Wageningen UR Livestock Research

Partner in livestock innovations

Report 533

Simplifying the Welfare Quality assessment protocol for broilers

November 2011

LIVESTOCK RESEARCH
WAGENINGEN UR

LIVESTOCK RESEARCH
WAGENINGEN UR

Colophon

Publisher
Wageningen UR Livestock Research
P.O. Box 65, 8200 AB Lelystad
Telephone +31 320 - 238238
Fax +31 320 - 238050
E-mail info.livestockresearch@wur.nl
Internet http://www.livestockresearch.wur.nl

Editing
Communication Services

Wageningen UR Livestock Research and Central Veterinary Institute of Wageningen UR, both part of Stichting Dienst Landbouwkundig Onderzoek (DLO Foundation), together with the Department of Animal Sciences of Wageningen University comprises the Animal Sciences Group of Wageningen UR (University & Research centre).

Single numbers can be obtained from the website.

Abstract
This report describes the results of a study to simplify the Welfare Quality® assessment protocol for broiler chickens.

Keywords
Broilers, welfare assessment

Reference
ISSN 1570 - 8616

Author(s)
Ingrid C. de Jong
Tomas Perez Moya
Henk Gunnink
Hans van den Heuvel
Vincent A. Hindle
Monique Mul
Kees van Reenen

Title
Simplifying the Welfare Quality assessment protocol for broilers

Report 533

Report 533

Simplifying the Welfare Quality assessment protocol for broilers

Vereenvoudiging van het Welfare Quality protocol voor het meten van welzijn bij vleeskuikens

Ingrid C. de Jong
Tomas Perez Moya
Henk Gunnink
Hans van den Heuvel
Vincent A. Hindle
Monique Mul
Kees van Reenen

November 2011

This study was financed by the Ministry of Economic Affairs, Agriculture and Innovation.

Ministerie van Economische Zaken,
Landbouw en Innovatie

Preface

Between 2004 and 2009 the Welfare Quality® project developed a method to assess animal welfare on cattle, pig and poultry farms. The resulting Welfare Quality® assessment protocols, published in 2009, provide a detailed account of the measures which need to be taken, and how these can be combined into a single overall statement about the level of welfare on the farm under assessment. The method has attracted a lot of interest from European and national policy makers, NGO's and the farming community in general, but has to date not been adopted in any commercial scheme nor is it used by farmers to improve animal welfare on their farm. The main drawback would appear to be the amount of time required to perform the assessment. In 2010 the Dutch ministry of Economic Affairs, Agriculture and Innovation commissioned Wageningen UR Livestock Research to conduct a series of studies aimed at simplifying the original three protocols. In collaboration with former Welfare Quality® partners and the Dutch broiler sector, farm visits were organised and a considerable amount of data was collected between February and July 2011. This report describes the results of the collection and analyses of the data, together with options for replacing time consuming measures with simpler alternatives. The results will be presented to the international Welfare Quality® Network, which is working towards further improvement of the protocols. They will also be recommended to the Dutch Ministry, for future implementation to improve on-farm animal welfare in collaboration with the Dutch broiler sector.

Without the co-opeeration of these three stakeholder groups; the Dutch Ministry, representatives of the broiler industry and the Welfare Quality® Network, this work would not have been possible. On behalf of the project team I would like to thank Bart Crijns, Amanda Manten, and Léon Arnts (Ministry of Economic Affairs, Agriculture and Innovation), Peter Vesseur (Nepluvi), Willem Tel (2 Sisters Storteboom), René Welpelo (De Kuikenaer BV) and Andy Butterworth (Welfare Quality® Network) for their contributions to this work.

Paul Vriesekoop
Director Wageningen UR Livestock Research

Summary

The European Welfare Quality® project developed standardized animal welfare assessment methods for different categories of farm animals, e.g. broiler chickens and laying hens, sows, growing pigs, veal calves and dairy cattle. Measurements and the integration of different individual measures into a final on-farm score for broiler chickens has been described in the Welfare Quality® assessment protocol for poultry (Welfare Quality®, 2009). One of the key characteristics of the Welfare Quality® assessment protocols is that it places more focus on animal based measures (i.e. injuries or behaviour) than on design or management criteria (i.e. pen size). Dutch stakeholders have expressed their interest in the assessment protocols for different types of farm animals, but have also suggested that a reduction in performance time may improve the practical applicability of the assessment protocol and improve the probability of adoption of the welfare assessment protocol in practice. The aim of the current project was to determine whether or not there is scope for simplification of the broiler assessment protocol by reduction in performance time. In addition, the Ministry of Economic Affairs, Agriculture and Innovation along with interested parties within the poultry sector have emphasized the requirement for robust testing of the assessment protocol prior to consideration for wider implementation.

According to the standard broiler assessment protocol, data were collected from 180 broiler flocks, of different breeds and housed under different conditions. The majority of the data were collected in 2011 in Dutch broiler flocks, but more recent Belgian data alongside data from UK, Italian and Dutch farms were included in the analysis. Slaughter plant visits were performed for 150 flocks in addition to the on-farm measurements as described in the full assessment protocol. Assessors were thoroughly trained prior to visiting. For the data collected in 2011, a personal digital assistant was used for scoring and all data were subsequently stored in an access database until required for analysis. Statistical analysis was performed as follows: a) data exploration, i.e. studying variability within the different data sets and analysing differences between standard rearing systems using fast growing broilers (housed at a stocking density of about 42 kg/m^2), and alternative rearing systems using slower growing genotypes (housed at lower stocking densities); b) analysis of correlations between animal-based measurements (on-farm as well as at the slaughter plant); c) calculation of end scores for all flocks, based on the calculations described in the full assessment protocol; d) analysis of possible strategies for simplification of the broiler assessment protocol.

