshold of 5%, whereas others were seldom recorded ("Below the warning threshold"), and are presented separately. The Chi-square test of independence showed that the relation between the prevalence of each animal-based indicator and farm category was not significant at the 0.05-level. Table 18 shows the overall most prevalent indicators and Table 19 presents the results of queuing at feeding and at drinking. The prevalence's variation of some of these indicators among farm categories is presented as a box-plot in Figure 11.

Figure 11. Variation of some of the most prevalent indicators among the 30 Portuguese commercial intensive dairy goat farms visited, divided by farm category. Points (o) and asterisks (*; extreme values) represent outliers.



All indicators categorised as less prevalent presented similar values across the three farm categories, except for 'Panting score 1', 'Shivering score 1', and 'Faecal soiling', as shown in Table 20. Indicators as Panting score 2', 'Shivering score 2' and 'Kneeling at feeding rack' presented zero prevalences. The prevalences of some of these indicators are given in Figure 12 as a box-plot.

Figure 12. Prevalence's variation of some of the less prevalent indicators among the 30 Portuguese commercial intensive dairy goat farms visited, divided by farm category. Points (o) and asterisks (*; extreme values) represent outliers.

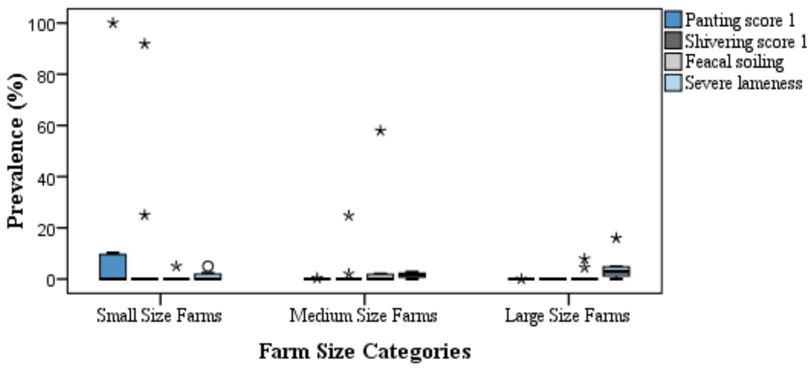


Table 18. Most prevalent indicators of the 30 Portuguese commercial intensive dairy goat farms visited : each indicator's prevalence is organized according to farm category.

Animal-based Indicator		N	Goats (%)	Farm Size Category (%)								
				Small			Medium			Large		
				N	Farms (%)	Goats (%)	N	Farms (%)	Goats (%)	N	Farms (%)	Goats (%)
Pen-level observations												
Improper Disbudding*		1778	256 (14.4)	453	6 (60)	41 (9.1)	557	6 (60)	130 (23.3)	768	7 (70)	85 (11.1)
Hair coat condition		2715	508 (18.7)	473	10 (100)	84 (17.8)	1122	10 (100)	197 (17.6)	1120	10 (100)	227 (20.3)
Individual Assessment												
BCS	Very thin	1172	60 (5.1)	305	5 (50)	16 (5.3)	407	7 (70)	19 (4.7)	460	5 (50)	25 (5.4)
	Very Fat		220 (18.8)		6 (60)	40 (13.1)		9 (90)	67 (16.5)		10 (100)	113 (24.6)
Cleanliness	Hindquarters	1172	229 (19.5)	305	8 (80)	53 (17.4)	407	5 (50)	38 (9.3)	460	8 (80)	138 (30.0)
	Lower legs		210 (17.9)		6 (60)	50 (16.4)		7 (70)	67 (16.5)		8 (80)	93 (20.2)
Lesions	Hindquarters	1172	124 (10.6)	305	4 (40)	16 (5.2)	407	7 (70)	46 (11.3)	460	7 (70)	62 (13.5)
	Lower legs		112 (9.6)		2 (20)	2 (0.7)		6 (60)	35 (8.6)		7 (70)	75 (16.3)
	Head		299 (25.5)		9 (90)	34 (11.1)	407	10 (100)	125 (30.7)		10 (100)	140 (30.4)
	Body		183 (15.6)		7 (70)	42 (13.8)		7 (70)	71 (17.4)		10 (100)	70 (15.2)
	Neck	1172	169 (14.4)	305	5 (50)	14 (4.6)	407	8 (80)	76 (18.7)	460	9 (90)	79 (17.2)
Udder asymmetry		1172	68 (5.8)	305	7 (70)	15 (4.9)	407	9 (90)	23 (5.7)	460	8 (80)	30 (6.5)
Knee calluses	Score 1	1172	968 (82.6)	305	10 (100)	286 (93.8)	407	10 (100)	321 (78.9)	460	10 (100)	361 (78.5)
	Score 2		135 (11.5)		4 (40)	15 (4.9)		6 (60)	44 (10.8)		7 (70)	76 (16.5)

Table 18. (continuation)

Overgrown claws		1172	402 (34.3)	305	7 (70)	36 (11.8)	407	10 (100)	155 (38.1)	460	10 (100)	211 (41.9)
Abscesses	Head	1172	91 (7.8)	305	5 (50)	18 (5.9)	407	7 (70)	14 (3.4)	460	10 (100)	59 (12.8)
	Body		126 (10.8)		5 (50)	23 (7.5)		7 (70)	38 (9.3)		8 (80)	65 (14.1)
Discharge	Nasal	1172	62 (5.3)	305	4 (40)	7 (2.3)	407	5 (50)	33 (8.1)	460	6 (60)	22 (4.8)
	Ocular		67 (5.7)		3 (30)	4 (1.3)		6 (60)	24 (5.9)		7 (70)	39 (8.5)

Table 19. Results of queuing at feeding and at drinking of the 30 Portuguese commercial intensive dairy goat farms according to farm category

Animal-based Indicator		N	Goats (%)	Farm Size Category (%)								
				Small			Medium			Large		
				N	Farms (%)	Goats (%)	N	Farms (%)	Goats (%)	N	Farms (%)	Goats (%)
Pen-level observations												
Queuing	At feeding	2715	721 (26.6)	473	8 (80)	131 (27.7)	1122	8 (80)	145 (12.9)	1120	9 (90)	445 (39.7)
	At drinking		130 (4.8)		0 (0)	0 (0)		7 (70)	67 (6.0)		4 (40)	63 (5.6)

Human-animal relation tests

The 'Latency to the first contact' time was lower in large farms (81.1 ± 117s; range from 0 to 300 s) than in small (139.7 ± 139s; range from 10 to 300 s) and medium farms (156.9 ±131.4s; range from 7 to 300 s). The percentage of goats that showed 'Acceptance' and 'Contact' during the Avoidance Distance test was below 2.2% in the three farm categories, with large farms having higher values (2.1%).

Qualitative Behaviour Assessment

The first two PCs explained 27.6% and 17.7% of the total variance. A loading plot of the QBA descriptors is shown in Figure 13.

Figure 13. Word chart of the QBA assessed in the 30 dairy goat farms. This 2-dimensional loading plot shows the relationship among the 13 QBA descriptors representing dairy goat behaviour on the two principal components.

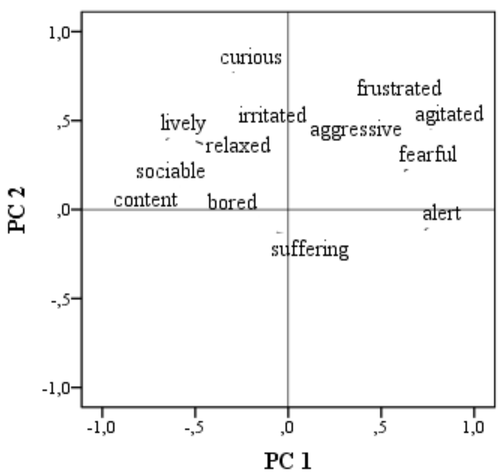
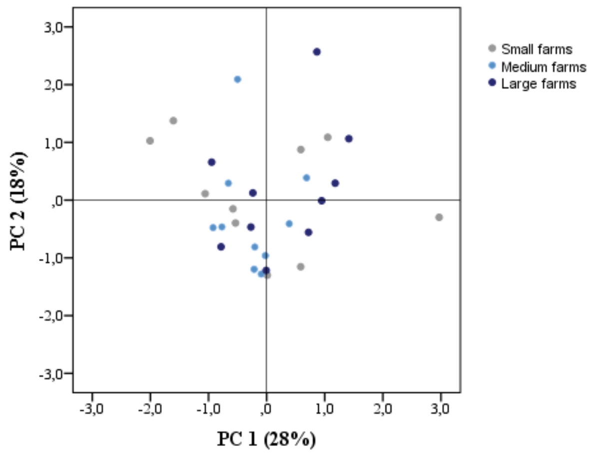


Table 20. Less prevalent indicators of the 30 Portuguese commercial intensive dairy goat farms visited: each indicator's prevalence is organized according to farm category.

Animal-based Indicator		N	Goats (%)	Farm Size Categories								
				Small			Medium			Large		
				N	Farms (%)	Goats (%)	N	Farms (%)	Goats (%)	N	Farms (%)	Goats (%)
Pen-level observations												
Oblivion		2715	13 (0.5)	473	1 (10)	2 (0.4)	1122	4 (40)	6 (0.5)	1120	3 (30)	5 (0.4)
Panting (score > 0)	Score 1	2715	69 (2.5)	473	3 (30)	66 (14.0)	1122	1 (10)	1 (0.1)	1120	1 (10)	2 (0.2)
	Score 2		0 (0)		0 (0)	0 (0)		0 (0)	0 (0)		0 (0)	0 (0)
Shivering (score > 0)	Score 1	2715	74 (2.7)	473	2 (20)	50 (10.6)	1122	2 (20)	24 (2.1)	1120	0 (0)	0 (0)
	Score 2		0 (0)		0 (0)	0 (0)		0 (0)	0 (0)		0 (0)	0 (0)
Severe lameness		2715	48 (1.8)	473	3 (30)	3 (0.6)	1122	8 (80)	15 (1.3)	1120	9 (90)	30 (2.7)
Kneeling	Feeding rack	2715	0 (0)	473	0 (0)	0 (0)	1122	0 (0)	0 (0)	1120	0 (0)	0 (0)
	Pen		13 (0.5)		1 (10)	1 (0.2)		5 (50)	5 (0.5)		2 (20)	7 (0.6)
Individual Assessment												
Cleanliness	Udder	1172	37 (3.2)	305	3 (30)	5 (1.6)	407	3 (30)	12 (2.9)	460	3 (30)	20 (4.3)
Abscesses	Udder	1172	29 (2.5)	305	3 (30)	4 (1.3)	407	5 (50)	12 (2.9)	460	6 (60)	13 (2.8)
	Neck		48 (4.1)		5 (50)	9 (3.0)		5 (50)	10 (2.5)		10 (100)	29 (6.3)
	Hindquarters		7 (0.6)		1 (10)	2 (0.7)		4 (40)	1 (0.2)		4 (40)	4 (0.9)
Discharge	Vulvar	1172	3 (0,3)	305	0 (0)	0 (0)	407	2 (20)	2 (0.5)	460	1 (10)	1 (0.2)
Faecal soiling		1172	31 (2.6)	305	1 (10)	1 (0.3)	407	3 (30)	24 (5.9)	460	2 (20)	6 (1.3)

Descriptors such as agitated (0.76), alert (0.72), aggressive (0.63), fearful (0.61), as well as lively (-0.67), sociable (-0.65), content (-0.58) and relaxed (-0.52), presented the highest weights on the first PC. The second PC was characterised by descriptors such as curious (0.75), frustrated (0.68), irritated (0.63). The position of each farm, classed by farm size, was scattered in a score plot (Figure 14), showing a homogenous overall distribution of farms throughout the two PCs, regardless of farm size.

Figure 14. Score plot of farms of different size categories on the first two PCs, based on QBA descriptors of dairy goats' behaviour.



Consistency over time of animal-based indicators

Indicators such as 'BCS - very fat', 'Faecal soiling', 'Cleanliness – hindquarters', 'Knee calluses - score 2', 'Head lesions', 'Ocular discharge' and 'Overgrown claws', showed a change in prevalences above 5%. However, according to the Wilcoxon signed rank test, only 'Head lesions' differed significantly between the two visits (P=0.037). Indicators as 'Improper disbudding', 'Oblivion', 'BCS – very thin', 'Hindquarters abscesses', 'Neck abscesses', 'Vulvar discharge', 'Severe lameness' and kneeling, both in the pen and at the feeding rack, presented a prevalence variation below 1%. 'Panting score 2', 'Shivering score 2' prevalences demonstrated no variation between visits, since there were no cases recorded in either visits.

10.4 DISCUSSION

As mentioned by several authors (Castel et al., 2010; de Rancourt et al., 2006; Dubeuf, 2005), European goat production is an important economic, environmental and sociological activity in Mediterranean countries. This study presents the first welfare assessment of dairy goats in the Mediterranean region, incorporating only animal-based indicators collected through pen-level observations and individual assessment of animals. All observations were performed on 30 commercial intensive dairy goat farms, allowing for comparisons with the studies of Anzuino et al. (2010) and Muri et al. (2013), which have published empirical data from overall welfare assessment of this species.

Warning threshold analysis

Claw overgrowth should be considered a major problem in Portuguese farms, due to the high prevalence detected. Although it was identified in large and medium farms at a higher prevalence (38.1-41.9%) compared with smaller farms (11.9%), it is a cause for general concern. This result is in agreement with the studies performed in British and Norwegian dairy goat farms, where severe claw overgrowth reached a prevalence of 80% and 66%, respectively (Anzuino et al., 2010; Muri et al., 2013). According to Smith and Sherman (2009), this is probably due to a lack of claw wear when animals are housed on straw bedding. Our results are also probably related to differences in access to outdoor grazing; in fact, in nine of the 10 small farms visited, goats had access to pasture during nine months of the year, while in large and medium farms this only happened in one and two farms, respectively, and for a period of time under two months. It is important to differentiate access to pasture from access to an exterior pen, which was also largely found in medium and large farms, since access to a small exterior pen does not necessarily imply sufficient claw wear. Although claw trimming was performed at different times before the visits, it was noticeable that in large farms a low human: animal ratio corresponded with less time to observe and treat or trim individual animals (Stafford & Gregory, 2008). This was confirmed in the present study, with small farms presenting a higher human: animal ratio (0.03) than large farms (0.004).

In both Anzuino et al. (2010) and Muri et al. (2013) studies, nearly all animals had knee calluses, which is in agreement with the prevalence found in our study (82.6%). As mentioned by Smith and Sherman (2009a), goats kept in confinement, especially on hard surfaces, can develop a marked thickening of the skin over the carpi, hocks, and sternum. However, 'Knee calluses -score 2' presented a higher prevalence (16.5%) in large farms, being significantly correlated with claw overgrowth (data not shown).

In British dairy goat farms, Anzuino et al. (2010) found a 3% prevalence of very thin animals, which is close to the prevalence found in our farms (5.1%), with no differences due to farm size. Compared with our results (18.8%), a much lower prevalence of very fat animals (3%) was reported in the British study. The inclusion of BCS in on-farm welfare assessment schemes allows for the identification of the endpoints of a scale (animals that are too thin or too fat), as these are the ones more likely related with welfare problems. Obesity is usually associated with higher predisposition for metabolic diseases, and emaciation may be either a sign (e.g., chronic disease such as paratuberculosis) or a cause (e.g., pregnancy toxaemia) for welfare problems (Smith & Sherman, 2009a).

Muri et al. (2013) reported an 18% prevalence of dirty hindquarters being also in line with our overall findings (19.5%). However, in Portugal, large farms presented higher prevalences (30%). Animal cleanliness is often used as a welfare indicator in several species, as pigs (Otten, 2013; Scott et al., 2007), poultry (de Jong et al., 2014; Sans et al., 2014), and cattle (Ellis et al., 2007; Whay et al., 2003b), and may provide information not only on animal comfort but also on stockpeople's attitudes and care for animals, as supported by De Rosa et al. (2009). In goats, cleanliness depends mostly on how often bedding is replaced or added, as animals are housed on straw bedding all year round. However, cleanliness assessment can be challenging in some breeds, being easier to assess in white breeds such as Saanen. This fact may account, to some extent, for the higher prevalences of hindquarters dirtiness found in large farms, as Saanen was the most common breed in this farm category.

Hair coat condition is an indicator recently developed by Battini et al. (2015b) and it reflects not only goats' nutritional status but also their health status. This indicator was found at a prevalence of around 20% in all farm categories, functioning as a first warning on goats' nutritional and health status.

Presence of queuing animals at the feeding rack showed a higher prevalence on large farms (39.7%) with high stocking densities corresponding to a high goats/feed space ratio. The natural synchronous behaviour of goats increases the probability of finding this indicator and its effect on welfare.

In our study we found an overall prevalence of 1.8% of severe lameness, which is in line with Anzuino et al. (2010) and Muri et al. (2013) studies, presenting 3% and 2% of severe lameness, respectively. Herd size can be considered a risk factor for lameness, as mentioned in several studies (Alban, 1995; Katsoulos & Christodoulopoulos, 2009). Individual observation and care for the animals may be more difficult in large herds than in smaller ones, which can lead to a lower detection of lame animals, and can explain the higher prevalence values observed in large farms. In terms of overall welfare assessment, the identification of only the severely lame animals may provide sufficient information, however, the reason why across studies this is the only information collected is because in goats, lameness assessment is often performed while the animals are housed in their pens with soft straw surfaces, which tends to hide the mild cases (Anzuino et al., 2010). Hence, the recording of the severe cases of lameness is the only feasible method of assessing this indicator; moreover, there are some studies (Green et al., 2002; Gundelach et al., 2013; Main et al., 2010) that show that prevalences of severe and mild cases are correlated, therefore delivering an accurate picture of the real welfare state of the animals.

Other welfare indicators with similar significance are the presence of faecal soiling and dirty udders. In the study of Muri et al. (2013) a 1% prevalence of faecal soiling was reported. In our study a 2.6% prevalence of faecal soiling was obtained, being this value in both studies below the warning threshold considered. However, medium farms did show a higher prevalence (5.9%). The presence of faecal soiling can be a sign of disease or nutrition errors, and generally has a direct effect on the animals' cleanliness. Regarding dirty udders, Anzuino et al. (2010) and Muri et al. (2013) found a 1% prevalence of very dirty udders, being also very much alike to our findings (3.2%). Schreiner and Ruegg (2003) found that cows with dirtier udders, teats and hindquarters had a higher prevalence of intramammary infection, as this may also occur with dairy goats.

Regarding the Chi-square test of independence performed, the fact that all farms were under an intensive production system with similar management, can partly account for the lack of significant differences at the 0.05-level between the animal-based indicators and the farm categories considered. However, the prevalence of most of the more common welfare indicators, i.e. exceeding the warning threshold (5%), was affected by farm size, being usually higher in larger farms. Although there was no significant differences, some of the indicators such as 'Knee calluses – score 1', 'Neck lesions', 'Head Abscesses', 'Neck Abscesses', Queuing at drinking' and 'BCS – Very fat' were very close to the 0.05 significance level (data not presented). A larger sample size might lead to different results, as small population effects can be highly significant if the sample is large enough, which points toward the need for further studies on the on-farm prevalence of these indicators.

Human-animal relation tests

Human-animal relation was generally better in large farms than in small farms, especially considering the latency period to the first contact between goat and assessor. This was unexpected, as HAR is usually better in small farms, where the relationship between the stockperson and the animals is very strict (Mattiello et al., 2010). However, our results can probably be explained by the breed differences among farms. Breeds such as Saanen and Murciano-Granadina, that are very common in large-size farms in Portugal, are reported to be docile and easier to handle, being more suited for intensive systems (Escareño et al., 2013; Martínez et al., 2010; Sinn & Rudenberg, 2008). A higher variety of breeds was found in small farms – Serrana, Alpine, Malagaña, Murciano-Granadina, Saanen, Florida, Charnequeira and crossbreds (Saanen with Alpine). Based on personal experience and observers' assessments, breeds such as Saanen and Murciano-Granadina accept being touched or even gently stroked more often, sometimes even complicating the assessment by grouping around the assessor; while breeds, such as Serrana, showed strong avoidance behaviour, making it difficult to carry out the test in a standardized way. Muri et al. (2013) also referred similar limitations when testing Norwegian dairy goats. Additionally, in small farms, goats had more access to outdoor grazing enhancing the expression of natural, foraging and exploratory behaviours. In a study performed by Battini et al., (2011), it was observed that the avoidance distance of dairy cows after the grazing period was significantly higher than both before and during the grazing period, reflecting the animals' lower level of confidence with humans and also supporting our findings.