Large variability between flocks was found for almost all measurements. Analysis of differences between flocks from standard rearing systems and flocks from alternative systems indicated large differences between rearing systems. In general, birds from flocks reared in alternative systems (alternative, slower growing genotype) displayed fewer incidences of contact dermatitis (foot pad dermatitis, hock burn and breast blisters), better mobility, better scores for cleanliness, less panting and more positive scores for qualitative behaviour assessment compared to birds reared in standard systems (fast growing genotypes). Birds from alternative systems however scored lower in the touch test compared with birds from standard systems.

Analysis of correlations between animal based measurements showed no high correlations for on-farm measurements ($r<0.7$). The highest correlation that was of interest with regard to further analysis for simplification was that between severe hock burn and high gait scores ($r=0.615$ overall , $r=0.44$ for standard and alternative flocks separately). High correlations ($r>0.7$) were found for foot pad dermatitis measured on- farm and at the slaughter plant. Therefore, a potential second strategy for simplification, i.e. replacing on-farm measures with slaughter plant measures, was analysed. In this simplification strategy, clinical scores (foot pad dermatitis, hock burn and cleanliness) and gait score were predicted from slaughter plant measurements (foot pad dermatitis and hock burn).

Calculations of end scores for flocks based on the full assessment protocol showed that nearly all flocks were classified in the same category, i.e. acceptable, despite high between-flock variability for individual measurements. The fact that Welfare Quality® methodology does not allow compensation between criteria and principles may be of influence on this.

Two strategies for simplification, i.e. predicting gait scores from hock burn scores and predicting on-farm measures from slaughter plant measures (i.e. predicting gait score, cleanliness, foot pad dermatitis and hock burn on-farm from foot pad dermatitis and hock burn measured at the slaughter plant), were analysed. Results are shown in relation to the golden standard, i.e. the full assessment protocol. Analysis of simplification strategies showed that there was in general close agreement on the

level of flock score, as well as on the level of principle and criterion scores (for principles and criteria affected by simplification). In addition, there was generally a high correlation between the golden standard and the simplified model on principle and criterion level. Where only a few farms were involved in a certain category, a larger confidence interval was found indicating that further study is required prior to drawing any final conclusions with respect to the simplification strategies.

It was concluded that the strategies for simplification of the broiler assessment protocol as analysed with the data collected during the current project are encouraging in terms of agreement with the golden standard for flock score, principle level and criterion level. Both strategies for simplification of the broiler assessment protocol appear promising regarding the potential for reduction in performance time essential for improvement of the probability of acceptance for implementation in practice. It is strongly advised to validate the results of the data-based simplification strategies in a further study, preferably in flocks that are more widely distributed over the different categories, before implementation of the simplification strategies in practice.

Samenvatting

Binnen het Europese Welfare Quality® project is een gestandaardiseerde methode ontwikkeld om het welzijn vast te stellen voor verschillende soorten landbouwhuisdieren, zoals bijvoorbeeld vleeskuikens en leghennen, zeugen en vleesvarkens, vleeskalveren en melkvee. Voor vleeskuikens zijn de afzonderlijke variabelen en de integratie van de verschillende variabelen tot een eindscore voor een koppel beschreven in het 'Welfare Quality® assessment protocol for poultry' (Welfare Quality®, 2009). Eén van de belangrijkste kenmerken van de Welfare Quality® protocollen is dat ze zoveel mogelijk uitgaan van dierkenmerken (zoals verwondingen of gedrag) in plaats van omgevingskenmerken (zoals afmetingen van de stal). Belanghebbenden uit de verschillende sectoren in Nederland hebben aangegeven zeer geïnteresseerd te zijn in de protocollen door Welfare Quality® ontwikkeld, maar ze hebben ook aangegeven dat het verminderen van de tijd benodigd voor een volledige beoordeling van een bedrijf of koppel de praktische toepasbaarheid sterk kan vergroten, en dus het omarmen van het protocol door de sector. Het doel van het hier beschreven project was om te bepalen of er mogelijkheden zijn om het protocol zoals omschreven voor vleeskuikens verder te vereenvoudigen zodat de benodigd tijd om alle metingen uit te voeren kan worden teruggebracht. Dit bevordert de praktische toepasbaarheid van het protocol. Daarnaast hebben het Ministerie van Economische Zaken, Landbouw en Innovatie en verschillende ketenpartners aangegeven dat het op robuuste wijze testen van het protocol noodzakelijk is voordat overgegaan kan worden tot implementatie in de praktijk.

Van 180 koppels zijn data verzameld volgens het volledige protocol zoals beschreven is voor vleeskuikens. Deze koppels bevatten verschillende typen dieren en waren gehuisvest onder verschillende condities. De meerderheid van de data is verzameld in 2011 bij Nederlandse koppels vleeskuikens, maar ook recente data van Belgische koppels, en data van koppels uit het Verenigd Koninkrijk, Italië en Nederland verzameld in 2008 werden meegenomen in de analyse. Naast metingen op het primaire bedrijf zijn ook bezoeken gebracht aan het slachthuis (bij 150 koppels) zoals beschreven in het protocol. De koppels werden beoordeeld door getrainde waarnemers. Voor de koppels bezocht in 2011 werden de data vastgelegd met behulp van een handcomputer en vervolgens in een database opgeslagen tot het moment van verdere analyse. Statistische analyse bestond uit de volgende stappen: a) data exploratie, d.w.z. nader bestuderen van de variabiliteit van de metingen en analyse van verschillen tussen 'standaard' koppels (reguliere, snel groeiende typen vleeskuikens meestal gehouden op een bezetting van rond 42 kg/m^2) en 'alternatieve' koppels (langzamer groeiende vleeskuikens, meestal gehouden bij een lagere bezetting, in systemen met (buiten)uitloop, daglicht en omgevingsverrijking); b) bepalen van de correlatie tussen dierkenmerken (gemeten op het primaire bedrijf en op het slachthuis), c) bepalen van de eindscore voor alle koppels, gebaseerd op de berekeningen zoals beschreven in het volledige protocol, d) analyseren van mogelijke strategieën om het protocol voor de vleeskuikens te vereenvoudigen.

Voor de meeste variabelen was er grote variatie tussen de individuele koppels. Analyse van de verschillen tussen standaard en alternatieve koppels liet zien dat deze koppels verschilden voor de meeste gemeten variabelen. In het algemeen hadden koppels van alternatieve systemen minder last van contact dermatitis (voetzoollaesies, hakdermatitis en borstirritatie), hadden ze minder problemen met lopen, waren ze minder bevuild, vertoonden ze minder hijggedrag en kregen ze meer positieve scores in de 'qualitative behaviour assessment' vergeleken met standaard koppels. Koppels in alternatieve systemen scoorden lager in de 'touch test'.

Analyse van de correlaties tussen de verschillende dierkenmerken liet zien dat voor de kenmerken gemeten op het bedrijf nooit zeer hoge correlaties ($r>0.7$) werden gevonden. De hoogste correlatie die mogelijkheden gaf voor verdere vereenvoudiging van het protocol was de correlatie tussen ernstige hakdermatitis en een slechte 'gait score' (dieren die slecht kunnen lopen; $r=0.615$ voor alle koppels, $r=0.44$ voor standaard en alternatieve koppels afzonderlijk). Hoge correlaties ($r>0.7$) werden wel gevonden tussen metingen van voetzoollaesies op het bedrijf en aan de slachtlijn. Een mogelijke andere strategie voor vereenvoudiging, namelijk het vervangen van metingen op het bedrijf door metingen aan de slachtlijn, is daarom ook verder doorgerekend. Hierbij is gerekend met het voorspellen van de klinische scores (bevuiling, voetzoollaesies en hakdermatitis) en de gait score op het bedrijf uit de waarnemingen van hakdermatitis en voetzoollaesies aan de slachtlijn.

Berekeningen van eindscores voor koppels volgens het volledige protocol lieten zien dat bijna alle koppels in dezelfde categorie (‘acceptable’) eindigden, ondanks de grote variatie in uitkomsten van de metingen tussen de koppels. Een oorzaak hiervoor kan zijn dat de methodiek van Welfare Quality® voor het berekenen van eindscores geen compensatie toelaat op niveau van criteria en principes.

Twee potentiële methoden voor vereenvoudiging, namelijk het voorspellen van de ‘gait score’ uit de scores voor hakdermatitis op het primaire bedrijf en het voorspellen van dierkenmerken op het bedrijf (gait score, bevuiling, hakdermatitis en voetzoollaesies) uit dierkenmerken (hakdermatitis en voetzoollaesies) gemeten aan de slachtlijn, zijn geanalyseerd. De resultaten zijn weergegeven in termen van overeenkomst met de gouden standaard (het volledige protocol). Analyse van de vereenvoudigingsstrategieën liet zien dat in het algemeen de overeenkomst tussen de gouden standaard en het vereenvoudigd protocol hoog was, zowel voor de eindscores voor koppels, als voor de principes en criteria (alleen de principes en criteria werden beïnvloed door de vereenvoudiging). Bovendien werd er in het algemeen een hoge correlatie gevonden tussen de uitkomst van de gouden standaard en de uitkomst van het vereenvoudigd protocol op niveau van de principes en criteria. Wanneer er sprake was van een laag aantal bedrijven in een bepaalde categorie was er soms een groot betrouwbaarheidsinterval voor de sensitiviteit en de specificiteit. Dit betekent dat verder onderzoek nodig is voordat we definitieve conclusies kunnen trekken over de vereenvoudigingsstrategieën.

Concluderend kan gesteld worden dat de beide strategieën voor vereenvoudiging van het meetprotocol voor vleeskuikens zoals geanalyseerd op basis van de data die in dit project zijn verzameld, kansrijk lijken. Dit is gebaseerd op de overeenkomst met de gouden standaard op basis van de eindscore van de koppels en op de scores voor criteria en principes. Beide vereenvoudigingsstrategieën zijn ook kansrijk in termen van het terugbrengen van de tijd benodigd voor de beoordeling van een koppel en kunnen dus de implementatie van de vleeskuikenmonitor in de praktijk faciliteren. Wij adviseren om de data-gedreven vereenvoudiging van het vleeskuikenprotocol verder te valideren, bij voorkeur in koppels die goed verdeeld zijn over de verschillende categorieën, voordat een vereenvoudigd protocol in de praktijk wordt geïmplementeerd.

Table of contents

1 Introduction

1.1 Welfare Quality® assessment protocols

The European Welfare Quality® project developed standardized assessment methods of animal welfare for different categories of farm animals, e.g. broiler chickens and laying hens, sows, growing pigs, veal calves and dairy cattle. For broiler chickens, the measurements and the integration of different individual measures into a final score for a flock has been described in the Welfare Quality® assessment protocol for poultry (Welfare Quality®, 2009). One of the key characteristics of the Welfare Quality® assessment protocols is that it focuses more on animal based measures (such as injuries or behaviour) than on design or management criteria (such as pen size) (Blokhuis et al., 2010).