Qualitative Behaviour Assessment

A PCA analysis revealed two dimensions of goat behaviour PC1 and PC2. PC1 of the QBA, which carries most of the relevant variance, allows for the differentiation between farms with animals that appeared to be in a more positive mood from farms that presented animals with a more negative mood. Agitated, alert, aggressive, fearful, as well as lively, sociable, content and relaxed showed the highest weights on PC1. PC2 was defined by descriptors as curious, frustrated and irritated, presenting a more difficult interpretation. The homogenous overall distribution of farms throughout the two axes might be supported by housing and management having a real effect on the animals' on-going behaviour, and these farms were only selected regarding their herd size, with all the animals being bred under an intensive production system. These particular results might be explained by the observers' moderate training, as a crucial requisite for applying QBA is adequate training, in which the observers should discuss the meaning of each QBA descriptor or watch video clips representing each descriptor, so as to standardize the evaluations (Napolitano et al., 2015).

Consistency over time of animal-based indicators

The consistency of indicators over time allows identifying real welfare issues that persistently continue on-farm. However, this was a preliminary analysis of the consistency of these animal-based indicators over time and further studies are required to robustly establish the significance of these variations across time. Analysing the variation between the two visits, it was possible to observe that only head lesions presented significantly different results, which might also be explained by the training intensity and the break period between visits. In a study performed by Gibbons et al. (2012), a five-day break in a training programme, to train observers to score injuries on dairy cows, resulted in decreased agreement for all injury scores, improving again in the next day after practice. This highlights the importance of continual practice in order to "recalibrate" the observers to a reference standard, as defended by (EFSA, 2012). An overall consistency of results was apparent, with common findings such as improper disbudding or presence of external abscesses remaining common and those conditions that occurred less often, such as oblivious animals, remaining at low levels of prevalence.

Constraints and future perspectives

According to Willeberg (1991), preventing disease is a major animal welfare topic, as describing disease occurrence and its consequences are parts of qualifying and quantifying animal welfare problems in herds, being key aspects in welfare research and practice. Most of the disease issues are not only production limiting events, but also welfare problems, having however become inherent pieces of the intensive animal production for which there are no easy solutions. Nevertheless, little epidemiological literature concerning such welfare considerations has been published. Although the goat sector is growing, there has been less research on goats than on other production species, especially regarding welfare aspects (Anzuino et al., 2010; Sahlu & Goetsch, 2005): this makes the development of welfare assessment protocols for this species a much more difficult task. The prevalence of the welfare indicators in dairy goat farms provides information on the general health and welfare status of the farms. Knowing the general prevalence of the indicators in the goat industry is paramount to be able to set thresholds of acceptability. These thresholds will allow farmers, vets and other technicians to identify main welfare issues in their own farms, and therefore set plans of action to improve the general welfare/health condition. Moreover, setting these thresholds might be very important for law implementation in animal welfare, when considering the use of animal-based indicators. In this sense, this study has contributed to the knowledge of the reality of the Mediterranean region. On the other hand, prevalence analysis can also provide data for benchmarking purposes. Recently, benchmarking has been used as an approach for helping farmers manage the welfare of their animals (e.g., planning programs on organic farms in Europe, dairy cow comfort and road transport practices in North America; Colditz et al., 2014). This approach proved to be educational, informative and empowered better management, allowing the farmers to determine how well or badly they were performing in relation to others (Blokhuis et al., 2013). However, further studies in different countries and different management systems are needed, as all the EU countries work under the same legislation and are therefore obliged to the same rules.

10.5 CONCLUSIONS

The present study is the first welfare assessment of dairy goats in Portugal, incorporating pen-level observations and individual assessment of the animals, and analysing indicators consistency over time. It is a first step to give an overview of the reality of the Mediterranean countries, however, further studies in different countries and different management systems are required. The main areas of concern were claw overgrowth, queuing at feeding, very fat animals, poor hair coat condition and improper disbudding. Some of the assessed indicators presented similar prevalences to indicators included in previous studies (Anzuino et al., 2010 in UK; Muri et al., 2013 in Norway), suggesting common problems in different countries.

The analysis of the variation in the indicators' prevalence between different seasons revealed an overall consistency of the indicators. This study contributes to an increased awareness of the main welfare issues affecting Portuguese intensively kept dairy goats. These findings can help to define intervention thresholds and guidelines for each welfare indicator, in order to improve the on-farm general welfare conditions.

PART III

General discussion

Following the work developed in this thesis, Part III is dedicated to the general discussion and conclusions, which will then lead to the proposal of some future research.

CHAPTER 11

Discussion and conclusions

Chapter 11 focuses on the main theoretical and methodological findings of this research, and highlights the practical implications of the different studies. Moreover, it discusses some of the limitations of this research work.

Knierim & Winckler (2009) have identified two major challenges in on-farm welfare assessment: an appropriate selection of welfare indicators and associated level of measurement; and, the identification of strategies to overcome the practical constraints associated with the indicators' integration on welfare protocols. The present research thesis provides a scientific contribute to both these challenges. Furthermore, knowledge generated by this thesis is particularly relevant given that EFSA is currently clarifying issues on animal--based indicators proprieties to ensure that they are fit for the purpose for overall welfare assessment.

The overall aim of this thesis was to develop and integrate animal-based welfare indicators, including pain, in goat farms. In order to accomplish this purpose, three main objectives were set. In summary, key contributions of this thesis consist in:

1. The identification of animal-based welfare indicators with potential to be included in welfare assessment protocols (chapter six);
2. The development of body condition scoring (chapter seven) and lameness (chapter eight and nine) indicators, based on the assessment of their psychometric properties and the level of measurement needed;
3. The integration and testing of animal-based indicators in welfare assessment protocols (chapter ten).

Adding to the discussion and main conclusions presented in the chapters within Part II, a general discussion for making a comprehensive analysis of the theoretical, methodological, and practical implications of this work is due. After this exercise future study perspectives can be highlighted.

11.1 MAIN THEORETICAL AND METHODOLOGICAL RESULTS

Identification of the most promising animal-based welfare indicators for dairy goat on-farm welfare assessment

The identification of animal-based indicators is of paramount importance within the context of the EU Strategy for the Protection and Welfare of Animals 2012-2015 (European Commission, 2012), as it adds value to the policy of introducing science-based indicators to the current legal framework. However, up to this date, the information regarding animal--based welfare indicators, particularly considering dairy goats, is scarce (Caroprese et al., 2009; Johnsen et al., 2001), and focuses on very particular aspects of animal welfare.

The extensive and comprehensive literature review conducted in chapter six allowed us to identify the most promising animal-based welfare indicators for dairy goats' welfare assessment. The fact that the review was made within the frame of the four principles and 12 criteria developed by the Welfare Quality project allowed us to deliver a list of indicators in line with the current concept of animal welfare. Moreover, given that it focused exclusively on indicators that could be used on-farm (excluding indicators that were too time consuming to collect or that required specific instruments or laboratory analysis or that were recorded on other locations that not the farm), the review supplied the foundations for the development of an on-farm welfare assessment protocol for dairy goats.

For example, the literature review identified BCS as one of the most promising indicators for the assessment of the "good feeding" principle, but also with application on the "good health" principle. For the "good housing" principle, indicators related to thermal comfort seemed promising. As to the "good health" several indicators (e.g., lameness, claw overgrowth) were found, probably due to their attributed economic relevance. As to "appropriate behaviour" principle a few promising indicators such as agonistic behaviour and latency to first contact were also found. Overall, only one criterion within the principle "expression of other behaviours" remained without indicators. Additionally, it is worth to mention that some indicators, such as BCS, hair coat condition and queuing (both at feeding and drinking) can be used to evaluate different criteria, providing information about several welfare aspects.

Considering the importance of choosing scientifically valid welfare indicators, special attention was given to the assessment of the indicator's validity, reliability, and feasibility. As a result, the review showed that the great majority of indicators required further adjustment or research, particularly at the reliability and feasibility level. Furthermore, most studies surveyed were designed and tested under experimental conditions; hence, to enhance the indicators' feasibility, the review suggested that further studies should be performed on commercial farms. The accomplishment of this first objective allowed us to identify which indicators should merit further scientific attention in the context of this thesis, and which should be listed for future research.

Development of the lameness and body condition scoring indicators for dairy goat welfare assessment

As said, BCS and lameness were considered promising welfare indicators for dairy goat welfare assessment, however they both required further adjustment, by means of the development of new scoring methods, to be included in a practical on-farm welfare assessment protocol. Streiner & Norman (2008) discuss in the beginning of their book on health measurement scales how one of the characteristics of health sciences literature is the enormous amount of available scales, and they show some concern on how it seems that every article published uses a different approach to assess a given indicator, with the concurrent difficulty in comparing different studies. Anzuino et al. (2010) also shared this concern while discussing how different scoring definitions are used across welfare studies. Similarly, Churchill (1979) in a different field of knowledge, marketing, has stated: "more stupefying than the sheer number of our measures, is the ease with which they are proposed, and the uncritical manner in which they are accepted" (Churchill, 1979, p.64). Therefore, before developing the new scoring systems, attention was given to all the available scoring systems for BCS and lameness assessment, exploring the need for the development of other scoring systems[14].

Regarding BCS, as it is extensively discussed on chapter seven, we confirmed the need to develop an exclusively visual scoring system for BCS assessment that would allow the identification of the animals in extreme nutritional conditions (i.e., too thin or too fat), without the need to restrain them. As for lameness, the review highlighted that the definition of clinical lameness was unclear, with no established scoring systems being available. Particularly, the review identified sensitivity problems associated with the existing scoring systems, and suggested the development of a modified VAS. While most of the experimental work in this thesis is related to scale building, the two experimental settings pursuit different outcomes: for BCS the objective was to develop a more simplified scoring system, whereas for lameness the need was for higher measure accuracy. Moreover, the available scoring systems were shown not to meet the standard measurement criteria of validity, reliability and feasibility.

As Streiner & Norman (2008) underline, the effort to develop another scoring system should only take place if the new system is either cheaper or simpler, or aims to provide a more refined assessment of the construct that is being assessed, or if the available scoring systems does not meet validity, reliability and feasibility criteria. These three psychometric proprieties, which are indistinguishably linked to the methodology of measurement, are the

[14] Information regarding other BCS and lameness scoring systems can be found, respectively, in chapters seven and nine.

focus of this discussion and one of extreme importance regarding on-farm welfare assessment.

Body condition scoring

The work developed in this thesis was based in the understanding that animal welfare is a multidimensional concept (Botreau et al., 2008), hence needing a multidisciplinary approach[15]. There are several advantages associated with multidisciplinary work, namely the potential gains we can obtain by importing concepts and models from one discipline to the other (Lund, Coleman, Gunnarsson, Appleby & Karkinen, 2006). Animal welfare can be seen as pioneer in interdisciplinary research (Blokhuis et al., 2013; Lund et al., 2006), with several examples of cooperation being successfully explored between natural and social scientists. A good example of multidisciplinary work was the development of “welfare criteria” by the Welfare Quality project (Botreau et al., 2009). More recently, partnerships are being established with scientists from more technological fields such as computer vision. Computer vision is an umbrella term for all tasks pertaining the automated processing of images, or other sources of visual information. The type of problems that computer vision focuses on, goes from low signal processing (e.g., denoising and colour equalization), to more sophisticating levels, where it is expected that an automated system (e.g., a computer or a robot) “understands and reasons” based on visual information. With this later aim, concepts from computer vision have been applied in lameness and lying behaviour detection studies (Pluk et al., 2010; Porto, Arcidiacono, Anguzza & Cascone, 2013; Viazzi et al., 2013, 2014), and in BCS (Azzaro et al., 2011; Bercovich et al., 2013; Van Hertem et al., 2013). So, to develop the BCS indicator, concepts depicted from computer vision were applied in an innovative way to create a visual scoring system[16]. There are no references in literature regarding the application of computer vision concepts for this purpose.

Alternatively to the development of a visual scoring system, an improved scoring system based on verbal descriptors of how observers should assess visually different body regions could have been tried. However, verbal descriptors are difficult to elaborate and even more difficult to understand, being therefore important to reduce the observers dependence on written descriptors (Bana e Costa & Beinat, 2005; Edmonson et al., 1989). Already with this notion in mind, Edmonson et al. (1989) developed a visual scoring system to assess BCS in dairy cows based on diagrams of visual cues of different body regions. The Edmonson et al. (1989) system was developed based on an interactive process that integrated literature review, interviews with experts, field testing, statistical analysis, and reviews from users. However, this methodology presents two potential flaws. The first is that it needs the

[15] On chapter two the multidimensionality of animal welfare is discussed.
[16] All the methodological details are clearly explained in chapter seven.

assessment to be performed considering a total of eight different anatomical locations. The second flaw is that it relies on expert assessment, which can be biased towards their own experience in the field. In goats this might be particularly problematic due to the different types of breeds used in intensive production systems.

The methodology developed in this thesis relied on the identification of only one anatomical location of the goat that would allow for a valid BCS, and subsequently on the development of a system that conveyed the appearance of a representative animal from each category being assessed. This methodology was performed with the belief that after defining a representative image for a given category, the observers would be successful in scoring subsequent sets of new images. The challenge was then to develop the concept of a representative image. To meet the goal, computer vision concepts were applied at two different levels: at a lower level for the manipulation of images with the aim of improving data quality, and at a higher level of analysis that allowed identifying what was a "typical" element of a given BCS category. A further addition to the method was the scientific illustration of the representative images. Scientific illustration is known in anatomical drawing, and was applied here with great success. The systematic approach for the creation of visual scoring systems has the potential to be applied in the building of similar scoring systems for other species, or on building scoring systems for other indicators that rely on visual cues.

Furthermore, in terms of methodology, the identification of the threshold images allowed users to interiorize the border of each category, which helped them to improve their recognition of the representative images, and consequently enhanced reliability. The concept of threshold images can be exported to other indicators training, representing a potential way to improve reliability, a common aim associated with the development and integration of welfare indicators (Brenninkmeyer et al., 2007; Gibbons et al., 2012; Knierim & Winckler, 2009; March et al., 2007).

The methodology here described was developed in parallel with a concurrent validation process that was possible due to the existence and use of an already established method for BCS, the Hervieu & Morand-Fehr (1999) method. As reviewed by Botreau et al. (2007), an overall animal welfare assessment system may only be valid if the indicators that integrate it have established validity. Therefore the experimental design for the development of the BCS indicator was important in terms of scientific acceptability of the new scoring system.

Another important asset of the experimental design was that it did not consider the reliability of the new scoring system to be an absolute propriety, but the result of an interaction between the system and the group of observers (Streiner & Norman, 2008). Therefore, when developing the reliability studies, attention was given to previous experience of the observers, and to the effect of training. This allowed us to identify how previous experience and training affects reliability results, which reinforced previous studies (Ferguson et al., 1994; Kristensen et al., 2006).

Lameness

Regarding the lameness indicator, the objective was to assess the grounds for developing a modified VAS. However, as discussed in chapter nine, before developing this scoring system it was imperative to investigate the distribution of different gait and posture signs along the scales' continuum. The strategy adopted to address this question was two-fold. First, the ability to recognize intensity increase or decrease in lameness signs was confirmed. This was followed by the investigation on how different gait and posture signs were scored individually in a VAS.

These questions could have been answered by selecting an expert in lameness scoring in goats that would provide the information needed. However, as pointed by Streiner (1993) while developing a scoring system it is better to take advantage of the clinical experience of a panel of experts, rather than relying on one person personal experience. Moreover, Streiner (1993) also points out that is important to take into account that there is always the danger that when the panel of experts is chosen it is biased towards a particular occupation and experience, as the panel is usually gathered by invitation, resulting in a limited range of ideas.

Reliability of lameness scoring systems in goats has never ben explored. However, reliability studies on dairy cows and ewes have shown that observers have different aptitudes to discriminate levels, with this being highly dependent on their experience and training (Brenninkmeyer et al., 2007; Kaler, Wassink & Green, 2009; Kaler et al., 2009; March et al., 2007; Nalon et al., 2014; Winckler & Willen, 2001). Although these studies were not performed on goats, there seems to be no reason to assume that these difficulties will not be found on lameness scoring in goats, and should therefore be explored.

For these two main reasons, for our experimental design it was important to collect a large number of lameness scorings. Looking into literature the most frequent way of collecting observers scorings are face-to-face meetings, generally workshop sessions where videos were shown to participants. However, considering the logistic costs associated with the organization of such meetings, the expected low rates of participation and subsequent low numbers of observations collected, it was decided to explore the use of a web-survey to collect observers' lameness scorings[17]. To the best knowledge of the author, this was the first study in which a web-survey was used with the objective of collecting information regarding the assessment of a health indicator in the veterinary field.

One important part of the development and application of the web-survey was the web--survey preparatory stage. The video collection and preparation, and their scoring by experts were an important part of the experimental design. The option to create 82 video clips and

[17] The design and implementation of the web-survey are described in chapter eight.

their random assignment in the web-survey was important to guarantee a diversity of examples for scoring, instead of only relying on a few cases.

To guarantee a large amount of answers, besides a good distribution of the web-survey through different associations and organizations, special attention was given to the building of an attractive platform design (e.g., intuitive graphical interface, help-buttons, photos that illustrated concepts), and provision of clear explanations of what was intended (e.g., definition of the scoring system and lameness signs). Still considering the platform technological advantages, the possibility to automatically measure and store the respondent's assessment allowed the research team to save a lot of time, and was also a more trustworthy approach as recording mistakes were prevented.

The web-survey was online for approximately five months and had 2,312 views that resulted in 600 participations, hence demonstrating a very good diffusion and showing how this topic elicited interest and curiosity. Still, the difference between views and participations should be further debated. Given that there is no data to properly discuss this, it can only be speculated that this difference was due to the species in assessment – when participants saw the goat videos, and not having a lot of experience with the species, chose not to participate. On the other hand it is a strong point for data quality as the participants that actually finished the web-survey showed interest and therefore their assessment was surely performed to be best of their knowledge. Another option for this non-finishing number was that the survey was too long. However, there was evidence from another study using an experimental design with some similarities, that the use of eight videos was considerate adequate in terms of time (Kaler & Green, 2008).

The web-survey allowed for the collection of 570 real participations, the highest number of participations in similar studies[18]. But more important was the diversity of the respondents' education, occupation and experience. It was the first time a tool for data collection was able to congregate such a large variety of backgrounds and experience. For this reason the web--survey was an effective way of collecting respondents scorings needed for the indicator development.

Regarding the development of the modified VAS we reached two important conclusions[19]. The first was that respondents presented different difficulties in discriminating lameness descriptors, and that these difficulties changed with the lameness sign. Therefore, while developing such scoring systems, the integration of lameness signs has to be carefully investigated as it might influence the respondents scoring. The second conclusion was that the scoring of the different lameness signs was not evenly spaced along the VAS continuum. These results suggest that it is relevant to set a new approach for the development of modified VAS, as already intuited by Tuyttens et al. (2009), showing that the exact location

[18] For more information regarding other studies collecting observer's ratings refer to chapter nine.
[19] The discussion leading to these conclusions is presented in chapter nine.

of the thresholds along the VAS continuum has to be carefully assessed, and that lameness signs to be included in the VAS segments should be carefully investigated. These results can be seen as the basis for the development of new modified VAS for lameness scoring in goats, but are also relevant for research in modified VAS in other species.

Regarding the data analysis methodology, the cardinal consistency assessment design proved to be an effective way of analysing lameness data, excluding respondents that were not true--respondents, but allowing for individual variation to be considered. Although this methodology represents an innovative way of assessing lameness data consistency, it is in line with the finding of inconsistencies in inter-observer reliability lameness studies. Therefore, the results found in this study can be compared with the aforementioned reliability studies. From this comparison it can be understood that the difficulties in lameness scoring in goats are similar to the ones found in other species. In this study, almost all the respondents were able to differentiate the severely lame from the non-lame goats, which is in line with results from previous studies in dairy cows and ewes (Brenninkmeyer et al., 2007; Kaler et al., 2009; March et al., 2007; Winckler & Willen, 2001). However, only less than half distinguished non-lame goats from the slightly lame ones, corroborating for goats what was already concluded for other species – most difficulties in lameness scoring are found in the lower end of the scales (Brenninkmeyer et al., 2007; Kaler et al., 2009; Nalon et al., 2014; Winckler & Willen, 2001). Additionally, it was possible to analyse that observers showed an aptitude to differentiate the moderate from the severe cases, contradicting the recommendation to aggregate lameness categories given by some authors (Brenninkmeyer et al., 2007; Channon et al., 2009), and showing that following that recommendation might represent an important loss of information, by failing to detect moderate cases before they become severely lame.