Welfare Quality® assessment protocols are based on the approach that welfare is a multidimensional concept, that comprises both physical as well as mental health. Within the Welfare Quality® project for all species the same framework has been used to measure welfare of animals. Species-specific measures of welfare, e.g. for broilers the number of foot pad lesions, are integrated into a score for twelve independent welfare criteria. These criteria are integrated into four principle scores which are subsequently integrated into an overall score for a flock. These twelve welfare criteria and the four principles are listed in table 1.

Table 1. The principles and criteria that are the basis for the Welfare Quality® assessment protocols (Blokhuis et al., 2010).

Welfare Quality® Principles	Welfare Quality® Criteria
Good feeding	1 Absence of prolonged hunger
	2 Absence of prolonged thirst
Good housing	3 Comfort around resting
	4 Thermal comfort
	5 Ease of movement
Good health	6 Absence of injuries
	7 Absence of disease
	8 Absence of pain induced by management procedures
Appropriate behaviour	9 Expression of social behaviour
	10 Expression of other behaviour
	11 Good human-animal relationship
	12 Positive emotional state

The welfare assessment protocol for broilers (Welfare Quality®, 2009) describes measures indicative of broiler welfare on-farm, as well as measures indicative of broiler welfare during transport and slaughter. For the assessment of broiler welfare on-farm, the calculations for the integration into an overall flock score are available. This needs to be developed for the measurements of broiler welfare during transport and slaughter. In this report we focus on the measures on-farm and slaughter plant measures that are indicative of broiler welfare on-farm. Measures indicative of welfare during transport and slaughter have not been assessed during this study.
On average, the broiler welfare assessment protocol on-farm takes about 3-4 hours per flock (Welfare Quality®, 2009) . However, the assessment protocol also provides the possibility to collect data concerning some of the measures (foot pad dermatitis and hock burn) at the slaughter plant. These may replace the necessity for collection of these data on-farm, but it has not yet been established how slaughter plant measures relate to on-farm measures. If slaughter plant and on-farm measures are shown to be closely related, replacement of on-farm assessments with slaughter plant measures will reduce performance time considerably.
Dutch stakeholders expressed their interest in the assessment protocols for the different types of farm animals, but they also emphasized that a reduction in performance time of assessment will improve the probability for practical applicability of the assessment protocol and encourage adoption of a welfare assessment protocol by stakeholders (Manten and De Jong, 2011). Although on-farm the broiler welfare assessment protocol can be performed in a reasonably short time, it remains important to study the possibilities to improve efficiency of performance, without compromising quality of measurement.

In addition to the request from stakeholders to reduce assessment time, there was little practical experience with the methodology. It was expressed by the Ministry of Economic Affairs, Agriculture and Innovation as well by other interested parties within the poultry sector that there is a necessity to perform robust testing on a small scale prior to consideration for wider implementation.

1.2 Aim of the project

The aim of the project was to determine the possibilities for simplification of the broiler assessment protocol in order to reduce performance time. This will improve the potential for practical applicability of the assessment protocol. In order to determine whether or not a simplification of the broiler assessment protocol is possible without compromising the outcome of the protocol, the welfare assessment protocol for broilers (Welfare Quality®, 2009) was applied to a large number of flocks. A minimum of 150 broiler flocks, preferably differing in housing conditions and breeds were estimated to be required to establish a reliable statistical analysis. Two possible ways of simplifying the broiler welfare assessment protocol were analysed:

a) Use of predictors. If there are significant and meaningful correlations between individual measures in the assessment protocol, the value of one measure can be predicted by using the value of another measure. A simplified assessment protocol may in that case consist of a limited set of measures, that can be used to predict the value of other measures. The structure of the assessment protocol (measures – criterion scores – principle scores) remains unchanged. The final outcome of a possible simplified protocol will be compared with the final outcome of the full assessment protocol (the 'golden standard');
b) Use of data sampled at the slaughter plant. The broiler assessment protocol already prescribes that some measures can be performed either at the slaughter plant or on- farm (foot pad dermatitis and hock burn). However, the relationship between these measures at the plant, and on-farm, remains unknown. Currently, in Dutch slaughterhouses, hock burn and in the future foot pad dermatitis will be measured at the slaughter plant for flocks housed at a stocking density of 42 kg/m^2 (Anonymus, 2009). If on-farm measures can be replaced by slaughter plant data, this will save time required for on-farm assessment. As described under a) relationships between on-farm measures and slaughter plant measures will be determined and the outcome of a possible simplified protocol will be compared with the final outcome of the full assessment protocol (the 'golden standard').

1.3 Content of this report

In the current project a large number of data of broiler flocks were collected. This report provides a general overview of the data, in terms of variability in individual measures and differences between farm/bird types (standard systems with fast growing broilers and alternative systems with slower growing breeds) and results of the analysis of possibilities for simplification of the broiler assessment protocol are also presented. In addition, adaptations in the assessment protocol due to practical constraints and practical experiences are described.
The large number of data collected also enabled a risk factor analysis, including not only animal-based measures but also factors such as stocking density and litter quality . However, due to time constraints this analysis is not described in the current report.

2 Methods

2.1 Broiler flocks

A total of 180 flocks were assessed for this project. A flock being defined as birds in a single house at a particular farm. Birds were reared under different conditions, ranging from standard rearing systems (animals housed exclusively indoors and fed diets that allow them to reach the target weight (2-2.5 kg) in 35 days at stocking densities beyond 42 kg/m², using so-called fast growing breeds), to birds reared in systems allowing more space per bird (approximately 20-32 kg/m²), with target weights achieved over longer periods (50-81 days), using natural lighting schemes, daylight in the house with outdoor access or a covered range ('winter garden'), occasional use of enrichment and using slower growing breeds (alternative systems). Data were provided from 4 different countries: 140 Dutch flocks: (122 assessed in 2011 and 18 flocks assessed in 2008), ten British flocks (assessed in 2008), 18 Italian flocks (assessed in 2008) and 12 Belgian flocks (assessed in 2011). As one farm could have more than one broiler house, more flocks could be assessed at a particular farm.

2.2 Visits

Assessments were performed between March and June 2008 or between March and June 2011. A total of 25 assessors (16 in the Netherlands, 3 in Italy, 2 in Belgium and 1 in United Kingdom), all trained in the theory and practice by experienced persons, performed the data collection. Assessment of the flocks was performed in the period between one and five days prior to slaughter according to the Welfare Quality® broiler assessment protocol (Welfare Quality®, 2009). If possible, additional measures were carried out at the slaughter plant. A total of 150 flocks was also assessed at the slaughter plant. Reasons for not assessing all flocks at slaughter were that assessors were not always allowed in the slaughter house or a last-minute change in the planning of the time or place of slaughter not allowing sufficient time for the assessor to react to these changes.

2.3 Training of assessors

All assessors received training to perform the measures as described in the assessment protocol. Training consisted of one day of theory, followed by one day on-farm training and, for slaughter plant assessors, half a day training at a slaughter plant. Two weeks later, an additional half day training was given using video clips. In addition, on this day assessors were trained in the use of the personal digital assistant (pda). Theoretical training was given by one trainer from the Welfare Quality® consortium. Training in practice was given by either a trainer from the Welfare Quality® consortium, or experienced researchers from WUR-LR that were previously trained and well experienced in the Welfare Quality® protocol for broilers. In addition to training, assessors were examined for the different measures and the agreement with an experienced assessor, the 'golden standard', was determined. At least 75% agreement between the assessor and the 'golden standard' was required before the assessors were sent to farms and slaughter plants. For the first visit two assessors were sent to a particular farm and performed the assessment together.

2.4 Measures

A brief description of the measures involved is given in the following pragraphs. For more detailed descriptions we refer to the Welfare Quality® broiler assessment protocol (Welfare Quality®, 2009). Specialised software to register the data in a personal digital assistant (pda) was developed. This enabled downloading of the data into an access database. Assessors had to download the data as soon as possible after their visit. Data were subsequently checked by a researcher and any missing or incomplete data were supplemented or corrected, after consultation with the assessor and/or the farmer.

2.4.1 On-farm measures

2.4.1.1 General questions

Each visit started with a short questionnaire for the farmer. The following information was registered: name and address of the farmer, the number of the house assessed, number of birds on site, number of birds in the house at placement, number of actual birds in the house (at time of visit), date of placement, age of the birds, age of the parent stock, breed, average actual bird weight, name of the hatchery of origin, name of the slaughter plant, dimensions of the house, drinker type(s) and number of drinkers, percentage mortality or number of animals that had died since placement, percentage culling or number of animals that were culled.
At the request of one of the participating slaughter plants, the type of litter and the type of heating in the house were also registered.
The name of the assessor was registered, the starting time of the visit when entering the house, and the time at the end of the visit when all assessments had been completed.

2.4.1.2 Absence of prolonged hunger

This measure is not scored on-farm, but only at the slaughter plant.

2.4.1.3 Absence of prolonged thirst

The type of drinker as well as the number of drinkers in the house was noted. From this the number of birds per type of drinker could be calculated.

2.4.1.4 Comfort around resting

2.4.1.4.1 Plumage cleanliness

A sample of at least 100 birds at 10 locations in the house was scored. Scoring of plumage cleanliness was performed on the same birds as for foot pad dermatitis and hock burn. Locations within the house were selected randomly. Birds were penned in a catching pen and all birds in the catching pen were scored. They received a score from 0 (feathers clean) to 3 (feathers very dirty), as described in the broiler assessment protocol (Welfare Quality®, 2009).

2.4.1.4.2 Litter quality

The quality of the bedding in the house was assessed at 6 locations, these were the same as for birds selected for gait scoring (see below). Classification ranged from 1 (completely dry and flaky, moves easily with the foot) to 5 (sticks to boot once the cap or compacted crust is broken) (Welfare Quality®, 2009). According to the protocol the scores range from 0-4, but we used 1-5 as in the pda software a score of 0 represented ‘not scored’.

2.4.1.4.3 Dust

A dust sheet test was performed by placing a black painted A6 size aluminium layer in the house at the start of the visit, on a location not in the vicinity of any machinery. At the end of the visit the amount of dust on the tray was scored on a scale from 1 (no evidence of dust) to 5 (colour of tray not visible) (Welfare Quality®, 2009).

2.4.1.5 Thermal comfort

2.4.1.5.1 Panting and huddling

The percentages of birds showing either panting (breathing rapidly in short gasps, indicator of heat stress) or huddling (birds grouping together in ‘clumps’, indicator of the environment being too cold) were estimated at 5 randomly chosen, locations distributed throughout the house (Welfare Quality®, 2009). This was performed along with observation of birds for the qualitative behaviour assessment (see below), at the start of the assessment.

2.4.1.6 Ease of movement

2.4.1.6.1 Stocking density

The stocking density was calculated according to the dimensions of the house, bird weight and the number of birds as provided by the farmer.

2.4.1.7 Absence of injuries

2.4.1.7.1 Lameness (gait score)

At least 150 birds, from at least 6 randomly chosen locations in the house, were penned in a catching pen and their gait was scored by letting them walk out of the pen one by one. Birds were classified according to six categories, ranging from 0 (normal, dextrous and agile) to 5 (incapable of walking) (Kestin et al., 1992; Welfare Quality®, 2009).