Integration of indicators in a dairy goat on-farm welfare assessment prototype

After developing the BCS and lameness indicators, these were integrated, together with other 22 promising ones, in a welfare assessment prototype developed by the AWIN project. This prototype, that included both pen and individual level observations, was tested in thirty Portuguese dairy goat farms, organized in three size categories. This was the first overall welfare assessment of dairy goats in Portugal, and to the author's knowledge, of any country in the Mediterranean region. Moreover, it was the first study to assess animal-based indicators while considering different farm size categories.

This study allowed us to have an insight on the main welfare problems affecting dairy goats in Portugal, identifying as main areas of concern: the presence of a large proportion of goats with claw overgrowth, very fat animals, poor hair coat condition, and improper disbudding. Moreover, a lot of animals were observed while queuing at feeding[20]. Although the prevalence were independent of farm size, it was shown that larger farms tended to present higher prevalences and so should be viewed with more concern. Likewise, Stafford & Gregory (2008) state that increase in herd size is often associated with a reduction in the human: animal ratio and so with less time to observe individual animals. Analysing the variation in the indicators' prevalence, an overall consistency of results was evident. Moreover, the majority of indicators presented similar prevalences when compared with previous studies (Anzuino et al., 2010 in UK; Muri et al., 2013 in Norway), suggesting common problems in different countries.

The knowledge of the prevalence of the welfare indicators in dairy goat farms provides information on the general health and welfare status of the farms, and enables the future determination of thresholds of acceptability, information that is of extreme importance to the dairy goat industry, as was previously highlighted by Muri et al. (2013). Setting these thresholds is important to stakeholders. For example, they allow farmers, veterinarians and other technicians to identify main welfare issues in their own farms, and therefore set plans of action to improve the general welfare/health condition. A similar tool for farmers developed with the purpose of monitoring and hence improving the animal welfare performance of their businesses was recently set by Colditz et al. (2014). The knowledge of thresholds is also important to improve legislation on animal welfare, when considering the use of animal-based indicators. Moreover, it is important for certification schemes and for consumers as a way to introduce transparency in welfare assessment and product differentiation. This is in line with the concept of benchmarking (welfare comparisons among a given farm population) emphasized in the 2012-2015 EU strategy (European Commission, 2012). By promoting benchmarking voluntary schemes among farmers, the European Commission expects to improve enforcement of the EU legislation in a consistent and transparent way across the Member States, better inform the general public on animal welfare issues, and at the same time encourage competitiveness among livestock producers.

11.2 PRACTICAL APPLICATIONS

It is hard to separate the practical applications of the work conducted during this thesis from the overall objective of the AWIN project, in which it was inserted: the development of an on-farm welfare assessment protocol for dairy goats in intensive systems. However, the scoring systems developed and the knowledge gained by both the literature review and the integration of the indicators, go further beyond the scope of the AWIN project and have multiple applications in the welfare/health assessment field.

[20] For data regarding all the indicators in assessment, refer to chapter ten.

The first practical application of the work was the identification of the 24 animal-based welfare indicators with inherent value to be included in welfare assessment protocols. This review set the basis for the development of the AWIN welfare assessment protocol, but can also aid other researchers to develop further systems for dairy goat welfare assessment with different aims. The 24 animal-based welfare indicators identified are: BCS, queuing at the feed barrier and at the drinker, hair coat condition, displacements at the feed barrier and at the drinker, resting in contact with the wall, shivering and panting score, kneeling at the trough and in the pen, lameness, claw overgrowth, lesions and swellings, teats and udder abnormalities and conformation traits, abscesses, vulvar, ocular, and nasal discharges, ear tear, improper disbudding and dehorning, avoidance distance test, latency to first contact, and qualitative behaviour assessment.

The development of the BCS indicator resulted in a valid and reliable visual BCS system that, with minimum animal handling, allows for a simple and expedite individual assessment of the goats, enhancing BCS feasibility as a welfare indicator. The reliability studies showed that the visual scoring system was by itself clear and unambiguous, and therefore suitable to be used even with low levels of training. This is a valuable characteristic, particularly in the context of welfare assessment, where it is likely that assessors have different levels of experience (Waiblinger et al., 2001). Nevertheless, training did have a positive effect on BCS assessment reliability. For that reason, the training material, innovative due to the use of threshold images, is also of practical value for potential users of the scoring system.

Because of what was mentioned above, the AWIN BCS system has the potential to be applied in different circumstances where welfare assessment schemes can be applied, namely on research, legislative requirements, and certification schemes. Thus, it is expected that the scoring system will have a high acceptability among stakeholders. However, its application within advisory/management tools requires some further attention as it is going to be detailed in the next sub-chapter.

The reliability studies allowed us to establish how previous experience and training influences the final results. This information is of practical value for both users of welfare protocols (it highlights special care that has to be followed on selecting and training observers), and scientists developing reliability studies.

While conducting the studies for the development of the modified VAS for lameness assessment, it was observed that almost all the respondents were able to differentiate the severely lame from the non-lame goats, and a considerable amount of respondents showed an aptitude to differentiate the moderate from the severe cases; whereas less than half were able to clearly separate the slightly-lame goats from the non-lame.

As already discussed for the BCS indicator, in welfare assessment it is common that observers have different levels of experience (Waiblinger et al., 2001), and consequently it is reasonable to considerer one should not expect a great aptitude in lameness assessment. Moreover, by visiting commercial intensive dairy goat farms during the course of the project it was possible to perceive how difficult it is to find a standardized location to assess lameness across different farms. In some farms it was possible to observe animals at the exit of the milking parlour, in others when goats were entering, and in others none of these options were possible and other locations had to be found. The only location that was common across farms were the pens. However, as Anzuino et al. (2010) mention in their study, the lameness prevalence observed while the goats are in their pens (with soft straw flooring) is usually lower than when goats are walking on hard surfaces. This difficulty in standardizing the location to conduct lameness assessment in goat farms, together with the fact that observers with different levels of experience in lameness assessment were only consistent in assessing the severe cases of lameness, as well as the high stocking density and limited time for assessment, led to the decision to only integrate the observation of the most severe lameness cases in the AWIN protocol. The authors understand this represents an important loss of information regarding the lameness status of a farm, however, given the constraints, it was the most feasible approach to the integration of this indicator in the welfare protocol.

A solution found to overcome this limitation and allow a higher level of lameness assessment in farms was the development of a learning object, namely an Application, on lameness called the WelGoat (the Application can be found in http://animalwelfarehub.com/Learning Materials/Details/120). This tool applies the conclusions drawn from the studies for the development of the modified VAS, in a practical way, to help users identify moderately and severely lame goats. The Application is intended for different stakeholders, namely farmers, veterinarians and other technicians; but it also has the potential to be used by researchers and students as it has an educational component. Other two great assets of the Application are that it allows for recording the evolution of the lameness status of different animals, and it also presents important information on claw overgrowth in its relation to lameness, a work developed by WP2 of the AWIN project. Additionally, this kind of approach represents an effort to communicate the research findings to a broader audience, which is valuable in terms of knowledge dissemination in science.

Finally, the knowledge of the main welfare issues affecting thirty Portuguese intensively kept dairy goats, contributed to determine intervention thresholds and guidelines for each welfare indicator that are of enormous value in the context of benchmarking of farms in general, and of the goat industry in Mediterranean countries in particular, helping to improve the on-farm general welfare condition of this production sector. As suggested by Muri et al. (2013), setting of acceptability thresholds can aid the development of welfare advice tailored specifically to each farm in assessment.

The biggest contribution of measuring “is in taking the guesswork out of scientific observation” (Nunnally & Bernstein, 1994, p.6), therefore leading to more objective observations. Replication of experiments is one of the key principles in science; however, for a team of scientists to replicate other team’s work, they have to trust the measurement tools the first team used in their experimental design (Nunnally & Bernstein, 1994). The biggest practical contribution of this thesis is the discussion of measurement requisites. This discussion aims to draw the attention of researchers about the importance of properly assessing existing measurement tools before using them, and discuss some of the necessary requisites when developing new scoring systems.

11.3 LIMITATIONS OF THE RESEARCH WORK AND THE RESULTS PRESENTED

The first limitation that has to be highlighted is the choice of the target of all the research. Taking into account the trend found in goat productive systems around Europe, and the challenges associated with very intensive systems in animal husbandry, special emphasis was given to intensively bred adult dairy goats. The literature review, and the indicators developed and tested for integration on the protocol were all concentrated on adult lactating goats, and did not consider other productive categories such as kids, young female or dry goats. For this reason the on-farm assessment performed may not deliver a full picture of the general welfare in a dairy goat farm. However, given that adult lactating goats are more numerous and constitute the productive category in higher welfare risk, it is possible to have a sufficiently accurate idea of the welfare status of the farm.

Other limitations are associated with inherent methodological weaknesses that can potentially hamper the results obtained. Considering the BCS system development, the main limitations include the decision to focus only on three assessment categories: very thin, normal and very fat goats. This decision made sense taking into consideration that the objective was to increase the indicator's feasibility by developing an exclusively visual BCS system. In this sense, as already presented in the sub-chapter on practical applications of our results, the AWIN BCS can be applied in research, legislative requirements, and certification schemes within the scope of animal welfare. However, in terms of production the very thin and very fat animals represent BCS scores that should not be present on a farm, and therefore are of limited use in terms of advisory and management information. In order to overcome this limitation, the visual scoring system can be extended to other BCS categories. This objective was not pursued given that this was not the overall objective for this particular indicator, but also because methodologically it is was not feasible.

The development of such a scoring system was possible given that there was a high correlation between the visual cues on the rump region and the BCS of the goats, and this allowed for the development of the concept of representative animals of a given category. For the development of such representative animals we used 171 goats from the most common breeds found in intensive production systems. The breed type introduces variability and influences the extent to which the rump visual cues reflect BCS. To attempt to overcome this variation, special attention was given to the unbiased body measurements and shape alignment procedures. However, even after these procedures, it was still possible to observe a large variability between breeds' visual cues and BCS, only allowing for animals' classification in the three categories presented. In order to overcome such limitation and develop a more sensible scoring system we would have to create separate scoring systems considering each breed. However, this option would render the system non-practical for the purpose for which it was developed. In this sense, it should be reinforced that the AWIN scoring system does not aim to replace the current BCS in use, the Hervieu & Morand-Fehr (1999) method that is valuable as an effective production monitoring system.

Another limitation regarding the development of this indicator is that it is potentially biased towards the breeds commonly found in Portuguese intensive farms (Saanen, Alpine and Murciana), and was found to have a more reduced application on other autochthonous breeds such as the Portuguese breed Serrana.

Concerning the lameness indicator, some limitations were also found and should be discussed. One of the first is that in the web-survey the experts used an ordinal scale to score the videos, but the respondents scored lameness signs using a different assessment tool. This previous scoring was essential to compare the survey respondent's scores with an overall lameness scoring, this being a strategy followed in other studies (Nalon et al., 2014)[21]. However, as already presented, there is no universally established scoring system for lameness scoring in goats, which introduces some validity questions with this overall scoring. The scoring system chosen, the one developed by Anzuino et al. (2010), was selected due to its expected face validity as it proved to be efficient in identifying and scoring different levels of lameness in goats housed in a considerable amount of similar farms. The experimental design accounted for this constraint as it has been demonstrated that, in the absence of a gold standard method, a scoring system can be evaluated by assessing its reliability (Shrout, 1998).

Regarding the web-survey methodology its main limitation was that the recruitment strategy did not ensure a true random sampling of respondents. However, as discussed in chapter seven, it is always a risk to make generalizations between a sample and the target population.

[21] Borderas, Fournier, Rushen & de Passillé (2008) and Chapinal, de Passillé, Weary, von Keyserlingk & Rushen (2009) found that observers gave higher correlations between an overall lameness score and some specific gait and posture signs. However, Van Nuffel et al. (2009) showed that observers detected some lameness signs more frequently than others.

The fact that the web-survey link had a wide distribution, including the main international farm animal associations, universities and research institutions on farm animal behaviour, allowed for a probable reduction of biases, at least when compared to the frequent way lameness assessment studies collect observer's ratings (small groups of observers in workshops or conferences or by experts opinions). However, the results presented should always assume some margin for mistake. Another limitation associated with the recruitment strategy was that because data was collected online, and therefore without supervision, it is not possible to guarantee that respondents understood the full range of what was being asked, or if they were answering the survey to the best of their knowledge and perception. In order to overcome this limitation, attention was given to a proper elaboration of guidelines, and the experimental design made use of the technological tools to enhance the respondents understanding. Additionally, it was established a rule that allowed us to verify true respondents. For example, if the participant did not change the default value of level of at least one individual VAS, he/she was not considered a respondent and his/hers assessment was not considered. Moreover, the data analysis strategy accounted for this possibility. While looking for consistency among the respondent's ratings, four competence levels based upon four cardinal consistency levels (Keeney, 1976; Keeney & Kirkwood, 1975) and literature found in lameness assessment (Brenninkmeyer et al., 2007; Kaler et al., 2009, 2009; March et al., 2007; Nalon et al., 2014; Winckler & Willen, 2001) were established. These competence levels determined the minimums and higher levels of competence in lameness assessment. For the purpose of assessing true respondents the first cardinal consistency level established that all respondents would have to distinguish between non-lame and severely lame animals.

The 10% of respondents that could not differentiate between these two levels of assessment should be analysed with care, as it is always important to discuss if they are in fact inconsistent in lameness assessment or should be considered as true non-respondents. The cardinal consistency levels assessment is therefore an adequate way of accounting for this type of problems.

Another limitation of the study on the on-farm welfare assessment of dairy goat farms using animal-based indicators was connected to sampling. The study population was "dairy goats in farms under intensive production system in Portugal". Accordingly to information delivered by DGAV, in the beginning of January 2014 there were a total of 269 farms under intensive production system. In order to have a representative sample of farms that would allow extrapolating the results of the study to the Portuguese reality, the number of sampling units to test would be drawn assuming for instance, a 50% expected ratio (as before data collection reference prevalence were not available), an absolute error of 5%, and calculating the minimum sample size for a 95 % confidence level. This would give a minimum number of

205 farms, a number much higher than the 30 dairy goat farms visited in the study. However, due to project execution constraints the number of Portuguese farms was pre-determined by the AWIN project, to a total of 30 farms. Therefore, this study can only allow us to have an indication on the main welfare problems affecting dairy goats in Portugal. Still, this study presents the first overview of this reality on Mediterranean countries, contributing to an increase awareness of the main welfare issues affecting intensively kept dairy goats, and enhancing the definition of intervention thresholds and guidelines for each welfare indicator, that can aid to improve the on-farm general welfare conditions.

CHAPTER 12

Future research perspectives

Chapter 12 presents some future research perspectives.

Throughout the last chapter, while discussing the main theoretical and methodological results and presenting its limitations, it was possible to identify that some of the work developed can be taken to another level. Therefore, before finishing, we would like to present a perspective of research that could follow the work in this thesis.

As discussed in chapter six, the literature review was a starting point for the identification of animal-based indicators in dairy goats. In this sense, animal-based indicators associated with other categories, such as dry goats, kids, and bucks, should be identified. Only after accomplishing this, will it be possible to add information to the AWIN protocol for lactating goats in order to develop a complete on-farm welfare assessment protocol for dairy farms. Furthermore, this research should be extended to other production systems (e.g., meat and fibber) and to extensive husbandry systems. But more than identifying animal-based indicators, future research regarding animal-based welfare indicators for goats should focus on the development and integration of already identified indicators as highlighted in chapter five, but also on the validation of new indicators to assess criteria that presently have fewer indicators, e.g., positive indicators[22].

Regarding the development of the BCS indicator, it is possible to identify two different future research paths. The first one, with the already developed scoring system, that would aim to integrate this indicator in welfare assessment schemes with advisory purposes. For this to be possible a connection would have to be established between the number of goats scored as very thin and very fat, and the overall BCS condition of a determined group of animals. In this sense, the AWIN scoring system would allow to set a maximum number of very thin and very fat animals that would trigger a second level assessment, which would be conducted using the Hervieu & Morand-Fehr (1999) method. The implementation of this two level evaluation would be extremely valuable for routine farm management.

[22] For a more detailed analysis on positive indicators' assessment see, e.g., Yeates and Main (2008).

The second path of research would be to pursuit the multidisciplinary work developed. With the aim of further refining the BCS system, more images would have to be collected. In this sense, the image acquisition conditions were efficient for their purpose but required careful control of conditions such as: i) keeping the animals still; and ii) the rumps aligned with the camera. An improvement in data collection would be to move towards a more realistic scenario, where a fixed RGB-D camera would be placed on top of the animals' normal path (e.g., at the exit of the milking parlour), and using the collected 3D meshes it would not be necessary any sort of animal handling[23]. The feasibility of using 3D cameras to estimate BCS is already being tested with success in dairy cows (Kuzuhara et al., 2015; Weber et al., 2014). This type of multidisciplinary research and development effort was considered by Morand--Fehr et al. (2004) as a strategic option for goat farming in our century. Also, considering animal welfare research, the possibility to move to a complete automated assessment system, would enhance the efficiency of on-farm protocols. Finally, the methodology used for the development of this indicator can be applied to BCS in other species, or even to other indicators that rely on visual cues.

Regarding the lameness indicator, an immediate focus of future research is the development of the modified VAS. The results presented showed that for the development of this scoring system it was paramount to assess the exact location of the thresholds along the continuum of the VAS[24], as its distribution was not even, and furthermore justify the inclusion of different lameness signs for the labelling of the resulting segments. To be able to explore these questions a possible path is to develop the new modified VAS based on machine learning, namely on data mining techniques. Data mining techniques are generally used for knowledge extraction from data with the objective of finding predictive patters, built upon different variables or features, that are useful to classify different individuals into categories (Provost & Fawcett, 2013). Data mining problems can be of two types: supervised learning, when there is a specific independent target variable that determines the grouping of the observations, and unsupervised learning when such target variable does not exist (Provost & Fawcett, 2013). Among the different data mining techniques for supervised classification, the techniques that apply to our context, decision tree methods are the most appealing because they represent results in a simple format that is easy to interpret and that allow to study the structure of decisions in the classification process (Stiglic, Kocbek, Pernek & Kokol, 2012). Data mining techniques are traditionally used in business problems (Provost & Fawcett, 2013), but in recent years have entered the realm of medical decision-making,

[23] This is also already being tested for features associated with lameness, namely in studies like Pluk et al. 2010 and Viazzi et al. (2013, 2014).

[24] The need to assess the distribution of the thresholds along the continuum of the VAS was intuited by Tuyttens et al. (2009), without, however, assessing for this assumption.

(Kutasi, Balogh, Lajos, Nagy & Szenci, 2011; Nowotny, Rospars, Martinez, Elbanna & Anton, 2013; Paul, Groza, Hunter & Zankl, 2012), animal behaviour (Kubinyi, Turcsán & Miklósi, 2009), identification of risk factors for behavioural problems (Nagy, Reiczigel, Harnos, Schrott & Kabai, 2010).

The results presented in chapter nine also showed that the respondents presented different difficulties in discriminating lameness descriptors, and that these difficulties changed with the lameness sign[25]. Therefore, it would be interesting and important to verify, by means of multivariate statistics, whether there are significant differences in the respondents' performance associated with their education, occupation, previous experience, and other characteristics of the respondents such as age or sex[26]. Additionally, the development of tools to avoid inconsistencies in evaluation and to enhance assessments would be valuable. For example, using the technological advantages of web-surveys, a tool might be created where the respondent received feedback on inconsistencies in lameness scoring, and where they would be given the possibility to repeat assessments. This would be especially valuable in training programmes. Furthermore, these results consider an individual assessment of each lameness sign, hence it would be interesting to verify what would be the ability of the respondents in distinguishing lameness if the three signs (gait, head nodding and arched-back) were considered simultaneously.

Finally, the study on the on-farm welfare assessment of dairy goat farms using animal-based indicators opens several lines of research. Although animal-based indicators are for several reasons preferred over resource-based indicators[27], they are generally more time-consuming, when compared with environmental-based observations. Thus, an important line of research is on sampling procedures that will allow focusing the assessment of a reduced, but representative, number of animals[28]. Additionally, the connection between animal and environmental-based indicators might be another strategic path to reduce the time, and hence cost, associated with the assessment of these indicators. Furthermore, different studies point that the most valid assessment of animal welfare is achieved when these two categories of indicators are used in association (Johnsen et al., 2001; Waiblinger et al., 2001), therefore it is important to direct research towards the identification of the most valid indicators in these two groups. This type of research is also important, in a trend context that some scientists defend, than the role of science in animal welfare should be in identifying and

[25] Further information is presented in chapter eight.