2.4.1.7.2 Hock burn

At least 100 birds were scored at 10 locations in the house, as described for cleanliness. The hocks of the birds were inspected and given a score ranging from 0 (no hock burn present) to 4 (evidence of hock burn, severe, dark coloured lesion of considerable size (Welfare Quality®, 2009). Both hocks were inspected and a score was given according to the most severely injured hock.

2.4.1.7.3 Foot pad dermatitis

Foot pads from the same birds scored for cleanliness and hock burn, were inspected for foot pad dermatitis and scored using the Bristol Foot Burn scale, ranging from 0 (no foot pad lesion) to 4 (severe lesion, large area of the feet injured) (Welfare Quality®, 2009). Both feet were scored and a score was given according to the most severely injured foot.

2.4.1.8 Absence of disease

2.4.1.8.1 Mortality and culls

The number of birds that had died and had been culled since placement in the house were registered according to the information provided by the farmer. Where culling had not been registered separately, total mortality was scored.

2.4.1.9 Expression of other behaviours

2.4.1.9.1 Cover on range and number of birds outdoors

For free range or extensive systems with an outdoor range, the proportion of birds in the outdoor range as well as the proportion of range cover (trees, maize, shelters) was estimated (Welfare Quality®, 2009).

2.4.1.10 Good human-animal relationship

2.4.1.10.1 Touch test

The assessor approached a group of at least 3 birds in the litter area, squatted for 10 seconds and then counted the number of birds within arm length (within 1 meter of the observer). This was repeated 21 times at different locations in the house (Welfare Quality®, 2009).

2.4.1.11 Positive emotional state

2.4.1.11.1 Qualitative behaviour assessment (QBA)

Between one and eight observation points in the house were selected (dependent on the size of the house) and the flock was observed for 20 minutes. The assessor left the house and scored the behaviour of the flock according to the terms as described in the broiler assessment protocol (Welfare Quality®, 2009).

2.4.2 Slaughter plant measures

2.4.2.1 Absence of prolonged hunger

2.4.2.1.1 Emaciated birds

The number of emaciated birds should be provided by the slaughterhouse. It should be noticed that Dutch slaughterhouses do not register the number of emaciated birds, they only indicate the major reason of rejections.

2.4.2.2 Absence of injuries

2.4.2.2.1 Breast blisters

Birds were observed for 5 minutes on the slaughter line. This was performed after approximately half of the flock had passed along the slaughter line. It was scored as a breast blister (score 1) or no breast blister (score 0). The broiler assessment protocol (Welfare Quality®, 2009) does not show any examples of breast blisters but an example of an abscess. Therefore, new pictures for classification of birds were provided. As real breast blisters are only seldom seen in broilers, it was decided not only to score real breast blisters, but also breast irritation (discolouration of the breast, brown colour, caused by wet and dirty litter). Pictures used for classification of the birds are shown in figure 1. A bird received a score 1 when any sign of breast irritation was observed, independent of size.

Figure 1. Examples of breast blisters (upper pictures) and breast irritation (lower picture).

2.4.2.2.2 Hock burn

Hock burn was assessed as the birds were passing on the slaughter line. When approximately 1/3 and 2/3 of the flock had passed the slaughter line, observations were performed for 5 minutes. Birds received a score 1 as a dark colouration of the hock of at least 0.5 cm^2 was observed (this could be either one big spot, or several smaller spots). In any other case birds received a score 0. This differs from the scoring as described in the broiler assessment protocol (Welfare Quality®, 2009). However, it is not possible to score hocks into 5 classes when the birds are passing on the slaughter line with high speed (120 birds per minute or more). Therefore, the methodology was adjusted.

2.4.2.2.3 Foot pad dermatitis

When approximately 1/3 and 2/3 of the flock had passed the slaughter line, a sample of 50 right feet was taken from the line and temporarily stored in a box until assessment (in total 100 feet were sampled). This was performed just after hock burn scoring. Feet were scored according to the Bristol Foot Burn scale as described previously for farm measures (Welfare Quality®, 2009).

2.4.2.3 Absence of disease

2.4.2.3.1 Rejections

The number of rejected birds was provided by the slaughter plant. In The Netherlands, only the main reasons for rejection are indicated on the form (by classifying them with +++, ++, +). Rejections are not quantified as numbers of birds per reason for rejection. According to the broiler assessment protocol, it should be distinguished if birds were rejected for reason of ascites, dehydration, septicaemia, hepatitis, pericarditis and abscesses (Welfare Quality®, 2009). This was not possible for Dutch flocks.

2.5 Feedback to the farmer

A report of the assessment of each house was sent to the farmer. This feedback report to the farmer, contained a score for the house together with an indication of the score in relation to the expected Dutch average for each measure. An example of such a report is given in appendix 1.

2.6 Data handling

All data were uploaded in an access database and exported to excel for further calculations after all visits had been completed. Dependent on the data type uploaded, additional calculations were performed. Examples of such calculations are transforming the 5-point scoring of foot pad dermatitis and hock burn into a 3-point score, necessary for calculations of farm scores. See the broiler welfare assessment protocol for description of additional calculations (Welfare Quality®, 2009).
In the broiler welfare assessment protocol (Welfare Quality®, 2009), the incorrect score sheet for the QBA is published (Annex A, page 89) that used the terms of the laying hen QBA. This was only noticed after all visits in 2011 had been performed, as appendix A was used for programming the software for the pda's. To overcome this difference in terms used for the QBA for the data collected in 2008 and 2011, new coefficients and a new constant were calculated for broiler flocks scored based on laying hen terminology . See table 2 for these new coefficients and constant. Flocks assessed in 2008 were scored with the correct terms and calculations for these flocks were performed according to the description in the broiler assessment protocol (Welfare Quality®, 2009).