[26] Previous work have identified that experience (Brenninkmeyer et al., 2007; March et al., 2007; O'Callaghan et al, 2003) and occupation (Kaler & Green, 2008; O'Callaghan, 2002; Whay et al., 2003) have influence on lameness scoring. These studies were conducted on dairy cows and ewes, to our knowledge there are no studies available for goats. There is also a study that shows how expectation bias may invalidate subjective recordings of behaviour (Tuyttens et al., 2014).

[27] For further information on advantages and disadvantages of animal and resource-based indicators refer to chapter three.

[28] This is a line of research that has been currently receiving a lot of attention For examples see Endres, Lobeck-Luchterhand, Espejo & Tucker (2014), Hoffman, Moore, Wenz & Vanegas (2013), Mullan et al., (2009), and Waiblinger & Menke (2003).

further preventing welfare problems, instead of focusing on attempts to measure animal welfare as a whole (Fraser, 1995).

Other important research questions related to reducing the time and costs associated with animal welfare assessment is the assessment of the stability of results associated with the assessment of farms (Knierim & Winckler, 2009; Sørensen et al., 2007). In chapter ten a preliminary assessment of the indicators consistency over time, is presented, but further research is needed to identify indicators that, when selected, will allow for a reduction of farm visits[29]. Finally, research aiming to assess the famers' motivation to adopt animal-based indicators to evaluate their daily practices is fundamental to increase the likelihood these measures enter the farms' routine.

[29] More information about this issue is discussed in chapter four.

REFERENCES

Adams, D.S., Klevjer-Anderson, P., Carlson, J.L., McGuire, T.C. & Gorham, J.R. (1983). Transmission and control of caprine arthritis-encephalitis virus. *American Journal of Veterinary Research*, 44(9), 1670–1675.

Ajuda, I. & Stilwell, G. (2014). Using farm routine procedures to reduce pain in dairy goats. [abstract] [electronic version]. In A.J. Zanella (Ed.), *Proceedings of Animal Welfare Indicators 3rd Annual Conference, Prague, 13-15th May 2014*, p. 12. Retrieved from http://www.animal-welfare-indicators.net/site/index.php/awin-events

Ajuda, I., Vieira, A. & Stilwell, G. (2014). Are there differences in dairy goats claws' temperature, before and after trimming? [electronic version]. In *2014 IEEE International Symposium on Medical Measurements and Applications (MeMeA), Lisbon, Portugal, 11-12 June*, pp. 1–5. doi:10.1109/MeMeA.2014.6860149

Alawa, J.P., Ngele, M.B. & Ogwu, D. (2000). Chronic caprine mastitis in Nigerian goat breeds: microbiological flora and histopathological findings. *Small Ruminant Research*, 35(3), 203–207. doi:http://doi.org/10.1016/S0921-4488(99)00099-1

Alban, L. (1995). Lameness in Danish dairy cows: frequency and possible risk factors. *Preventive Veterinary Medicine*, 22(3), 213–225. doi:http://doi.org/10.1016/0167-5877(94)00411-B

Albright, J. & Arawe, C. (1997). *The behaviour of cattle*. Wallingford, Oxon, UK ; New York, NY, USA: CABI.

Al-Ramamneh, D., Riek, A. & Gerken, M. (2012). Effect of water restriction on drinking behaviour and water intake in German black-head mutton sheep and Boer goats. *Animal: An International Journal of Animal Bioscience*, 6(1), 173–178. doi:http://doi.org/10.1017/S1751731111001431

Amao, O.A., Osinowo, O.A., Lakpini, C.A. M., Dipeolu, M.A., Abiola, S.S. & Onwuka, C.F.I. (2003). Types and frequency of udder shapes and abnormalities in West African Dwarf and Red Sokoto goats. *Nigerian Journal of Animal Production*, 30(2), 253–258. doi:http://doi.org/10.4314/njap.v30i2.3303

Ameh, J. A., & Tari, I. S. (1999). Observations on the prevalence of caprine mastitis in relation to predisposing factors in Maiduguri. *Small Ruminant Research*, 35(1), 1–5. doi:http://doi.org/10.1016/S0921-4488(99)00047-4

Ameh, J.A., Addo, P.B., Adekeye, J.O. & Gyang, E.O. (1993). Prevalence of clinical mastitis and of intramammary infections in Nigerian goats. *Preventive Veterinary Medicine*, 17(1–2), 41–46. doi:http://doi.org/10.1016/0167-5877(93)90053-V

Andersen, I.L. & Bøe, K.E. (2007). Resting pattern and social interactions in goats—The impact of size and organisation of lying space. *Applied Animal Behaviour Science*, 108(1-2), 89–103. doi:http://doi.org/10.1016/j.applanim.2006.10.015

Andersen, I.L., Bøe, K.E. & Hove, K. (2000). Behavioural and physiolgical thermoregulation in groups of pregnant sows housed in a kennel system at low temperatures. *Canadian Journal of Animal Science*, 80(1), 1–8. doi:http://doi.org/10.4141/A98-075

Andersen, I.L., Roussel, S., Ropstad, E., Braastad, B.O., Steinheim, G., Janczak, A.M., Jørgensen, G.M. & Bøe, K.E. (2008). Social instability increases aggression in groups of dairy goats, but with minor consequences for the goats' growth, kid production and development. *Applied Animal Behaviour Science*, 114(1–2), 132–148. doi:http://doi.org/10.1016/j.applanim.2008.01.007

Andreasen, S.N. & Forkman, B. (2012). The welfare of dairy cows is improved in relation to cleanliness and integument alterations on the hocks and lameness when sand is used as stall surface. *Journal of Dairy Science*, 95(9), 4961–4967. doi:http://doi.org/10.3168/jds.2011-5169

Anzuino, K., Bell, N.J., Bazeley, K.J. & Nicol, C.J. (2010). Assessment of welfare on 24 commercial UK dairy goat farms based on direct observations. *Veterinary Record*, 167(20), 774 –780. doi:http://doi.org/10.1136/vr.c5892

Arrow, K.J. (1951). *Social choice and individual values*. New Jersey, USA: Wiley.

Aschwanden, J., Gygax, L., Wechsler, B. & Keil, N.M. (2008). Social distances of goats at the feeding rack: Influence of the quality of social bonds, rank differences, grouping age and presence of horns. *Applied Animal Behaviour Science*, 114(1–2), 116–131. doi:http://doi.org/10.1016/j.applanim.2008.02.002

Aschwanden, J., Gygax, L., Wechsler, B. & Keil, N.M. (2009a). Loose housing of small goat groups: Influence of visual cover and elevated levels on feeding, resting and agonistic behaviour. *Applied Animal Behaviour Science*, 119(3–4), 171–179. doi:http://doi.org/10.1016/j.applanim.2009.04.005

Aschwanden, J., Gygax, L., Wechsler, B. & Keil, N.M. (2009b). Structural modifications at the feeding place: Effects of partitions and platforms on feeding and social behaviour of goats. *Applied Animal Behaviour Science*, 119(3–4), 180–192. doi:http://doi.org/10.1016/j.applanim.2009.04.004

Averbuch, M. & Katzper, M. (2004). Assessment of visual analog versus categorical scale for measurement of osteoarthritis pain. *Journal of Clinical Pharmacology*, 44(4), 368–372. doi:http://doi.org/10.1177/0091270004263995

AWIN. (2015). *AWIN welfare assessment protocol for goats*. doi:http://dx.doi.org/10.13130/AWIN_goats_2015

Aziz, M.A. (2010). Present status of the world goat populations and their productivity. *Lohmann Information*, 45(2), 42. Retrieved from http://lohmann-information.com/content/l_i_45_artikel17.pdf

Azzaro, G., Caccamo, M., Ferguson, J.D., Battiato, S., Farinella, G.M., Guarnera, G.C., Puglisi, G., Petriglieri, R. & Licitra, G. (2011). Objective estimation of body condition score by modeling cow body shape from digital images. *Journal of Dairy Science*, 94(4), 2126–2137. doi:http://doi.org/10.3168/jds.2010-3467

Baird, G.J. & Fontaine, M.C. (2007). Corynebacterium pseudotuberculosis and its role in ovine caseous lymphadenitis. *Journal of Comparative Pathology*, 137(4), 179–210. doi:http://doi.org/10.1016/j.jcpa.2007.07.002

Bana e Costa, C.A. & Beinat, E. (2005). *Model-structuring in public decision-aiding*. Working paper, London School of Economics and Political Science.

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(2–3), 179–192. doi:http://doi.org/10.1016/S0301-6226(99)00067-6

Battini, M. (2013). *Animal-based indicators for on-farm welfare assessment in cattle and goats* (PhD Thesis). University of Milan, Italy.

Battini, M., Andreoli, E., Barbieri, S. & Mattiello, S. (2011). Long-term stability of Avoidance Distance tests for on-farm assessment of dairy cow relationship to humans in alpine traditional husbandry systems. *Applied Animal Behaviour Science*, 135(4), 267–270. doi:http://doi.org/10.1016/j.applanim.2011.10.013

Battini, M., Barbieri, S., Fioni, L. & Mattiello, S. (2015). Feasibility and validity of animal-based indicators for on-farm welfare assessment of thermal stress in dairy goats. *International Journal of Biometeorology*. doi:http://doi.org/doi:10.1007/s00484-015-1025-7

Battini, M., Grosso, L., Barbieri, S., Ajuda, I., Vieira, A., Stilwell, G. & Mattiello, S. (2013). Le condizioni del mantello sono un valido indicatore di benessere nella capra? *Large Anim. Rev*, 5 (Suppl. 1), 10. Retrieved from http://www.sipaoc.it/documenti/2012/XX%20Congresso%20Nazionale%20SIPAOC%20-%20Atti.pdf

Battini, M., Vieira, A., Barbieri, S., Ajuda, I., Stilwell, G. & Mattiello, S. (2014). Invited review: Animal-based indicators for on-farm welfare assessment for dairy goats. *Journal of Dairy Science*, 97(11), 6625–6648. doi:http://doi.org/10.3168/jds.2013-7493

Battini, M., Peric, T., Ajuda, I., Vieira, A., Grosso, L., Barbieri, S. Stilwell, G., Prandi, A., Comin, A., Tubaro, F. & Mattiello, S. (2015). Hair coat condition: A valid and reliable indicator for on-farm welfare assessment in adult dairy goats. *Small Ruminant Research*, 123(2–3), 197–203. doi:http://doi.org/10.1016/j.smallrumres.2014.12.009

Bell, N.J., Bell, M.J., Knowles, T.G., Whay, H.R., Main, D.J. & Webster, A.J.F. (2009). The development, implementation and testing of a lameness control programme based on HACCP principles and designed for heifers on dairy farms. *Veterinary Journal (London, England: 1997)*, 180(2), 178–188. doi:http://doi.org/10.1016/j.tvjl.2008.05.020

Belongie, S., Malik, J. & Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4), 509–522. doi:http://doi.org/10.1109/34.993558

Bercovich, A., Edan, Y., Alchanatis, V., Moallem, U., Parmet, Y., Honig, H., Maltz, E., Antler, A. & Halachmi, I. (2013). Development of an automatic cow body condition scoring using body shape signature and Fourier descriptors. *Journal of Dairy Science*, 96(12), 8047–8059. doi:http://doi.org/10.3168/jds.2013-6568

Bergonier, D., Berthelot, X. & Poumarat, F. (1997). Contagious agalactia of small ruminants: current knowledge concerning epidemiology, diagnosis and control. *Revue Scientifique Et Technique (International Office of Epizootics)*, 16(3), 848–873. Retrieved from http://www.oie.int/doc/ged/d9177.pdf

Bewley, J.M., Peacock, A.M., Lewis, O., Boyce, R.E., Roberts, D.J., Coffey, M.P., Kenyon, S.J. & Schutz, M.M. (2008). Potential for estimation of body condition scores in dairy cattle from digital images. *Journal of Dairy Science*, 91(9), 3439–3453. doi:http://doi.org/10.3168/jds.2007-0836

Blackie, N., Amory, J., Bleach, E. & Scaife, J. (2011). The effect of lameness on lying behaviour of zero grazed Holstein dairy cattle. *Applied Animal Behaviour Science*, 134(3–4), 85–91. doi:http://doi.org/10.1016/j.applanim.2011.08.004

Bland, J.M. & Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 1(8476), 307–310. doi:10.1016/s0140-6736(86)90837-8

Blokhuis, H.J., Keeling, L.J., Gavinelli, A. & Serratosa, J. (2008). Animal welfare's impact on the food chain. *Trends in Food Science & Technology*, 19, (Supplement 1), S79–S87. doi:http://doi.org/10.1016/j.tifs.2008.09.007

Blokhuis, H.J., Veissier, I., Miele, M. & Jones, B. (2010). The Welfare Quality® project and beyond: Safeguarding farm animal well-being. *Acta Agriculturae Scandinavica, Section A - Animal Science*, 60(3), 129–140. doi:http://doi.org/10.1080/09064702.2010.523480

Blokhuis, H.J., Miele, M., Veissier, I. & Jones, B. (2013). *Improving farm animal welfare - Science and society working together: the Welfare Quality approach*. Wageningen Academic Publishers. Retrieved from http://www.springer.com/life+sciences/book/978-90-8686-770-7

Bock, B.B. (2009). Farmers' perspectives. In A. Butterworth, H. Blokhuis, B. Jones & I. Veissier (Eds.), *Proceedings Conference Delivering animal welfare and quality: transparency in the food production chain: Including the final results of the Welfare Quality® project, 8 – 9 October 2009, Uppsala, Sweden,* pp. 73–75.

Bøe, K. E., Berg, S. & Andersen, I. L. (2006). Resting behaviour and displacements in ewes—effects of reduced lying space and pen shape. *Applied Animal Behaviour Science*, 98(3–4), 249–259. doi:http://doi.org/10.1016/j.applanim.2005.10.001

Bøe, K.E., Andersen, I.L., Buisson, L., Simensen, E. & Jeksrud, W.K. (2007). Flooring preferences in dairy goats at moderate and low ambient temperature. *Applied Animal Behaviour Science*, 108(1–2), 45–57. doi:http://doi.org/10.1016/j.applanim.2006.12.002

Bøe, K.E. & Ehrlenbruch, R. (2013). Thermoregulatory behavior of dairy goats at low temperatures and the use of outdoor yards. *Canadian Journal of Animal Science*, 93(1), 35–41. doi:http://doi.org/10.4141/cjas2012-028

Borderas, T.F., Fournier, A., Rushen, J. & de Passillé, A.M.B. (2008). Effect of lameness on dairy cows' visits to automatic milking systems. *Canadian Journal of Animal Science*, 88(1), 1–8. doi:http://doi.org/10.4141/CJAS07014

Botreau, R., Bonde, M., Butterworth, A., Perny, P., Bracke, M.B.M., Capdeville, J. & Veissier, I. (2007). Aggregation of measures to produce an overall assessment of animal welfare. Part 1: A review of existing methods. *Animal*, 1(08), 1179-1187. doi:10.1017/s1751731107000535

Botreau, R., Bracke, M., Perny, P., Butterworth, A., Capdeville, J., Van Reenen, C. & Veissier, I. (2007). Aggregation of measures to produce an overall assessment of animal welfare. Part 2: Analysis of constraints. *Animal*, 1(8), 1188-1197. doi:10.1017/s1751731107000547

Botreau, R., Veissier, I., Butterworth, A., Bracke, M. & Keeling, L. (2007). Definition of criteria for overall assessment of animal welfare. *Animal Welfare*, 16(2), 225–228.

Botreau, R., Capdeville, J., Perny, P. & Veissier, I. (2008). Multicriteria evaluation of animal welfare at farm level: An application of mcda methodologies. Foundation of Computing and Decision Sciences, 33, 1–18.

Botreau, R., Veissier, I. & Perny, P. (2009). Overall assessment of animal welfare: strategy adopted in Welfare Quality®. *Animal Welfare*, 18(4), 363–370.

Boyazoglu, J., Hatziminaoglou, I. & Morand-Fehr, P. (2005). The role of the goat in society: Past, present and perspectives for the future. *Small Ruminant Research*, 60(1–2), 13–23. doi:http://doi.org/10.1016/j.smallrumres.2005.06.003

Bracke, M.B.M., Spruijt, B.M., Metz, J.H.M. & Schouten, W.G.P. (2002). Decision support system for overall welfare assessment in pregnant sows A: Model structure and weighting procedure. *Journal of Animal Science*, 80(7), 1819–1834. doi:/2002.8071819x

Brambell Committee. (1965). *Report of the technical committee to enquire into the welfare of animals kept under intensive livestock husbandry systems. Command report 2836*. London, UK: Her Majesty's Stationery Office.

Braun, U., Rihs, T. & Schefer, U. (1992). Ruminal lactic acidosis in sheep and goats. *Veterinary Record*, 130(16), 343–349. doi:http://doi.org/10.1136/vr.130.16.343

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.

Breuer, K., Hemsworth, P., Barnett, J., Matthews, L. & Coleman, G. (2000). Behavioural response to humans and the productivity of commercial dairy cows. *Applied Animal Behaviour Science*, 66(4), 273–288. doi:10.1016/s0168-1591(99)00097-0

Broom, D.M. (1986). Indicators of poor welfare. *British Veterinary Journal*, 142(6), 524–526. doi:http://doi.org/10.1016/0007-1935(86)90109-0

Broom, D.M. (1991). Animal welfare: concepts and measurement. *Journal of Animal Science*, 69(10), 4167–4175.

Broom, D.M. (1998). Welfare, stress and the evolution of feelings. *Advances in the Study Behavior*, 27, 371–403. doi:10.1016/s0065-3454(08)60369-1

Broom, D.M. & Johnson, K.G. (1994). *Stress and Animal Welfare*. (2nd ed.). London ; New York: Springer.