Table 2. Coefficients and constant for broiler flocks scored on broiler terms (Welfare Quality®, 2009) and on laying hen terms (data provided by F. Welmelsfelder).

terms	Coefficients (broiler terms)	Coefficients (laying hen terms)
Active	0.00593	0.003746904
Relaxed	0.00528	0.010794761
Comfortable	0.01274	0.010599243
Confident	0.00916	0.011039545
Calm	0.00449	0.0085481
Content	0.01321	0.011400917
Energetic	0.00726	0.003330148
Friendly	0.00676	0.012305522
Positively occupied	0.01018	0.00784211
Happy	Not in broiler terms	0.010403183
Playful	0.00746	Not in layer terms
Fearful	-0.00295	-0.007208854
Agitated	-0.00148	-0.009674835
Depressed	-0.01651	-0.01068085
Drowsy	-0.01105	Not in layer terms
Tense	-0.00283	-0.009073022
Unsure	-0.00114	-0.011726853
Frustrated	-0.01062	-0.010989451
Helpless	-0.04383	Not in layer terms
Inquisitive	0.00625	Not in layer terms
Bored	-0.01367	-0.007189465
Scared	0.00011	-0.008385268
Nervous	-0.00039	-0.009315655
Distressed	-0.03121	-0.011127293
Constant pc1	-2.7938	-5.010301444

2.7 Additional expert consultation and calculations

According to the current model used to summarize measures obtained at broiler farms into scores for welfare Criteria (see Botreau et al., 2009 - Part 3 – Subcriteria construction for broilers on farm) , criterion 7 is determined by 6 measures that are obtained at the slaughter house (ascites, dehydration, hepatitis, pericarditis, septicaemia, and abscess), as well as two measures of mortality (mortality – found dead on farm, and culls – actively destroyed on farm).
In the Netherlands (and possibly other European countries), the above 6 measures obtained at the slaughterhouse are not recorded. However, the total prevalence of birds rejected at the slaughterhouse because of any pathological condition (i.e., ascites or dehydration or hepatitis, etc.) is available. In addition, the total mortality rate for each flock (% birds) is available, but not separate figures for mortality and culls.
If the integration model could only be described according to the current WQ protocol, at least for Dutch broiler farms it would be impossible to generate scores for criterion 7. We therefore developed an alternative method of calculating a score for criterion 7, based on the measures that are available, i.e. total rejections and total mortality.
In order to be able to use these latter two measures, another method of summarizing prevalence into a criterion score would be needed. An appropriate method would be, first, to generate spline functions for each measure and, second, to summarize welfare scores obtained with these splines with the use of a Choquet integral. In order to do this, additional expert opinion would be needed on each of these measures, and on the combination of the two.
A new expert consultation was therfore performed for the total percentage mortality and total percentage of rejected birds. The tables used for this expert consultation are shown in Appendix 2. Three experts gave scores for mortality and four experts gave scores for percentage of rejections.
For criterion 1, absence of hunger, the spline function as shown in Botreau et al. (2009, Figure 1.1) is not correct, so a new spline function and Choquet integral were calculated based on the data in Botreau et al. (2009).
For criterion 6, the Choquet integral as presented in Botreau et al. (2009) is not correct, because the scores for expert 1, 4 and 5 are not correct (table 6.4 in Botreau et al., 2009). A new Choquet integral was calculated based on Table 6.4, expert 2 and expert 3 in Botreau et al., 2009.

2.8 Statistical analysis

2.8.1 Data exploration, correlations and differences between rearing systems

All statistical analyses were performed using GENSTAT13 (VSN International Limited, 2010). Statistical analysis was performed in different steps. First, variation in individual measures was checked by making distribution curves of flock scores.
Due to the size of the data file and because the amount of flocks from certain production systems were too low to allow a statistical comparison, the data were divided into two main groups according to bird type and production system. This meant that the standard rearing system using fast growing bird types included animals reaching 2 kg bodyweight in less than 40 days, whereas the alternative rearing system using slow growing types included those that took longer to achieve a bodyweight of 2 kg (Van Middelkoop et al., 2002). Thus the database consisted of 139 flocks fast growing birds and 41 flocks slow growing .

Data were classified as:

- Factors - including rearing system, dust and nationality. Due to confounding between some factors, e.g. between nationality and bird type, only analysis for differences between rearing system were performed. For instance, data from Belgian and Italian flocks were only on fast growing birds whereas the British flocks were exclusively slow growing birds in alternative systems.
- Covariables (X) - environmental-based measurements such as litter score, age, birds per drinker, stocking density at day zero, stocking density in kilograms per square meter and stocking density at day of visit.
- Dependent variables (Y) - those collected directly from the animals such as gait score, cleanliness, foot pad and hock lesions, panting and huddling, etc.

Thereafter, data exploration of on- farm measures was performed based on the following analyses:

a) An Analysis of Variance model (ANOVA) was used to assess differences between bird types;
b) Spearman's ranks correlations were used to analyse the relationships between variables at overall level and for each type of rearing system (standard or alternative) separately.

2.8.2 Calculations of flock scores

For each measure, a score was calculated according to the description in the broiler assessment protocol (Welfare Quality®, 2009). Thereafter, these scores were combined to calculate criterion-scores. These criterion-scores were combined to calculate principle scores. Finally each flock was assigned to a welfare category according to the principle scores it attained. A description of the calculation method for the end score of a flock can be found in the assessment protocol (Welfare Quality®, 2009).