Brozos, C., Mavrogianni, V.S. & Fthenakis, G.C. (2011). Treatment and control of peri-parturient metabolic diseases: pregnancy toxemia, hypocalcemia, hypomagnesemia. *The Veterinary Clinics of North America. Food Animal Practice*, 27(1), 105–113. doi:http://doi.org/10.1016/j.cvfa.2010.10.004

Brunberg, E., Wallenbeck, A. & Keeling, L.J. (2011). Tail biting in fattening pigs: Associations between frequency of tail biting and other abnormal behaviours. *Applied Animal*

Behaviour Science, 133(1–2), 18–25. doi:http://doi.org/10.1016/j.applanim.2011.04.019

Bruno-de-Sousa, C., Martinez, A.M., Ginja, C., Santos-Silva, F., Carolino, M.I., Delgado, J.V. & Gama, L.T. (2011). Genetic diversity and population structure in Portuguese goat breeds. *Livestock Science*, 135(2–3), 131–139. doi:http://doi.org/10.1016/j.livsci.2010.06.159

Buchanan, E.A. & Hvizdak, E.E. (2009). Online survey tools: ethical and methodological concerns of human research ethics committees. *Journal of Empirical Research on Human Research Ethics: JERHRE*, 4(2), 37–48. doi:http://doi.org/10.1525/jer.2009.4.2.37

Burges, C.J.C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2), 121–167. doi:http://doi.org/10.1023/A:1009715923555

Burn, C.C. & Weir, A.A.S. (2011). Using prevalence indices to aid interpretation and comparison of agreement ratings between two or more observers. *Veterinary Journal (London, England: 1997)*, 188(2), 166–170. doi:http://doi.org/10.1016/j.tvjl.2010.04.021

Canali, E. & Keeling, L. (2010). Welfare Quality® project: from scientific research to on farm assessment of animal welfare. *Italian Journal of Animal Science*, 8(2s), 900. doi:http://doi.org/10.4081/ijas.2009.s2.900

Canali, E., Whay, H.R. & Leach, K.A. (2009). Cattle health status. In B. Forkman & L. Keeling (Eds.), *Assessment of Animal Welfare Measures for Dairy Cattle, Beef Bulls and Veal Calves. Welfare Quality Reports, Vol. 11.* (pp. 77–88). Cardiff, UK:

Capdeville, J. & Veissier, I. (2001). A method of assessing welfare in loose housed dairy cows at farm level, focusing on animal observations. *Acta Agriculturae Scandinavica, Section A - Animal Science*, 51(sup030), 62–68. doi:http://doi.org/10.1080/090647001316923081

Carbonaro, D.A., Friend, T.H., Dellmeier, G.R. & Nuti, L C. (1992). Behavioral and physiological responses of dairy goats to food thwarting. *Physiology & Behavior*, 51(2), 303–308. doi:10.1016/0031-9384(92)90145-r

Caroprese, M., Casamassima, D., Rassu, S.P.G., Napolitano, F. & Sevi, A. (2009). Monitoring the on-farm welfare of sheep and goats. *Italian Journal of Animal Science*, 8(1s), 343–354. doi:http://doi.org/10.4081/ijas.2009.s1.343

Castel, J.M., Ruiz, F.A., Mena, Y. & Sánchez-Rodríguez, M. (2010). Present situation and future perspectives for goat production systems in Spain. *Small Ruminant Research*, 89(2–3), 207–210. doi:http://doi.org/10.1016/j.smallrumres.2009.12.045

Channon, A.J., Walker, A.M., Pfau, T., Sheldon, I.M. & Wilson, A.M. (2009). Variability of Manson and Leaver locomotion scores assigned to dairy cows by different observers. *The Veterinary Record*, 164(13), 388–392. doi:10.1136/vr.164.13.388

Chapinal, N., de Passillé, A.M., Weary, D.M., von Keyserlingk, M.A.G. & Rushen, J. (2009). Using gait score, walking speed, and lying behavior to detect hoof lesions in dairy cows. *Journal of Dairy Science*, 92(9), 4365–4374. doi:http://doi.org/10.3168/jds.2009-2115

Christodoulopoulos, G. (2009). Foot lameness in dairy goats. *Research in Veterinary Science*, 86(2), 281–284. doi:http://doi.org/10.1016/j.rvsc.2008.07.013

Churchill, G.A. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, 16(1), 64. doi:http://doi.org/10.2307/3150876

Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4), 213–220. doi:10.1037/h0026256

Colditz, I.G., Ferguson, D.M., Collins, T., Matthews, L. & Hemsworth, P.H. (2014). A prototype tool to enable farmers to measure and improve the welfare performance of the farm animal enterprise: The unified field index. *Animals*, 4(3), 446–462. doi:http://doi.org/10.3390/ani4030446

Contreras, A., Sierra, D., Sánchez, A., Corrales, J.C., Marco, J.C., Paape, M.J. & Gonzalo, C. (2007). Mastitis in small ruminants. *Small Ruminant Research*, 68(1), 145–153. doi:http://doi.org/10.1016/j.smallrumres.2006.09.011

Cornetto, T. & Estevez, I. (2001). Influence of vertical panels on use of space by domestic fowl. *Applied Animal Behaviour Science*, 71(2), 141–153. doi:10.1016/s0168-1591(00)00171-4

Costa-Santos, C., Bernardes, J., Ayres-de-Campos, D., Costa, A., Amorim-Costa, C. & Costa, C. (2011). The limits of agreement and the intraclass correlation coefficient may be inconsistent in the interpretation of agreement. *Journal of Clinical Epidemiology*, 64(3), 264–269. doi:http://doi.org/10.1016/j.jclinepi.2009.11.010

Cronbach, L.J. & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302. doi:10.1037/h0040957

Cronin, G.M., Rault, J.L. & Glatz, P.C. (2014). Lessons learned from past experience with intensive livestock management systems. *Revue Scientifique Et Technique (International Office of Epizootics)*, 33(1), 139–151. Retrieved from http://www.oie.int/doc/ged/D13661.PDF

Dalla Villa, P., Matthews, L.R., Alessandrini, B., Messori, S. & Migliorati, G. (2014). Drivers for animal welfare policies in Europe. *Revue Scientifique Et Technique (International Office of Epizootics)*, 33(1), 39–46. Retrieved from http://www.oie.int/doc/ged/D13649.PDF

Darcan, N., Cedden, F. & Cankaya, S. (2009). Spraying effects on some physiological and behavioural traits of goats in a subtropical climate. *Italian J. Anim. Sci.*, 7, 77–85. doi:10.4081/ijas.2008.77

Dawkins, M. (2004). Using behaviour to assess animal welfare. *Animal Welfare*, 13(1), 3–7. Retrieved from http://www.fao.org/fileadmin/user_upload/animalwelfare/Dawkins%20behavior%202004.pdf

De Passillé, A.M. & Rushen, J. (2005). Can we measure human–animal interactions in on-farm animal welfare assessment?: Some unresolved issues. *Applied Animal*

Behaviour Science, 92(3), 193–209. doi:http://doi.org/10.1016/j.applanim.2005.05.006

De Rancourt, M., Fois, N., Lavín, M. P., Tchakérian, E. & Vallerand, F. (2006). Mediterranean sheep and goats production: An uncertain future. *Small Ruminant Research*, 62(3), 167–179. doi:http://doi.org/10.1016/j.smallrumres.2005.08.012

De Rosa, G., Grasso, F., Pacelli, C., Napolitano, F. & Winckler, C. (2009). The welfare of dairy buffalo. *Italian Journal of Animal Science*, 8(1s), 103–116. doi:http://doi.org/10.4081/ijas.2009.s1.103

De Rosa, G., Napolitano, F., Grasso, F., Pacelli, C. & Bordi, A. (2005). On the development of a monitoring scheme of buffalo welfare at farm level. *Italian Journal of Animal Science*, 4(2), 115–125. doi:http://doi.org/10.4081/ijas.2005.115

De Vet, H.C.W., Terwee, C.B., Knol, D.L. & Bouter, L.M. (2006). When to use agreement versus reliability measures. Journal of Clinical Epidemiology, 59(10), 1033–1039. doi:http://doi.org/10.1016/j.jclinepi.2005.10.015

De Vries, M., Bokkers, E.A.M., Dijkstra, T., van Schaik, G. & de Boer, I.J.M. (2011). Invited review: associations between variables of routine herd data and dairy cattle welfare indicators. *Journal of Dairy Science*, 94(7), 3213–3228. doi:http://doi.org/10.3168/jds.2011-4169

De Vries, M., Bokkers, E.A.M., van Schaik, G., Engel, B., Dijkstra, T. & de Boer, I.J.M. (2014). Exploring the value of routinely collected herd data for estimating dairy cattle welfare. *Journal of Dairy Science*, 97(2), 715–730. doi:http://doi.org/10.3168/jds.2013-6585

Deinhofer, M. & Pernthaner, A. (1995). Staphylococcus spp. as mastitis-related pathogens in goat milk. *Veterinary Microbiology*, 43(2-3), 161–166. doi:10.1016/0378-1135(95)92532-G

Dohoo, I., Martin, W. & Stryhn, H. (2009). Screening and diagnostic tests. In *Veterinary Epidemiologic Research*. (2nd ed.). (pp. 91–134). Charlottetown, Prince Edward Island, Canada: VER, Incorporated.

Dubeuf, J.P. (2005). Structural, market and organisational conditions for developing goat dairy production systems. *Small Ruminant Research*, 60(1–2), 67–74. doi:http://doi.org/10.1016/j.smallrumres.2005.06.015

Dubeuf, J.P. (2010). Characteristics and diversity of the dairy goat production systems and industry around the world. *Tecnol. & Ciên. Agropec*, 4(4), 25–31. Retrieved from http://emepa.org.br/revista/volumes/tca_v4_n4_dez/tca03_characteristic.pdf

Dubeuf, J.P., Morand-Fehr, P. & Rubino, R. (2004). Situation, changes and future of goat industry around the world. *Small Ruminant Research*, 51(2), 165–173. doi:http://doi.org/10.1016/j.smallrumres.2003.08.007

Duncan, I.J.H. (1993). Welfare is to do with what animals feel. *Journal of Agricultural and Environmental Ethics*, 6(Supplement 2), 8–14.

Duncan, I.J.H. (1996). Animal welfare defined in terms of feelings. *Acta Agriculturae Scandinavica Section A, Animal Science, Supplement* 27, 29–35.

Edmonson, A.J., Lean, I.J., Weaver, L.D., Farver, T. & Webster, G. (1989). A Body Condition Scoring Chart for Holstein Dairy Cows. *Journal of Dairy Science*, 72(1), 68–78. doi:http://doi.org/10.3168/jds.S0022-0302(89)79081-0

European Food Safety Authority (2012). Statement on the use of animal-based measures to assess the welfare of animals. Panel on Animal Health and Welfare (AHAW). *EFSA Journal*, 10(6), 2767. doi:http://dx.doi.org/10.2903/j.efsa.2012.2767

Ehrlenbruch, R., Jørgensen, G.H.M., Andersen, I.L. & Bøe, K.E. (2010a). Provision of additional walls in the resting area—The effects on resting behaviour and social interactions in goats. *Applied Animal Behaviour Science*, 122(1), 35–40. doi:http://doi.org/10.1016/j.applanim.2009.11.004

Ehrlenbruch, R., Pollen, T., Andersen, I. L. & Bøe, K. E. (2010b). Competition for water at feeding time—The effect of increasing number of individuals per water dispenser. *Applied Animal Behaviour Science*, 126(3–4), 105–108. doi:http://doi.org/10.1016/j.applanim.2010.06.004

Ellis, K.A., Innocent, G.T., Mihm, M., Cripps, P., McLean, W.G., Howard, C.V. & Grove-White, D. (2007). Dairy cow cleanliness and milk quality on organic and conventional farms in the UK. *The Journal of Dairy Research*, 74(3), 302–310. doi:http://doi.org/10.1017/S002202990700249X

Endres, M.I., Lobeck-Luchterhand, K.M., Espejo, L.A. & Tucker, C.B. (2014). Evaluation of the sample needed to accurately estimate outcome-based measurements of dairy welfare on farm. *Journal of Dairy Science*, 97(6), 3523–3530. doi:http://doi.org/10.3168/jds.2013-7464

Engel, B., Bruin, G., Andre, G. & Buist, W. (2003). Assessment of observer performance in a subjective scoring system: visual classification of the gait of cows. *The Journal of Agricultural Science*, 140(03), 317–333. doi:http://doi.org/10.1017/S0021859603002983

Escareño, L., Salinas-Gonzalez, H., Wurzinger, M., Iñiguez, L., Sölkner, J. & Meza-Herrera, C. (2013). Dairy goat production systems: status quo, perspectives and challenges. *Tropical Animal Health and Production*, 45(1), 17–34. doi:http://doi.org/10.1007/s11250-012-0246-6

European Commission. (2011). Evaluation of CAP measures for the sheep and goat sector. *Tender N° AGRI 2010 EVAL 02*. Retrieved from http://ec.europa.eu/agriculture/eval/reports/sheep-goat/executive_summary_en.pdf

European Commission. (2012). *Commission staff working paper impact assessment – Communication from the Commission to the European Parliament, the Council and the European Economic and Social Committee on the European Union strategy for the protection and welfare of animals*. Brussels: EC. Retrieved from http://ec.europa.eu/food/animal/welfare/actionplan/docs/impact_assesment_19012012_en.pdf

Eze, C.A. (2002). Lameness and reproductive performance in small ruminants in Nsukka Area of the Enugu State, Nigeria. *Small Ruminant Research*, 44(3), 263–267. doi:http://doi.org/10.1016/S0921-4488(02)00030-5

Færevik, G., Andersen, I.L. & Bøe, K.E. (2005). Preferences of sheep for different types of pen flooring. *Applied Animal Behaviour Science*, 90(3–4), 265–276. doi:http://doi.org/10.1016/j.applanim.2004.08.010

Farm Animal Welfare Council. (2009). *FAWC - Farm Animal Welfare Council*. Retrieved April 1, 2015, from http://webarchive.nationalarchives.gov.uk/20121007104210/http:/www.fawc.org.uk/freedoms.htm

Ferguson, J.D., Galligan, D.T. & Thomsen, N. (1994). Principal descriptors of body condition score in Holstein cows. *Journal of Dairy Science*, 77(9), 2695–2703. doi:http://doi.org/10.3168/jds.S0022-0302(94)77212-X

Ferguson, J.D., Azzaro, G. & Licitra, G. (2006). Body condition assessment using digital images. *Journal of Dairy Science*, 89(10), 3833–3841. doi:http://doi.org/10.3168/jds.S0022-0302(06)72425-0

Fernández, M.A., Alvarez, L. & Zarco, L. (2007). Regrouping in lactating goats increases aggression and decreases milk production. *Small Ruminant Research*, 70(2–3), 228–232. doi:http://doi.org/10.1016/j.smallrumres.2006.03.008

Ferreira, L.M.M., Celaya, R., Benavides, R., Jáuregui, B.M., García, U., Sofia Santos, A., García,R.R. Rodrigues, M.A.M. & Osoro, K. (2013). Foraging behaviour of domestic herbivore species grazing on heathlands associated with improved pasture areas. *Livestock Science*, 155(2–3), 373–383. doi:http://doi.org/10.1016/j.livsci.2013.05.007

Fioni, L. (2014). *Validazione di indicatori di comfort termico per la valutazione del benessere negli allevamenti caprini da latte*. University of Milan, Italy.

Flower, F.C. & Weary, D.M. (2009). Gait assessment in dairy cattle. *Animal*, 3(01), 87–95. doi:http://doi.org/10.1017/S1751731108003194

Foddai, A., Green, L.E., Mason, S.A. & Kaler, J. (2012). Evaluating observer agreement of scoring systems for foot integrity and footrot lesions in sheep. *BMC Veterinary Research*, 8(1), 65. doi:http://doi.org/10.1186/1746-6148-8-65

Food and Agriculture Organization of the United Nations. (2015). *Dairy production and products: Small ruminants*. Retrieved from http://www.fao.org/agriculture/dairy-gateway/milk-production/dairy-animals/smalruminants/en/#.VQDOBOHHLQM

Food and Agriculture Organization of the United Nations Statistics Division. (2015, May 5). *FAOSTAT*. Retrieved from http://faostat3.fao.org/browse/Q/QA/E

Forkman, B. (1996). The Social Facilitation of Drinking: What is Facilitated, and Who is Affected? *Ethology*, 102(2), 252–258. doi:http://doi.org/10.1111/j.1439-0310.1996.tb01122.x

Fournier, F. & Festa-Bianchet, M. (1995). Social dominance in adult female mountain goats. *Animal Behaviour*, 49(6), 1449–1459. doi:http://doi.org/10.1016/0003-3472(95)90066-7

Fraser, D. (1995). Science, values and animal welfare: Exploring the “inextricable connection”. *Animal Welfare*, 4(2), 103–117.

Fraser, D. (2006). Animal welfare assurance programs in food production: A framework for assessing the options. *Animal Welfare*, (15), 93–104. Retrieved from http://www.fao.org/fileadmin/user_upload/animalwelfare/Animal%20welfare%20assurance%20programs.pdf

Fraser, D. (2009). Animal behaviour, animal welfare and the scientific study of affect. *Applied Animal Behaviour Science*, 118(3–4), 108–117. doi:http://doi.org/10.1016/j.applanim.2009.02.020

Fraser, D., Weary, D.M., Pajor, E.A. & Milligan, B.N. (1997). A scientific conception of animal welfare that reflects ethical concerns. *Animal Welfare*, 6(3), 187–205. Retrieved from http://www.fao.org/fileadmin/user_upload/animalwelfare/Fraser1997.pdf

Fuller, C.J., Bladon, B.M., Driver, A.J. & Barr, A.R.S. (2006). The intra- and inter-assessor reliability of measurement of functional outcome by lameness scoring in horses. *The Veterinary Journal*, 171(2), 281–286. doi:http://doi.org/10.1016/j.tvjl.2004.10.012

Galindo, F., Broom, D.M. & Jackson, P.G.G. (2000). A note on possible link between behaviour and the occurrence of lameness in dairy cows. *Applied Animal Behaviour Science*, 67(4), 335–341. doi:http://doi.org/10.1016/S0168-1591(99)00114-8

Garner, J.P., Falcone, C., Wakenell, P., Martin, M. & Mench, J.A. (2002). Reliability and validity of a modified gait scoring system and its use in assessing tibial dyschondroplasia in broilers. *British Poultry Science*, 43(3), 355–363. doi:http://doi.org/10.1080/00071660120103620

Gaughan, J.B. (2003). Using panting scores to assess heat load in cattle. In N. Lacetera, U. Bernabucci, H. H. Khalifa, B. Ronchi & A. Nardone*, EAAP European Association for Animal Production Technical Series: Interactions between climate & animal production - EaapTS07,* EAAP Technical Series, 7, p.99. Wageningen, The Netherlands: Wageningen Academic Publishers.

Gaynor, J.S. & Muir, W.W. (2008). *Handbook of Veterinary Pain Management*. Elsevier Health Sciences.

Gibbons, J., Vasseur, E., Rushen, J. & Passille, A.M. de. (2012). A training programme to ensure high repeatability of injury scoring of dairy cows. *Animal Welfare*, 21(3), 379–388. doi:http://doi.org/10.7120/09627286.21.3.379

Goldhawk, C., Chapinal, N., Veira, D.M., Weary, D.M. & von Keyserlingk, M.A.G. (2009). Prepartum feeding behavior is an early indicator of subclinical ketosis. *Journal of Dairy Science*, 92(10), 4971–4977. doi:http://doi.org/10.3168/jds.2009-2242

González-Pachón, J., Diaz-Balteiro, L. & Romero, C. (2014). How to combine inconsistent ordinal and cardinal preferences: A satisficing modelling approach. *Computers & Industrial Engineering*, 67, 168–172. doi:http://doi.org/10.1016/j.cie.2013.11.008

Görgülü, M., Boğa, M., Şahin, A., Serbester, U., Kutlu, H.R. & Şahinler, S. (2008). Diet selection and eating behaviour of lactating goats subjected to time restricted feeding in choice and single feeding system. *Small Ruminant Research*, 78(1–3), 41–47. doi:http://doi.org/10.1016/j.smallrumres.2008.04.004

Green, L.E., Hedges, V.J., Schukken, Y.H., Blowey, R.W. & Packington, A.J. (2002). The impact of clinical lameness on the milk yield of dairy cows. *Journal of Dairy Science*, 85(9), 2250–2256. doi:http://doi.org/10.3168/jds.S0022-0302(02)74304-X

Griffioen, J.K D.H.M. (2012, May 17). *Does self-sucking in dairy goats develop in the raising period* (PhD Thesis). Retrieved from http://dspace.library.uu.nl/handle/1874/240726

Grove-White, D. (2004). Rumen healthcare in the dairy cow. *In Practice*, 26(2), 88–95. doi:http://doi.org/10.1136/inpract.26.2.88

Gundelach, Y., Schulz, T., Feldmann, M. & Hoedemaker, M. (2013). Effects of increased vigilance for locomotion disorders on lameness and production in dairy cows. *Animals*, 3(3), 951–961. doi:http://doi.org/10.3390/ani3030951

Hadjigeorgiou, I. (2011). Past, present and future of pastoralism in Greece. *Pastoralism: Research, Policy and Practice*, 1(1), 24. doi:http://doi.org/10.1186/2041-7136-1-24

Halachmi, I., Polak, P., Roberts, D.J. & Klopcic, M. (2008). Cow body shape and automation of condition scoring. *Journal of Dairy Science*, 91(11), 4444–4451. doi:http://doi.org/10.3168/jds.2007-0785

Hallgren, K.A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, 8(1), 23–34. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3402032/

Harwood, D. (2006). *Goat health and welfare: A veterinary huide*. Wiltshire, UK: The Crowood Press Ltd.

Hemsworth, P.H. (2003). Human–animal interactions in livestock production. *Applied Animal Behaviour Science*, 81(3), 185–198. doi:http://doi.org/10.1016/S0168-1591(02)00280-0

Hemsworth, P.H. & Barnett, J.L. (1991). The effects of aversively handling pigs, either individually or in groups, on their behaviour, growth and corticosteroids. *Applied Animal Behaviour Science*, 30(1–2), 61–72. doi:http://doi.org/10.1016/0168-1591(91)90085-C

Hemsworth, P.H., Coleman, G.J., Barnett, J.L. & Borg, S. (2000). Relationships between human-animal interactions and productivity of commercial dairy cows. *Journal of Animal Science*, 78(11), 2821–2831. Retrieved from https://www.animalsciencepublications.org/publications/jas/pdfs/78/11/2821

Hernandez, J.A., Garbarino, E.J., Shearer, J.K., Risco, C.A. & Thatcher, W. W. (2005). Comparison of the calving-to-conception interval in dairy cows with different degrees of lameness during the prebreeding postpartum period. *Journal of the American Veterinary Medical Association*, 227(8), 1284–1291.

Hervieu, J. & Morand-Fehr, P. (1999). Comment noter l'état corporel des chèvres. *Réussir La Chèvre. La Revue Des Éleveurs de Chèvres. Istitut de L' Elevage*, 231, 26–32.

Hewetson, M., Christley, R.M., Hunt, I.D. & Voute, L.C. (2006). Investigations of the reliability of observational gait analysis for the assessment of lameness in horses. *The Veterinary Record*, 158(25), 852–857. doi:10.1136/vr.158.25.852

Hill, N.P., Murphy, P.E., Nelson, A.J., Mouttotou, N., Green, L.E. & Morgan, K.L. (1997). Lameness and foot lesions in adult British dairy goats. *The Veterinary Record*, 141(16), 412–416. doi:10.1136/vr.141.16.412

Hoffman, A.C., Moore, D.A., Vanegas, J. & Wenz, J.R. (2014). Association of abnormal hind-limb postures and back arch with gait abnormality in dairy cattle. *Journal of Dairy Science*, 97(4), 2178–2185. doi:http://doi.org/10.3168/jds.2013-7528

Horgan, R. & Gavinelli, A. (2006). The expanding role of animal welfare within EU legislation and beyond. *Livestock Science*, 103(3), 303–307. doi:http://doi.org/10.1016/j.livsci.2006.05.019

Hughes, B.O. & Duncan, I.J.H. (1988). The notion of ethological "need", models of motivation and animal welfare. *Animal Behaviour*, 36(6), 1696–1707. doi:http://doi.org/10.1016/S0003-3472(88)80110-6

Hughes, J. (2001). A system for assessing cow cleanliness. *In Practice*, 23(9), 517–524. doi:http://doi.org/10.1136/inpract.23.9.517

Huzzey, J.M., Veira, D.M., Weary, D.M. & von Keyserlingk, M.A.G. (2007). Prepartum behavior and dry matter intake identify dairy cows at risk for metritis. *Journal of Dairy Science*, 90(7), 3220–3233. doi:http://doi.org/10.3168/jds.2006-807

Ingenbleek, P.T.M., Immink, V.M., Spoolder, H.A.M., Bokma, M.H. & Keeling, L.J. (2012). EU animal welfare policy: Developing a comprehensive policy framework. *Food Policy*, 37(6), 690–699. doi:http://doi.org/10.1016/j.foodpol.2012.07.001

Instituto Nacional de Estatística. (2011). *Recenseamento Agrícola 2009: Análise dos principais resultados*. Lisboa: INE.

Instituto Nacional de Estatística. (2015). *Portal Nacional de Estatística: Base de dados*. Retrieved from http://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&indOcorrCod=0000963&contexto=bd&selTab=tab2

Ito, K., Chapinal, N., Weary, D.M. & von Keyserlingk, M.A.G. (2014). Associations between herd-level factors and lying behavior of freestall-housed dairy cows. *Journal of Dairy Science*, 97(4), 2081–2089. doi:http://doi.org/10.3168/jds.2013-6861

Jackson, K.M.A. & Hackett, D. (2007). A note: The effects of human handling on heart girth, behaviour and milk quality in dairy goats. *Applied Animal Behaviour Science*, 108(3–4), 332–336. doi:http://doi.org/10.1016/j.applanim.2007.01.011

Jago, J.G., Krohn, C.C. & Matthews, L.R. (1999). The influence of feeding and handling on the development of the human–animal interactions in young cattle. *Applied Animal Behaviour Science*, 62(2–3), 137–151. doi:http://doi.org/10.1016/S0168-1591(98)00219-6

Jefferies, B.C. (1961). Body condition scoring and its use in management. *Tasmanian J. Agr.*, 32, 19–21.

Jensen, M.B. & Kyhn, R. (2000). Play behaviour in group-housed dairy calves, the effect of space allowance. *Applied Animal Behaviour Science*, *67*(1-2), 35–46.

Jimenez-Granado, R., Sanchez-Rodriguez, M., Arce, C. & Rodriguez-Estevez, V. (2014). Factors affecting somatic cell count in dairy goats: a review. *Span. J. Agric. Res.*, 12, 133–150. doi:http://doi.org/http://dx.doi.org/10.5424/sjar/2014121-3803

Johnsen, P.F., Johannesson, T. & Sandøe, P. (2001). Assessment of Farm Animal Welfare at Herd Level: Many Goals, Many Methods. *Acta Agriculturae Scandinavica, Section A — Animal Science*, 51(sup030), 26–33. doi:http://doi.org/10.1080/090647001316923027

Jong, de I.C., Gunnink, H. & Harn, van J. (2014). Wet litter not only induces footpad dermatitis but also reduces overall welfare, technical performance, and carcass yield in broiler chickens. *Journal of Applied Poultry Research*, 23(1), 51–58. doi:10.3382/japr.2013-00803

Jørgensen, G.H.M., Andersen, I.L. & Bøe, K.E. (2007). Feed intake and social interactions in dairy goats—The effects of feeding space and type of roughage. *Applied Animal Behaviour Science*, 107(3), 239–251. doi:http://doi.org/10.1016/j.applanim.2006.10.007

Juarez, S.T., Robinson, P.H., DePeters, E.J. & Price, E. O. (2003). Impact of lameness on behavior and productivity of lactating Holstein cows. *Applied Animal Behaviour Science*, 83(1), 1–14. doi:http://doi.org/10.1016/S0168-1591(03)00107-2

Kaler, J. & Green, L.E. (2008). Recognition of lameness and decisions to catch for inspection among sheep farmers and specialists in GB. *BMC Veterinary Research*, 4(1), 41. doi:http://doi.org/10.1186/1746-6148-4-41

Kaler, J., Wassink, G.J. & Green, L.E. (2009). The inter- and intra-observer reliability of a locomotion scoring scale for sheep. *Veterinary Journal (London, England: 1997)*, 180(2), 189–194. doi:http://doi.org/10.1016/j.tvjl.2007.12.028

Kamboj, M.L., Vishwakarma, V.K. & Singh, A. (2007). Abnormal behaviours in dairy cattle and buffaloes: causes and their management. *Indian Dairyman*, 59, 27–33.

Kamphaus, R.W. & Frick, P.J. (2005). *Clinical assessment of child and adolescent personality and behavior*. (2nd ed.). New York, USA: Springer Science & Business Media.

Katsoulos, P.D. & Christodoulopoulos, G. (2009). Prevalence of lameness and of associated claw disorders in Greek dairy cattle industry. *Livestock Science*, 122(2–3), 354–358. doi:http://doi.org/10.1016/j.livsci.2008.09.012

Kaufman, A.B. & Rosenthal, R. (2009). Can you believe my eyes? The importance of interobserver reliability statistics in observations of animal behaviour. *Animal Behaviour*, 78(6), 1487–1491. doi:http://doi.org/10.1016/j.anbehav.2009.09.014

Keeling, L.J. (2005). Healthy and happy: animal welfare as an integral part of sustainable agriculture. *Ambio*, 34(4-5), 316–319. doi:10.1579/0044-7447-34.4.316

Keeney, R.L. (1976). A Group Preference Axiomatization with Cardinal Utility. *Management Science*, 23(2), 140–145. doi:http://doi.org/10.1287/mnsc.23.2.140

Keeney, R.L. & Kirkwood, C.W. (1975). Group Decision Making Using Cardinal Social Welfare Functions. *Management Science*, 22(4), 430–437. doi:http://doi.org/10.1287/mnsc.22.4.430

Kelly, P.C., More, S.J., Blake, M. & Hanlon, A.J. (2011). Identification of key performance indicators for on-farm animal welfare incidents: possible tools for early warning and prevention. *Irish Veterinary Journal*, 64(1), 13. doi:http://doi.org/10.1186/2046-0481-64-13

Klaas, I.C., Enevoldsen, C., Vaarst, M. & Houe, H. (2004). Systematic clinical examinations for identification of latent udder health types in Danish dairy herds. *Journal of Dairy Science*, 87(5), 1217–1228. doi:http://doi.org/10.3168/jds.S0022-0302(04)73272-5

Knierim, U. & Winckler, C. (2009). On-farm welfare assessment in cattle: validity, reliability and feasibility issues and future perspectives with special regard to the Welfare Quality approach. *Animal Welfare*, 18(4), 451–458.

Koop, G., Collar, C.A., Toft, N., Nielen, M., van Werven, T., Bacon, D. & Gardner, I. A. (2013). Risk factors for subclinical intramammary infection in dairy goats in two longitudinal field studies evaluated by Bayesian logistic regression. *Preventive Veterinary Medicine*, 108(4), 304–312. doi:http://doi.org/10.1016/j.prevetmed.2012.11.007

Koop, G., Nielen, M. & van Werven, T. (2012). Diagnostic tools to monitor udder health in dairy goats. *The Veterinary Quarterly*, *32*(1), 37–44. doi:http://doi.org/10.1080/01652176.2012.675634

Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B.J., Hróbjartsson, A., Roberts, C., Shoukri, M. & Streiner, D. L. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *Journal of Clinical Epidemiology*, 64(1), 96–106. doi:http://doi.org/10.1016/j.jclinepi.2010.03.002

Kougioumtzis, A., Valergakis, G.E., Oikonomou, G., Arsenos, G. & Banos, G. (2014). Profile and genetic parameters of dairy cattle locomotion score and lameness across lactation. *Animal: An International Journal of Animal Bioscience*, 8(1), 20–27. doi:http://doi.org/10.1017/S1751731113001717

Krieg, A. & Peterhans, E. (1990). Caprine arthritis-encephalitis in Switzerland: epidemiological and clinical studies. *Schweiz. Arch. Tierh*, 132, 345–352.

Kristensen, E., Dueholm, L., Vink, D., Andersen, J E., Jakobsen, E.B., Illum-Nielsen, S., Petersen, F.A. & Enevoldsen, C. (2006). Within- and across-person uniformity of body condition scoring in Danish Holstein cattle. *Journal of Dairy Science*, 89(9), 3721–3728. doi:http://doi.org/10.3168/jds.S0022-0302(06)72413-4

Kubinyi, E., Turcsán, B. & Miklósi, A. (2009). Dog and owner demographic characteristics and dog personality trait associations. *Behavioural Processes*, 81(3), 392–401. doi:http://doi.org/10.1016/j.beproc.2009.04.004

Kutasi, O., Balogh, N., Lajos, Z., Nagy, K. & Szenci, O. (2011). Diagnostic Approaches for the Assessment of Equine Chronic Pulmonary Disorders. *Journal of Equine Veterinary Science*, 31(7), 400–410. doi:http://doi.org/10.1016/j.jevs.2011.03.018

Kuzuhara, Y., Kawamura, K., Yoshitoshi, R., Tamaki, T., Sugai, S., Ikegami, M., Kurokawa, Y., Obitsu, T., Okita, M., Sugino, T. & Yasuda, T. (2015). A preliminarily study for predicting body weight and milk properties in lactating Holstein cows using a three-dimensional camera system. *Computers and Electronics in Agriculture*, 111, 186–193. doi:http://doi.org/10.1016/j.compag.2014.12.020

Laister, S., Brorkens, N., Lolli, S., Zucca, D., Knierim, U., Minero, M., Canali & Winckler, C. (2009). Reliability of measures of agonistic behaviour in dairy and beef cattle. In B. Forkman & L. Keeling (Eds.), *Assessment of Animal Welfare Measures for Dairy Cattle, Beef Bulls and Veal Calves. Welfare Quality Reports, Vol. 11.* Cardiff, UK.

Lantz, C.A. (1997). Application and evaluation of the kappa statistic in the design and interpretation of chiropractic clinical research. *Journal of Manipulative and Physiological Therapeutics*, 20(8), 521–528.

Laporte-Broux, B., Duvaux-Ponter, C., Roussel, S., Promp, J., Chavatte-Palmer, P. & Ponter, A.A. (2011). Restricted feeding of goats during the last third of gestation modifies both metabolic parameters and behaviour. *Livestock Science*, *138*(1–3), 74–88. doi:http://doi.org/10.1016/j.livsci.2010.12.008

Leach, K., Knierim, U. & Whay, H. (2009). Cleanliness scoring for dairy and beef cattle and veal calves. In B. Forkman & L. Keeling (Eds.), *Assessment of Animal Welfare Measures for Dairy Cattle, Beef Bulls and Veal Calves. Welfare Quality Reports, Vol. 11*. (pp. 25–30). Cardiff, UK.

Leach, K.A., Knierim, U. & Whay, H.R. (2009). Condition scoring for dairy and beef cattle and veal calves. In B. Forkman & L. Keeling (Eds.), *Assessment of Animal Welfare Measures for Dairy Cattle, Beef Bulls and Veal Calves. Welfare Quality Reports, Vol. 11*. (pp. 1–6). Cardiff, UK.

Leitner, G., Silanikove, N. & Merin, U. (2008). Estimate of milk and curd yield loss of sheep and goats with intrammamary infection and its relation to somatic cell count. *Small Ruminant Research*, 74(1–3), 221–225. doi:http://doi.org/10.1016/j.smallrumres.2007.02.009

Lengarite, M.I., Mbugua, P.N., Gachuiri, C.K. & Kabuage, L.W. (2012). Herders' knowledge on mineral nutrition and implication on sheep and goat productivity in Marsabit South District, Kenya. *Livestock Research for Rural Development*, 24(4). Retrieved from http://www.lrrd.org/lrrd24/4/leng24057.htm

Lensink, B.J., Fernandez, X., Boivin, X., Pradel, P., Le Neindre, P. & Veissier, I. (2000). The impact of gentle contacts on ease of handling, welfare, and growth of calves and on quality of veal meat. *Journal of Animal Science*, 78(5), 1219–1226. Retrieved from https://www.animalsciencepublications.org/publications/jas/pdfs/78/5/1219

Lensink, B.J., Raussi, S., Boivin, X., Pyykkönen, M. & Veissier, I. (2001). Reactions of calves to handling depend on housing condition and previous experience with humans. *Applied Animal Behaviour Science*, *70*(3), 187–199. doi:10.1016/S0168-1591(00)00152-0

Lensink, B.J., van Reenen, C.G., Engel, B., Rodenburg, T.B. & Veissier, I. (2003). Repeatability and reliability of an approach test to determine calves' responsiveness to humans: "a brief report". *Applied Animal Behaviour Science*, 83(4), 325–330. doi:http://doi.org/10.1016/S0168-1591(03)00138-2

Ley, S.J., Livingston, A. & Waterman, A.E. (1989). The effect of chronic clinical pain on thermal and mechanical thresholds in sheep. *Pain*, 39(3), 353–357. doi:10.1016/0304-3959(89)90049-3

Lievaart, J.J. & Noordhuizen, J.P.T.M. (2011). Ranking experts' preferences regarding measures and methods of assessment of welfare in dairy herds using Adaptive Conjoint Analysis. *Journal of Dairy Science*, 94(7), 3420–3427. doi:http://doi.org/10.3168/jds.2010-3954

Lima, M.S., Pascoal, R.A. & Stilwell, G.T. (2012a). Glycaemia as a sign of the viability of the foetuses in the last days of gestation in dairy goats with pregnancy toxaemia. *Irish Veterinary Journal*, 65(1), 1. doi:http://doi.org/10.1186/2046-0481-65-1

Lima, M.S., Pascoal, R.A., Stilwell, G.T. & Hjerpe, C.A. (2012b). Clinical findings, blood chemistry values, and epidemiologic data from dairy goats with pregnancy toxemia. *The Bovine Practitioner*, 46, 102–110.

Loretz, C., Wechsler, B., Hauser, R. & Rüsch, P. (2004). A comparison of space requirements of horned and hornless goats at the feed barrier and in the lying area. *Applied Animal Behaviour Science*, 87(3–4), 275–283. doi:http://doi.org/10.1016/j.applanim.2004.01.005

Lowe, D.G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), 91–110. doi:http://doi.org/10.1023/B:VISI.0000029664.99615.94

Lowman, B.G., Scott, N.A. & Somerville, S.H. (1976). *Condition scoring of cattle*. Edinburgh School of Agriculture.

Lund, V., Coleman, G., Gunnarsson, S., Appleby, M.C. & Karkinen, K. (2006). Animal welfare science—Working at the interface between the natural and social sciences. *Applied Animal Behaviour Science*, 97(1), 37–49. doi:http://doi.org/10.1016/j.applanim.2005.11.017

Main, D.C.J. (2009). Application of welfare assessment to commercial livestock production. *Journal of Applied Animal Welfare Science: JAAWS*, 12(2), 97–104. doi:http://doi.org/10.1080/10888700902719658

Main, D.C.J., Webster, A.J.F. & Green, L.E. (2001). Animal Welfare Assessment in Farm Assurance Schemes. *Acta Agriculturae Scandinavica, Section A - Animal Science*, 51(sup030), 108–113. doi:http://doi.org/10.1080/090647001316923171

Main, D.C.J., Kent, J.P., Wemelsfelder, F., Ofner, E. & Tuyttens, F.A.M. (2003). Applications for methods of on-farm welfare assessment. *Animal Welfare*, 12(4), 523–528.

Main, D.C.J., Barker, Z.E., Leach, K.A., Bell, N.J., Whay, H.R. & Browne, W.J. (2010). Sampling strategies for monitoring lameness in dairy cattle. *Journal of Dairy Science*, 93(5), 1970–1978. doi:http://doi.org/10.3168/jds.2009-2500

Manteuffel, G., Puppe, B. & Schön, P.C. (2004). Vocalization of farm animals as a measure of welfare. *Applied Animal Behaviour Science*, 88(1–2), 163–182. doi:http://doi.org/10.1016/j.applanim.2004.02.012

Mantova, E. (2012). *La presenza di ascessi come indicatore di benessere negli allevamenti caprini da latte* (BSc Thesis). University of Milan, Italy.

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.

Marsden, M.D. & Wood-Gush, D.G.M. (1986). The use of space by group housed sheep. *Applied Animal Behaviour Science*, 15, 178. doi:10.1016/0168-1591(86)90064-x

Martínez, A.M., Vega-Pla, J.L., León, J.M., Camacho, M.E., Delgado, J.V. & Ribeiro, M.N. (2010). Is the Murciano-Granadina a single goat breed? A molecular genetics approach. *Arquivo Brasileiro de Medicina Veterinária E Zootecnia*, 62(5), 1191–1198. doi: http://doi.org/10.1590/S0102-09352010000500023

Martínez-de la Puente, J., Moreno-Indias, I., Morales-Delanuez, A., Ruiz-Díaz, M.D., Hernández-Castellano, L.E., Castro, N. & Argüello, A. (2011). Effects of feeding management and time of day on the occurrence of self-suckling in dairy goats. *The Veterinary Record*, 168(14), 378. doi:http://doi.org/10.1136/vr.c6483

Martin, P. & Bateson, P. (2007). *Measuring behaviour: an introductory guide*. Cambridge; New York: Cambridge University Press.

Mason, G. & Mendl, M. (1993). Why is there no simple way of Measuring Animal Welfare? *Animal Welfare*, 2(4), 301–319. Retrieved from https://atrium.lib.uoguelph.ca/xmlui/bitstream/handle/10214/4673/Mason_%26_Mendl_1993.pdf?sequence=1

Mason, G. & Rushen, J. (2008). *Stereotypic Animal Behaviour: Fundamentals and Applications to Welfare*. Wallingford, UK: CABI.

Mattiello, S., Canali, E., Ferrante, V., Caniatti, M., Gottardo, F., Cozzi, G., Andrighetto, I. & Verga, M. (2002). The provision of solid feeds to veal calves: II. Behavior, physiology, and abomasal damage. *Journal of Animal Science*, 80(2), 367–375. Retrieved from https://www.animalsciencepublications.org/publications/jas/pdfs/80/2/367

Mattiello, S., Villa, S. & Cioccarelli, G. (2008). Effetto delle modalità di gestione dell'alimentazione sul benessere del capretto. *Large Anim. Rev*, 4.

Mattiello, S., Battini, M., Andreoli, E., Minero, M., Barbieri, S. & Canali, E. (2010). Avoidance distance test in goats: A comparison with its application in cows. *Small Ruminant Research*, 91(2–3), 215–218. doi:http://doi.org/10.1016/j.smallrumres.2010.03.002

Mattiello, S., Battini, M., Andreoli, E. & Barbieri, S. (2011). Short communication: Breed differences affecting dairy cattle welfare in traditional alpine tie-stall husbandry systems. *Journal of Dairy Science*, 94(5), 2403–2407. doi:http://doi.org/10.3168/jds.2010-3606

Mavrogianni, V.S., Alexopoulos, C. & Fthenakis, G. C. (2004). Field evaluation of flunixin meglumine in the supportive treatment of caprine mastitis. *Journal of Veterinary Pharmacology and Therapeutics*, 27(5), 373–375. doi:http://doi.org/10.1111/j.1365-2885.2004.00590.x

Mavrogianni, V.S., Menzies, P.I., Fragkou, I.A. & Fthenakis, G.C. (2011). Principles of mastitis treatment in sheep and goats. *The Veterinary Clinics of North America. Food Animal Practice*, 27(1), 115–120. doi:http://doi.org/10.1016/j.cvfa.2010.10.010

Mazurek, M., Marie, M. & Desor, D. (2005). Animal-centred indicators of dairy goat welfare. In *Proceedings XIIth Congress of International Society for Animal Hygiene* (ISAH, Vol. 1). Warsaw, Poland. Retrieved from http://www.isah-soc.org/documents/2005/sections/28_vol_1.pdf

Mazurek, M., Marie, M. & Desor, D. (2007). Potential animal-centred indicators of dairy goat welfare. *Animal Welfare*, 16(2), 161–164.

Mazurek, M., McGee, M., Crowe, M. A., Prendiville, D. J., Boivin, X. & Earley, B. (2011). Consistency and stability of behavioural fear responses of heifers to different fear-eliciting situations involving humans. *Applied Animal Behaviour Science*, 131(1–2), 21–28. doi:http://doi.org/10.1016/j.applanim.2011.01.004

McGregor, B.A. (2002). Extent and source of short and cotted mohair. *RIRDC Research Paper*, 02(108), 60. Retrieved from https://rirdc.infoservices.com.au/downloads/02-108

McGregor, B.A. & Butler, KL. (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. *Australian Veterinary Journal*, 86(1-2), 12–17. doi:http://doi.org/10.1111/j.1751-0813.2007.00249.x

Meagher, R.K. (2009). Observer ratings: Validity and value as a tool for animal welfare research. *Applied Animal Behaviour Science*, 119(1–2), 1–14. doi:http://doi.org/10.1016/j.applanim.2009.02.026

Mendl, M., Burman, O.H.P., Parker, R.M.A. & Paul, E.S. (2009). Cognitive bias as an indicator of animal emotion and welfare: Emerging evidence and underlying mechanisms. *Applied Animal Behaviour Science*, 118(3–4), 161–181. doi:http://doi.org/10.1016/j.applanim.2009.02.023

Menzies, P.I. & Ramanoon, S.Z. (2001). Mastitis of sheep and goats. *The Veterinary Clinics of North America. Food Animal Practice*, 17(2), 333–358, vii.

Merbitz, C. & Morris, J. (1989). Ordinal scales and foundations of misinference. *Archives of Physical Medicine and Rehabilitation*, 70(4), 308–12.

Minero, M., Tosi, M.V., Canali, E. & Wemelsfelder, F. (2009). Quantitative and qualitative assessment of the response of foals to the presence of an unfamiliar human. *Applied Animal Behaviour Science*, 116(1), 74–81. doi:http://doi.org/10.1016/j.applanim.2008.07.001

Miranda-de la Lama, G.C. & Mattiello, S. (2010). The importance of social behaviour for goat welfare in livestock farming. *Small Ruminant Research*, 90(1–3), 1–10. doi:http://doi.org/10.1016/j.smallrumres.2010.01.006

Mononen, J., Møller, S.H., Hansen, S.W., Hovland, A.L., Koistinen, T., Lidfors, L., Malmkvist, J., Vinke, C.M. & Ahola, L. (2012). The development of on-farm welfare assessment protocols for foxes and mink: the WelFur project. *Animal Welfare*, 21(3), 363. doi:10.7120/09627286.21.3.363

Morand-Fehr, P. (2005). Recent developments in goat nutrition and application: A review. *Small Ruminant Research*, 60(1–2), 25–43. doi:http://doi.org/10.1016/j.smallrumres.2005.06.004

Morand-Fehr, P., Boutonnet, J. P., Devendra, C., Dubeuf, J.P., Haenlein, G.F.W., Holst, P., Mowlemg, L. & Capote, J. (2004). Strategy for goat farming in the 21st century. *Small Ruminant Research*, 51(2), 175–183. doi:http://doi.org/10.1016/j.smallrumres.2003.08.01

Morgan, F., Massouras, T., Barbosa, M., Roseiro, L., Ravasco, F., Kandarakis, I., Bonnin, V., Fistakoris, M., Anifantakis, E., Jaubert, G. & Raynal-Ljutovac, K. (2003). Characteristics of goat milk collected from small and medium enterprises in Greece, Portugal and France. *Small Ruminant Research*, 47(1), 39–49. doi:http://doi.org/10.1016/S0921-4488(02)00252-3

Mount, L.E. (1979). *Adaptation to thermal environment. Man and his productive animals.* London, UK: Edward Arnold.

Mullan, S., Browne, W.J., Edwards, S.A., Butterworth, A., Whay, H.R. & Main, D.C.J. (2009). The effect of sampling strategy on the estimated prevalence of welfare outcome measures on finishing pig farms. *Applied Animal Behaviour Science*, 119(1–2), 39–48. doi:http://doi.org/10.1016/j.applanim.2009.03.008

Mülleder, C., Troxler, J., Laaha, G. & Waiblinger, S. (2007). Can environmental variables replace some animal-based parameters in welfare assessment of dairy cows? *Animal Welfare*, 16(2), 153–156.

Muri, K., Stubsjoen, S. & Valle, P. (2013). Development and testing of an on-farm welfare assessment protocol for dairy goats. *Animal Welfare*, 22(3), 385–400. doi:http://doi.org/10.7120/09627286.22.3.385

Muri, K., Leine, N. & Valle, P.S. (2015). Welfare effects of a disease eradication programme for dairy goats. *Animal: An International Journal of Animal Bioscience*, 1–9. doi:http://doi.org/10.1017/S1751731115000762

Nagy, D.W., & Pugh, D.G. (2012). Handling and Examining Sheep and Goats. In D.G. Pugh & A.N. Baird (Eds.), *Sheep and Goat Medicine* (2nd ed). (pp. 1-17). Maryland Heights, Missouri: Elsevier Saunders.

Nagy, K., Reiczigel, J., Harnos, A., Schrott, A., & Kabai, P. (2010). Tree-based methods as an alternative to logistic regression in revealing risk factors of crib-biting in horses. *Journal of Equine Veterinary Science*, 30(1), 21–26. doi:http://doi.org/10.1016/j.jevs.2009.11.005

Nalon, E., Maes, D., Van Dongen, S., van Riet, M.M.J., Janssens, G.P.J., Millet, S. & Tuyttens, F.A.M. (2014). Comparison of the inter- and intra-observer repeatability of three gait-scoring scales for sows. *Animal: An International Journal of Animal Bioscience*, 8(4), 650–659. doi:http://doi.org/10.1017/S1751731113002462

Napolitano, F., De Rosa, G., Girolami, A., Scavone, M. & Braghieri, A. (2011). Avoidance distance in sheep: Test–retest reliability and relationship with stockmen attitude. *Small Ruminant Research*, 99(2–3), 81–86. doi:http://doi.org/10.1016/j.smallrumres.2011.03.044

Napolitano, F., De Rosa, G., Serrapica, M. & Braghieri, A. (2015). A continuous recording approach to qualitative behaviour assessment in dairy buffaloes (Bubalus bubalis). *Applied Animal Behaviour Science*, 166, 35–43. doi:http://doi.org/10.1016/j.applanim.2015.01.017

Napolitano, F., Knierim, U., Grasso, F. & De Rosa, G. (2009). Positive indicators of cattle welfare and their applicability to on-farm protocols. *Italian Journal of Animal Science*, 8(1s), 355. doi:http://doi.org/10.4081/ijas.2009.s1.355

Negretti, P., Bianconi, G., Bartocci, S., Terramoccia, S. & Verna, M. (2008). Determination of live weight and body condition score in lactating Mediterranean buffalo by Visual Image Analysis. *Livestock Science*, 113(1), 1–7. doi:http://doi.org/10.1016/j.livsci.2007.05.018

Nielsen, B.L., Lawrence, A.B. & Whittemore, C.T. (1995). Effect of group size on feeding behaviour, social behaviour, and performance of growing pigs using single-space feeders. *Livestock Production Science*, 44(1), 73–85. doi:http://doi.org/10.1016/0301-6226(95)00060-X

Nordmann, E., Keil, N.M., Schmied-Wagner, C., Graml, C., Langbein, J., Aschwanden, J.,von Hof, J., Maschat, K., Palme, R. & Waiblinger, S. (2011). Feed barrier design affects behaviour and physiology in goats. *Applied Animal Behaviour Science*, 133(1–2), 40–53. doi:http://doi.org/10.1016/j.applanim.2011.04.016

Nowotny, T., Rospars, J.P., Martinez, D., Elbanna, S. & Anton, S. (2013). Machine learning for automatic prediction of the quality of electrophysiological recordings. *PLoS ONE*, 8(12), e80838. doi:http://doi.org/10.1371/journal.pone.0080838

Nunnally, J.C. & Bernstein, I.H. (1994). *Psychometric theory*. New York, USA: McGraw-Hill.

O'Callaghan, K. (2002). Lamneness and associated pain in cattle - challenging traditional perceptions. *In Practice*, 24(4), 212–219. doi:http://doi.org/10.1136/inpract.24.4.212

O'Callaghan, K.A., Cripps, P.J., Downham, D.Y. & Murray, R.D. (2003). Subjective and objective assessment of pain and discomfort due to lameness in dairy cattle. *Animal Welfare*, 12(4), 605–610. Retrieved from ftp://s173-183-201-52.ab.hsia.telus.net/Inetpub/wwwroot/HoofHealth/refs/AW12_605.pdf

Ogebe, P.O., Ogunmodede, B.K. & McDowell, L.R. (1996). Behavioral and physiological responses of Nigerian dwarf goats to seasonal changes of the humid tropics. *Small Ruminant Research*, 22(3), 213–217. doi:http://doi.org/10.1016/S0921-4488(96)00877-2

Olofsson, J. (1999). Competition for total mixed diets fed for ad libitum intake using one or four cows per feeding station. *Journal of Dairy Science*, 82(1), 69–79. doi:http://doi.org/10.3168/jds.S0022-0302(99)75210-0

Otten, D. (2013). The application of animal welfare standards in intensive production systems using the assessment protocols of Welfare Quality: Fattening pig husbandry in Northwest Germany. *International Journal of Livestock Production*, 4(4), 49–59. doi:http://doi.org/10.5897/IJLP12.037

Paape, M.J., Wiggans, G.R., Bannerman, D.D., Thomas, D.L., Sanders, A.H., Contreras, A., Moroni, P. & Miller, R.H. (2007). Monitoring goat and sheep milk somatic cell counts. *Small Ruminant Research*, 68(1–2), 114–125. doi:http://doi.org/10.1016/j.smallrumres.2006.09.014

Palmer, M.A., Law, R. & O'Connell, N.E. (2012). Relationships between lameness and feeding behaviour in cubicle-housed Holstein–Friesian dairy cows. *Applied Animal Behaviour Science*, 140(3–4), 121–127. doi:http://doi.org/10.1016/j.applanim.2012.06.005

Paterna, A., Contreras, A., Gómez-Martín, A., Amores, J., Tatay-Dualde, J., Prats-van der Ham, M. & De la Fe, C. (2014). The diagnosis of mastitis and contagious agalactia in dairy goats. Small Ruminant Research, 121(1), 36–41. doi:10.1016/j.smallrumres.2013.12.002

Paul, R., Groza, T., Hunter, J. & Zankl, A. (2012). Decision support methods for finding phenotype — Disorder associations in the bone dysplasia domain. *PLoS ONE*, 7(11), e50614. doi:http://doi.org/10.1371/journal.pone.0050614

Paul-Dauphin, A., Guillemin, F., Virion, J.M. & Briançon, S. (1999). Bias and precision in Visual Analogue Scales: A randomized controlled trial. *American Journal of Epidemiology*, 150(10), 1117–1127. doi:10.1093/oxfordjournals.aje.a009937

Perrin, G.G., Mallereau, M.P., Lenfant, D. & Baudry, C. (1997). Relationships between California mastitis test (CMT) and somatic cell counts in dairy goats. *Small Ruminant Research*, 26(1–2), 167–170. doi:http://doi.org/10.1016/S0921-4488(96)00990-X

Phythian, C.J., Michalopoulou, E., Jones, P.H., Winter, A.C., Clarkson, M.J., Stubbings, L.A. Grove-White, D., Cripps, P.J. & Duncan, J.S. (2011). Validating indicators of sheep welfare through a consensus of expert opinion. *Animal*, 5(06), 943–952. doi:http://doi.org/10.1017/S1751731110002594

Phythian, C.J., Cripps, P.J., Michalopoulou, E., Jones, P.H., Grove-White, D., Clarkson, M.J., Winter, A.C., Stubbings, L.A. & Duncan, J.S. (2012). Reliability of indicators of sheep welfare assessed by a group observation method. *Veterinary Journal (London, England: 1997)*, 193(1), 257–263. doi:http://doi.org/10.1016/j.tvjl.2011.12.006

Phythian, C., Michalopoulou, E., Duncan, J. & Wemelsfelder, F. (2013). Inter-observer reliability of Qualitative Behavioural Assessments of sheep. *Applied Animal Behaviour Science*, 144(1–2), 73–79. doi:http://doi.org/10.1016/j.applanim.2012.11.011

Pirisi, A., Lauret, A. & Dubeuf, J. P. (2007). Basic and incentive payments for goat and sheep milk in relation to quality. *Small Ruminant Research*, 68(1), 167–178. doi:http://doi.org/10.1016/j.smallrumres.2006.09.009

Plesch, G., Broerkens, N., Laister, S., Winckler, C. & Knierim, U. (2010). Reliability and feasibility of selected measures concerning resting behaviour for the on-farm welfare assessment in dairy cows. *Applied Animal Behaviour Science*, 126(1–2), 19–26. doi:http://doi.org/10.1016/j.applanim.2010.05.003

Pluk, A., Bahr, C., Leroy, T., Poursaberi, A., Song, X., Vranken, E., Maertens, W., Van Nuffel, A. & Berckmans, D. (2010). Evaluation of step overlap as an automatic measure in dairy cow locomotion. *Transactions of the ASABE*, 53(4), 1305–1312. doi:10.13031/2013.32580

Plummer, P.J. & Schleining, J.A. (2013). Assessment and management of pain in small ruminants and camelids. *The Veterinary Clinics of North America. Food Animal Practice*, 29(1), 185–208. doi:http://doi.org/10.1016/j.cvfa.2012.11.004

Porto, S.M.C., Arcidiacono, C., Anguzza, U. & Cascone, G. (2013). A computer vision-based system for the automatic detection of lying behaviour of dairy cows in free-stall barns. *Biosystems Engineering*, 115(2), 184–194. doi:http://doi.org/10.1016/j.biosystemseng.2013.03.002

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking*. Sebastopol, California, USA: O'Reilly Media.

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Rahardja, D.P., Toleng, A.L. & Lestari, V.S. (2011). Thermoregulation and water balance in fat-tailed sheep and Kacang goat under sunlight exposure and water restriction in a hot and dry area. *Animal: An International Journal of Animal Bioscience*, 5(10), 1587–1593. doi:http://doi.org/10.1017/S1751731111000577

Reimert, I., Bolhuis, J.E., Kemp, B. & Rodenburg, T. B. (2013). Indicators of positive and negative emotions and emotional contagion in pigs. *Physiology & Behavior*, 109, 42–50. doi:http://doi.org/10.1016/j.physbeh.2012.11.002

Reneau, J.K., Seykora, A.J., Heins, B.J., Endres, M.I., Farnsworth, R.J. & Bey, R.F. (2005). Association between hygiene scores and somatic cell scores in dairy cattle. *Journal of the American Veterinary Medical Association*, 227(8), 1297–1301. doi:10.2460/javma.2005.227.1297

Revelle, W. (2014). *Psych: Procedures for personality and psychological research (Version 1.4.8.)*. Northwestern University, Evanston, Illinois, USA. Retrieved from http://CRAN.R-project.org/package=psych

Rhodes, S., Bowie, D. & Hergenrather, K. (2003). Collecting behavioural data using the world wide web: considerations for researchers. *Journal of Epidemiology and Community Health*, 57(1), 68–73. doi:http://doi.org/10.1136/jech.57.1.68

Roche, J.R., Friggens, N.C., Kay, J.K., Fisher, M.W., Stafford, K.J. & Berry, D.P. (2009). Invited review: Body condition score and its association with dairy cow productivity, health, and welfare. *Journal of Dairy Science*, 92(12), 5769–5801. doi:http://doi.org/10.3168/jds.2009-2431

Rollin, B.E. (1993). Animal welfare, science, and value. *Journal of Agricultural and Environmental Ethics*, 6(Supplement 2), 44–50.

Rook, A.J. & Penning, P.D. (1991). Synchronisation of eating, ruminating and idling activity by grazing sheep. *Applied Animal Behaviour Science*, 32(2–3), 157–166. doi:http://doi.org/10.1016/S0168-1591(05)80039-5

Rossi, R. & Scharrer, E. (1992). Circadian patterns of drinking and eating in pygmy goats. *Physiology & Behavior*, 51(5), 895–897. doi:10.1016/0031-9384(92)90067-c

Rousing, T. & Waiblinger, S. (2004). Evaluation of on-farm methods for testing the human–animal relationship in dairy herds with cubicle loose housing systems—test–retest and inter-observer reliability and consistency to familiarity of test person. *Applied Animal Behaviour Science*, 85(3–4), 215–231. doi:http://doi.org/10.1016/j.applanim.2003.09.014

Rushen, J., Butterworth, A. & Swanson, J.C. (2011). Animal behavior and well-being symposium: Farm animal welfare assurance: Science and application. *Journal of Animal Science*, 89(4), 1219–1228. doi:http://doi.org/10.2527/jas.2010-3589

Russel, A.J.F., Doney, J. M. & Gunn, R.G. (1969). Subjective assessment of body fat in live sheep. *The Journal of Agricultural Science*, 72(03), 451–454. doi:http://doi.org/10.1017/S0021859600024874

Rutherford, K.M.D., Donald, R.D., Lawrence, A.B. & Wemelsfelder, F. (2012). Qualitative Behavioural Assessment of emotionality in pigs. *Applied Animal Behaviour Science*, 139(3-4), 218–224. doi:http://doi.org/10.1016/j.applanim.2012.04.004

Sahlu, T. & Goetsch, A. L. (2005). A foresight on goat research. *Small Ruminant Research*, 60(1–2), 7–12. doi:http://doi.org/10.1016/j.smallrumres.2005.06.002

Sakamoto, Y., Ishiguro, M. & Kitagawa, G. (1986). *Akaike information criterion statistics*. Dordrecht, The Netherlands: D. Reidel. Retrieved from http://amstat.tandfonline.com/doi/pdf/10.1080/01621459.1988.10478680#page=6

Sandgren, C., Lindberg, A. & Keeling, L. (2009). Using a national dairy database to identify herds with poor welfare. *Animal Welfare*, 18(4), 523–532.

Sandøe, P., Munksgaard, L., Badsgard, N.P. & Jensen, K.H. (1997). How to manage the management factor – assessing animal welfare at the farm level. In *Proceedings of the Fourth International Symposium on Livestock Farming Systems, Foulum, Denmark,* EAAP Publication, 89. Wageningen, The Netherlands:Wageningen Pers.

Sans, E.C.O., Federici, J.F., Dahlke, F. & Molento, C.F.M. (2014). Evaluation of free-range broilers using the Welfare Quality® protocol. *Revista Brasileira de Ciência Avícola*, 16(3), 297–306. doi:http://doi.org/10.1590/1516-635x1603297-306

Santucci, P., Branca, A., Napoleone, M., Bouche, R., Aumont, G., Poisot, F. & Alexandre, G. (1991). Body condition scoring of goats in extensive conditions. In P. Morand- Fehr (Ed.*), Goat Nutrition*. (pp. 240–255). Wageningen, The Netherlands: Wageningen Academic Publishers.

Santucci, P. & Maestrini, O. (1985). Body composition of dairy goats in extensive systems of production: Method of estimation. *Ann. Zoothech*, 34(4), 473–474. doi:10.1051/animres:19850409

Sarkar, M., Rahman, S., Sarker, B., Anisuzzaman, A., Begum, N. & Mondal, M. (2010). Epidemiology and pathology of ectoparasitic infestations in Black Bengal goats in Gaibandha and Mymensingh districts of Bangladesh. *Bangladesh Journal of Veterinary Medicine*, 8(1), 41–50. doi:http://doi.org/10.3329/bjvm.v8i1.7718

Schlageter-Tello, A., Bokkers, E.A.M., Koerkamp, P.W.G.G., Van Hertem, T., Viazzi, S., Romanini, C.E.B., Halachmi, I., Bahr, C., Berckmans, D. & Lokhorst, K. (2014). Manual and automatic locomotion scoring systems in dairy cows: A review. *Preventive Veterinary Medicine*, 116(1-2), 12-25. doi:http://doi.org/10.1016/j.prevetmed.2014.06.006

Schlumbohm, C. & Harmeyer, J. (2008). Twin-pregnancy increases susceptibility of ewes to hypoglycaemic stress and pregnancy toxaemia. *Research in Veterinary Science*, 84(2), 286–299. doi:http://doi.org/10.1016/j.rvsc.2007.05.001

Schnettler, B., Vidal, R., Silva, R., Vallejos, L. & Sepúlveda, N. (2009). Consumer willingness to pay for beef meat in a developing country: The effect of information regarding country of origin, price and animal handling prior to slaughter. *Food Quality and Preference*, 20(2), 156–165. doi:http://doi.org/10.1016/j.foodqual.2008.07.006

Schreiner, D.A. & Ruegg, P.L. (2003). Relationship between udder and leg hygiene scores and subclinical mastitis. *Journal of Dairy Science*, 86(11), 3460–3465. doi:http://doi.org/10.3168/jds.S0022-0302(03)73950-2

Scott, E.M., Nolan, A.M. & Fitzpatrick, J.L. (2001). Conceptual and methodological issues related to welfare assessment: A framework for measurement. *Acta Agriculturae Scandinavica, Section A - Animal Science*, 51(sup030), 5–10. doi:http://doi.org/10.1080/090647001316922983

Scott, E.M., Fitzpatrick, J.L., Nolan, A.M., Reid, J. & Wiseman, M.L. (2003). Evaluation of welfare state based on interpretation of multiple indices. *Animal Welfare*, 12(4), 457–468. Retrieved from http://www.velferdsprotokoller.org/uploads/5/4/3/3/5433758/evaluation_of_welfare_state_based_on_interpretation_of_multiple_indices.pdf

Scott, J. & Huskisson, E.C. (1976). Graphic representation of pain. *Pain*, 2(2), 175–184. doi:10.1016/0304-3959(76)90113-5

Scott, K., Chennells, D., Armstrong, D., Taylor, L., Gill, B. & Edwards, S. (2007). The welfare of finishing pigs under different housing and feeding systems: liquid versus dry feeding in fully-slatted and straw-based housing. *Animal Welfare*, 16(1), 53–62.

Sejian, V. & Srivastava, R.S. (2010). Effects of melatonin on adrenal cortical functions of indian goats under thermal stress. *Veterinary Medicine International*. doi:http://doi.org/10.4061/2010/348919

Shackleton, D.M. & Shank, C.C. (1984). A review of the social behavior of feral and wild sheep and goats. *Journal of Animal Science*, 58, 500–509. doi:10.2134/jas1984.582500x

Shamay, A., Shapiro, F., Barash, H., Bruckental, I. & Silanikove, N. (2000). Effect of dexamethasone on milk yield and composition in dairy cows. *Animal Research*, 49(4), 10. doi:http://doi.org/10.1051/animres:2000125

Shinde, A.K., Verma, D.L. & Singh, N.P. (2004). Social dominance-subordinate relationship in a flock of Marwari goats. *Indian Journal of Animal Sciences (India)*, 74. Retrieved from http://agris.fao.org/agris-search/search.do?recordID=IN2005000067

Shrout, P.E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research*, 7(3), 301–317. doi:10.1191/096228098672090967

Shrout, P.E. & Fleiss, J.L. (1979). Intraclass correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin*, 89(2), 420–428. doi:10.1037/0033-2909.86.2.420

Siebert, K., Langbein, J., Schön, P.C., Tuchscherer, A. & Puppe, B. (2011). Degree of social isolation affects behavioural and vocal response patterns in dwarf goats (Capra hircus). *Applied Animal Behaviour Science*, 131(1), 53–62. doi:http://doi.org/10.1016/j.applanim.2011.01.003

Silanikove, N. (2000). Effects of heat stress on the welfare of extensively managed domestic ruminants. *Livestock Science, Formerly: Livestock Production Science,* 67, 1–18. doi:http://doi.org/10.1016/S0301-6226(00)00162-7

Sim, J. & Wright, C.C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy*, 85(3), 257–268. Retrieved from http://www.physicaltherapyjournal.com/content/85/3/257.full.pdf+html

Sinn, R. & Rudenberg, P. (2008). *Raising goats for milk and meat*. Little Rock, Arkansas, USA: Heifer International.

Smith, M.C. & Sherman, D.M. (2009a). *Goat Medicine* (2nd ed.). Ames, Iowa, USA: John Wiley & Sons.

Smith, M.C. & Sherman, D.M. (2009b). Nutrition and metabolic diseases. In *Goat Medicine* (2nd ed.). 733–786. Ames, Iowa, USA: John Wiley & Sons.

Sørensen, J., Rousing, T., Møller, S., Bonde, M. & Hegelund, L. (2007). On-farm welfare assessment systems: what are the recording costs? *Animal Welfare*, 16(2), 237–239.

Spoolder, H., De Rosa, G., Hörning, B., Waiblinger, S. & Wemelsfelder, F. (2003). Integrating parameters to assess on-farm welfare. *Animal Welfare*, 12(4), 529–534.

Stafford, K. & Gregory, N. (2008). Implications of intensification of pastoral animal production on animal welfare. *New Zealand Veterinary Journal*, 56(6), 274–280. doi:http://doi.org/10.1080/00480169.2008.36847

Stelwagen, K. & Knight, C.H. (1997). Effect of unilateral once or twice daily milking of cows on milk yield and udder characteristics in early and late lactation. *The Journal of Dairy Research*, 64(4), 487–494. doi:10.1017/s0022029997002458

Stevens, S.S. (1946). On the theory of scales of measurement. *Science*, 103(2684), 677–680.

Stiglic, G., Kocbek, S., Pernek, I. & Kokol, P. (2012). Comprehensive decision tree models in bioinformatics. *PLoS ONE*, 7(3), e33812. doi:http://doi.org/10.1371/journal.pone.0033812

Streiner, D.L. (1993). A checklist for evaluating the usefulness of rating scales. *Canadian Journal of Psychiatry. Revue Canadienne De Psychiatrie*, 38(2), 140–148.

Streiner, D.L. & Norman, G.R. (2008). *Health measurement scales: A practical guide to their development and use*. (4th ed.). Oxford, UK: Oxford University Press.

Stricklin, W.R., Graves, H.B. & Wilson, L.L. (1979). Some theoretical and observed relationships of fixed and portable spacing behavior of animals. *Applied Animal Ethology*, 5(3), 201–214. doi:http://doi.org/10.1016/0304-3762(79)90056-7

Temple, D., Manteca, X., Dalmau, A. & Velarde, A. (2013). Assessment of test–retest reliability of animal-based measures on growing pig farms. *Livestock Science*, 151(1), 35–45. doi:http://doi.org/10.1016/j.livsci.2012.10.012

The MathWorks, Inc. (2012). *MATLAB and Statistics Toolbox Release*. Natick, Massachusetts, United States.

Thomsen, P.T., Munksgaard, L. & Tøgersen, F.A. (2008). Evaluation of a Lameness Scoring System for Dairy Cows. *Journal of Dairy Science*, 91(1), 119–126. doi:http://doi.org/10.3168/jds.2007-0496

Tiberio, M.L. & Diniz, F. (2014). Sheep and goat production in Portugal: A dynamic view. Modern Economy, 05(06), 703–722. doi:10.4236/me.2014.56066

Torres, A., Castro, N., Hernández-Castellano, L.E., Argüello, A. & Capote, J. (2013). Short communication: Effects of milking frequency on udder morphology, milk partitioning, and milk quality in 3 dairy goat breeds. *Journal of Dairy Science*, 96(2), 1071–1074. doi:http://doi.org/10.3168/jds.2012-5435

Toussaint, G. (1997). The housing of milk goats. *Livestock Production Science*, 49(2), 151–164. doi:http://doi.org/10.1016/S0301-6226(97)00011-0

Tuyttens, F., Sprenger, M., Van Nuffel, A., Maertens, W. & Van Dongen, S. (2009). Reliability of categorical versus continuous scoring of welfare indicators: lameness in cows as a case study. *Animal Welfare*, 18(4), 399–405.

Tuyttens, F.A.M., Vanhonacker, F., Van Poucke, E. & Verbeke, W. (2010). Quantitative verification of the correspondence between the Welfare Quality® operational definition of farm animal welfare and the opinion of Flemish farmers, citizens and vegetarians. *Livestock Science*, 131(1), 108–114. doi:http://doi.org/10.1016/j.livsci.2010.03.008

Tuyttens, F.A.M., de Graaf, S., Heerkens, J.L.T., Jacobs, L., Nalon, E., Ott, S., Stadig, L., Van Laer, E. & Ampe, B. (2014). Observer bias in animal behaviour research: can we believe what we score, if we score what we believe? *Animal Behaviour*, 90, 273–280. doi:http://doi.org/10.1016/j.anbehav.2014.02.007

Tu, Z. & Yuille, A.L. (2004). Shape Matching and Recognition – Using Generative Models and Informative Features. [electronic version]. In T. Pajdla & J. Matas (Eds.), *Computer Vision - ECCV 2004, 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004, Proceedings, Part IV*, pp. 195–209. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-24672-5_16

Vaccari, G., Panagiotidis, C.H., Acin, C., Peletto, S., Barillet, F., Acutis, P., Bossers, A., Langeveld, J., Van Keulen, L., Sklaviadis, T., Badiola, J., Andre´Oletti, O., Groschup, M., Agrimi, U., Foster, J. & Goldmann, W. (2009). State-of-the-art review of goat TSE in the European Union, with special emphasis on PRNP genetics and epidemiology . Veterinary Research, 40(5), 48. doi:10.1051/vetres/2009031

Van, D.T.T., Mui, N.T. & Ledin, I. (2007). Effect of group size on feed intake, aggressive behaviour and growth rate in goat kids and lambs. *Small Ruminant Research*, 72(2–3), 187–196. doi:http://doi.org/10.1016/j.smallrumres.2006.10.010

Van Hertem, T., Alchanatis, V., Antler, A., Maltz, E., Halachmi, I., Schlageter-Tello, A., Lokhorstb, C., Viazzi, S., Romanini, C.E.B., Pluk, A. Bahr C. & Berckmans, D. (2013). Comparison of segmentation algorithms for cow contour extraction from natural barn background in side view images. *Computers and Electronics in Agriculture*, 91, 65–74. doi:http://doi.org/10.1016/j.compag.2012.12.003

Van Nuffel, A., Sprenger, M., Tuyttens, F. & Maertens, W. (2009). Cow gait scores and kinematic gait data: can people see gait irregularities? *Animal Welfare*, *18*(4), 433–439.

Vanhonacker, F., Verbeke, W., Van Poucke, E. & Tuyttens, F.A.M. (2007). Segmentation based on consumers' perceived importance and attitude toward animal welfare. *International Journal of Sociology of Food and Agriculture*, 15(3).

Varner, M. A., & Johnson, B. H. (1983). Influence of adrenocorticotropin upon milk production, milk constituents, and endocrine measures of dairy cows. *Journal of Dairy Science*, 66(3), 458–465. doi:http://doi.org/10.3168/jds.S0022-0302(83)81814-1

Veissier, I., Butterworth, A., Bock, B. & Roe, E. (2008). European approaches to ensure good animal welfare. *Applied Animal Behaviour Science*, 113(4), 279–297. doi:http://doi.org/10.1016/j.applanim.2008.01.008

Veit, H.P., McCarthy, F., Friedericks, J., Cashin, M. & Angert, R. (1992). A survey of goat and cattle diseases in the Artibonite Valley, Haiti, West Indies. *Revue D'elevage et de Medecine Veterinaire Des Pays Tropicaux*, 46(1-2), 27–38.

Viazzi, S., Bahr, C., Schlageter-Tello, A., Van Hertem, T., Romanini, C. E. B., Pluk, A., Halachmi, I., Lokhorst, C. & Berckmans, D. (2013). Analysis of individual classification of lameness using automatic measurement of back posture in dairy cattle. *Journal of Dairy Science*, 96(1), 257–266. doi:http://doi.org/10.3168/jds.2012-5806

Viazzi, S., Bahr, C., Van Hertem, T., Schlageter-Tello, A., Romanini, C.E.B., Halachmi, I., Lokhorstc, C. & Berckmans, D. (2014). Comparison of a three-dimensional and two-dimensional camera system for automated measurement of back posture in dairy cows. *Computers and Electronics in Agriculture*, 100, 139–147. doi:http://doi.org/10.1016/j.compag.2013.11.005

Vieira, A., Brandão, S., Monteiro, A., Ajuda, I., & Stilwell, G. (in press). *Development and validation of a visual body condition scoring system for dairy goats with picture-based training*.

Viñuela-Fernández, I., Jones, E., Welsh, E.M. & Fleetwood-Walker, S.M. (2007). Pain mechanisms and their implication for the management of pain in farm and companion animals. *Veterinary Journal (London, England: 1997)*, 174(2), 227–239. doi:http://doi.org/10.1016/j.tvjl.2007.02.002

Waiblinger, S., Knierim, U. & 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(sup030), 73–77. doi:http://doi.org/10.1080/090647001316923108

Waiblinger, S. & Menke, C. (2003). Influence of sample size and experimenter on reliability of measures of avoidance distance in dairy cows. *Animal Welfare*, 12(4), 585–589.

Waiblinger, S., Boivin, X., Pedersen, V., Tosi, M.V., Janczak, A.M., Visser, E.K. & Jones, R. B. (2006). Assessing the human–animal relationship in farmed species: A critical review. *Applied Animal Behaviour Science*, 101(3–4), 185–242. doi:http://doi.org/10.1016/j.applanim.2006.02.001

Waiblinger, S., Schmied-Wagner, C., Mersmann, D. & Nordmann, E. (2011). Social behaviour and injuries in horned and hornless dairy goats. In *Proc. XV Congress of the International Society for Animal Hygiene,* pp. 421–422. Vienna, Austria.

Walter, S.D., Eliasziw, M. & Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statistics in Medicine*, 17(1), 101–110. doi:http://dx.doi.org/10.1002/(sici)1097-0258(19980115)17:1<101::aid-sim727>3.3.co;2-5

Wassink, G.J., Grogono-Thomas, R., Moore, L.J. & Green, L.E. (2003). Risk factors associated with the prevalence of footrot in sheep from 1999 to 2000. *The Veterinary Record*, 152(12), 351–358. doi:10.1136/vr.152.12.351

Weber, A., Salau, J., Haas, J.H., Junge, W., Bauer, U., Harms, J., Suhr, O., Schönrock, K., Rothfuß, H., Bieletzkic, S. & Thaller, G. (2014). Estimation of backfat thickness using extracted traits from an automatic 3D optical system in lactating Holstein-Friesian cows. *Livestock Science*, 165, 129–137. doi:http://doi.org/10.1016/j.livsci.2014.03.022

Webster, A.J. (2001). Farm animal welfare: the five freedoms and the free market. *Veterinary Journal (London, England: 1997)*, 161(3), 229–237. doi:http://doi.org/10.1053/tvjl.2000.0563

Webster, J. (2005). The assessment and implementation of animal welfare: theory into practice. *Revue Scientifique Et Technique (International Office of Epizootics)*, 24(2), 723–734. Retrieved from http://www.oie.int/doc/ged/D2705.PDF

Webster, J. (Ed.). (2011). *Management and Welfare of Farm Animals: The UFAW Farm Handbook*. (5th ed.). Wheathampstead, Herts, UK: Wiley-Blackwell.

Weir, J.P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research / National Strength & Conditioning Association*, 19(1), 231–240. doi:http://doi.org/10.1519/15184.1

Welfare Quality. (2009a). *Welfare Quality® Assessment Protocol for cattle*. Lelystad, The Netherlands: Welfare Quality® Consortium.

Welfare Quality. (2009b). *Welfare Quality® Assessment Protocol for pigs*. Lelystad, The Netherlands: Welfare Quality® Consortium.

Welsh, E.M., Gettinby, G. & Nolan, A.M. (1993). Comparison of a visual analogue scale and a numerical rating scale for assessment of lameness, using sheep as a model. *American Journal of Veterinary Research*, *54*(6), 976–983.

Wemelsfelder, F. (2007). How animals communicate quality of life: the qualitative assessment of behaviour. *Animal Welfare*, 16(2), 25–31.

Wemelsfelder, F., Millard, F., De Rosa, G. & Napolitano, F. (2009). Qualitative behaviour assessment. *Assessment of Animal Welfare Measures for Dairy Cows, Beef Bulls and Veal Calves. Welfare Quality® Reports*, *Vol. 11*. (pp. 215–224). Cardiff, UK.

Wewers, M.E. & Lowe, N.K. (1990). A critical review of visual analogue scales in the measurement of clinical phenomena. *Research in Nursing & Health*, 13(4), 227–236. doi:10.1002/nur.4770130405

Whay, H. (2007). The journey to animal welfare improvement. *Animal Welfare*, 16(2), 117–122.

Whay, H.R., Main, D.C.J., Green, L.E. & Webster, A J.F. (2003). Assessment of the welfare of dairy cattle using animal-based measurements: direct observations and investigation of farm records. *Veterinary Record*, 153(7), 197–202. doi:http://doi.org/10.1136/vr.153.7.197

Wildman, E.E., Jones, G.M., Wagner, P E., Boman, R.L., Troutt Jr., H.F. & Lesch, T.N. (1982). A Dairy Cow Body Condition Scoring System and Its Relationship to Selected Production Characteristics. *Journal of Dairy Science*, 65(3), 495–501. doi:http://doi.org/10.3168/jds.S0022-0302(82)82223-6

Willeberg, P. (1991). Animal welfare studies: Epidemiological considerations. In M.V. Thrusfield (Ed.), *Proceedings of the Society for Veterinary Epidemiology and Preventive Medicine, London, 17-19, April*, pp. 76-82.

Winckler, C. (2006). On-farm welfare assessment in cattle: From basic concepts to feasible assessment systems. In *Proc. XXIV World Buiatrics Congress, Nice, France*, pp.493–500. Wilmington, USA: Gloyd Group. Retrieved from http://www.ivis.org/proceedings/wbc/wbc2006/Winckler.pdf?LA=1

Winckler, C., Capdeville, J., Gebresenbet, G., Hörning, B., Roiha, U., Tosi, M. & Waiblinger, S. (2003). Selection of parameters for on-farm welfare-assessment protocols in cattle and buffalo. *Animal Welfare*, 12(4), 619–624.

Winckler, C. & Willen, S. (2001). The Reliability and Repeatability of a Lameness Scoring System for Use as an Indicator of Welfare in Dairy Cattle. *Acta Agriculturae Scandinavica, Section A - Animal Science*, 51(sup030), 103–107. doi:http://doi.org/10.1080/090647001316923162

Windschnurer, I., Schmied, C., Boivin, X. & Waiblinger, S. (2009). Assessment of human–animal relationships in dairy cows. In B. Forkman & L. Keeling (Eds.), *Assessment of Animal Welfare Measures for Dairy Cows, Beef Bulls and Veal Calves. Welfare Quality® Reports, Vol. 11.* (pp. 137–152). Cardiff, UK

Winter, A.C. (2008). Lameness in sheep. *Small Ruminant Research*, 76(1), 149–153. doi:http://doi.org/10.1016/j.smallrumres.2007.12.008

Winter, A.C. (2011). Treatment and control of hoof disorders in sheep and goats. *The Veterinary Clinics of North America. Food Animal Practice*, 27(1), 187–192. doi:http://doi.org/10.1016/j.cvfa.2010.10.018

Winter, M., Fry, C. & Carruthers, S.P. (1998). European agricultural policy and farm animal welfare. *Food Policy*, 23(3–4), 305–323. doi:http://doi.org/10.1016/S0306-9192(98)00036-0

Wooldrigde, J.M. (2008). Multiple regression analysis with qualitative information: binary (or dummy) variables. In *Introductory Econometrics - A Modern approach.* (4th ed.). Mason, USA: Cengage Learning.

Yeates, J.W. & Main, D.C.J. (2008). Assessment of positive welfare: A review. *The Veterinary Journal*, 175(3), 293–300. doi:http://doi.org/10.1016/j.tvjl.2007.05.009