A flock scored 'excellent' if at least two principle scores were ≥80 and all principle scores were ≥55. A flock scored 'enhanced' if at least two principle scores were ≥55 and all principle scores were ≥20. A flock scored 'acceptable' if at least three principle scores were ≥20 and all principle scores were ≥10. All other flocks were scored 'not classified' (Welfare Quality®, 2009).
In the current project we assessed foot pad dermatitis and hock burn on- farm and at the slaughter plant. The end score of a flock was calculated based on the assessment of foot pad dermatitis and hock-burn on-farm. Modifications of the end score for a flock as described in the assessment protocol were (1) a modified Choquet integral for criterion 1, see results section and appendix 6; (2) a modified spline function and Choquet integral for criterion 6, see results section and appendix 5; (3) a modified spline function and Choquet integral for criterion 7, see results section and appendix 4.

Flocks that did not have one or more principle scores due to missing measures received the classification 'missing'

2.8.3 Calculations of possibilities for simplification of the assessment protocol

To compare possible strategies of simplification with the scores according to the full assessment protocol, we defined the calculation according to the full assessment protocol as the golden standard 1 (GS1). In this calculation, for criterion 8 (absence of pain induced by management procedures), all flocks receive a score of 100 (this criterion is not applicable for broilers). Thus, GS1: score criterion 8 =100 (Welfare Quality®, 2009).

In addition, we considered a second golden standard, i.e. golden standard 2 (GS2). In the calculation of this golden standard, the score for criterion 8 is defined as the minimum score for criterion 6 and 7 (absence of injuries or diseases respectively). Thus, GS2=min(criterion 6,criterion 7). This calculation can be regarded as less tolerant than GS1.

Based on the correlations between the animal based measures, it was chosen to analyse two possible methods of simplification of the broiler assessment protocol:

1. To predict gait score from the hock burn score assessed on-farm. This will reduce the assessment time on-farm considerably, as hock burn is assessed in 100 birds (same birds as for foot pad dermatitis and cleanliness) while gait score is assessed in 150 other birds;
2. To predict on-farm measures from measures assessed at the slaughter plant. This involves four different steps: (a) to predict hock burn assessed on-farm from hock burns assessed at the slaughter plant; (b) to predict foot pad dermatitis on-farm from foot pad dermatitis assessed at the slaughter plant; (c) to predict gait score from hock burn assessed at the slaughter plant; (d) to predict cleanliness from hock burn and foot pad dermatitis assessed at the slaughter plant. All these four steps together constitute the second strategy of simplification.

These two possible strategies of simplification of the broiler assessment protocol are analysed for two different sets of data. The first set, called dataset 1 in the results section, are all original data collected. The second dataset, called dataset 2 in the results section, are the original data where measurements of breast blisters are replaced by measurements of hock burn at the slaughter plant (severe hock burn, scored 1 at the plant). The reason for this is that we doubted if the measurements of breast blisters were completely in accordance with the description in the broiler assessment protocol (Welfare Quality®, 2009), and thus might have been less accurate.

2.8.3.1 Calculations of predictions of criterion scores for each strategy of simplification from prevalences for individual measures

Predictions were calculated for scores for the criteria (these are according to the Welfare Quality® model calculated from the individual measures). A logistic regression analysis was carried out on the criterion score of the original model (CScore) for the logistic transformed prevalences of prediction variables x1, x2, .. to calculate predictions of these criterion scores (CScore). Predictions on a logistic scale are subsequently transformed back to fractions p on the original scale, and finally the prediction for the criterion score. CScore is taken from pred.Cscore=100 *p. To illustrate this, the procedure to predict CScore from prevalences of measures x1,x2, …xk is:

- Logistic regression:

with

- transformation back to fractions according to:

- Calculation of predction of Cscore:
  Prediction=100*p

2.8.3.2 Comparison of golden standards GS1 and GS2 with simplification strategy 1 and 2 at the level of final flock score

The comparison of the golden standards with simplification strategies is shown in 4x4 tables (tables 12-15 in the results section), where the rows of the table represent the classifications for the golden standard and the columns represent the classes for the simplified models. Cells in the tables represent the number of farms. This is shown in the left half of the tables. In the right half of the tables 'summarising measures' of the simplification are shown. These are:

1. %equal=%farms classified correct;
2. %sp (%specificity)=%(enhanced+excellent) in simplified model given % (enhanced+excellent) in golden standard. This is the likelihood that a farm classified as enhanced/excellent receives the same classification in the simplified model;
3. %se=%(not classified+acceptable) in simplified model given the % (not classified+acceptable) in the golden standard. This is the likelihood that a farm classified as not classified/acceptable receives the same classification in the simplified model;
4. %fn (%false negative)=100-%se=percentage (not classified+acceptable) that is scored incorrect as (enhanced+excellent);
5. %fp (%false positive)=100-%sp=percentage (enhanced+excellent) that is scored incorrect as (not classified+acceptable).

For each percentage, the 90% confidence interval according to the binomial distribution is also shown as illustration of the inaccuracy of the estimation of the summarising measures. If only a few flocks are in a certain category the estimated sensitivity or specificity will have a large confidence interval.

2.8.3.3 Comparison of golden standards GS1 and GS2 with simplification strategy 1 and 2 at the level of principles and criteria

The simplification strategies that were analysed at the level of flock score were also analysed at the level of principles and the level of criteria. The same summarising measures were calculated as described above for the flock scores. For the flock scores, farms were divided in two groups, either being good (enhanced + excellent) or moderate (not classified + acceptable). For principles and criteria, farms were divided in three groups, that more or less correspond with the scores for the different classifications at the level of principle or criterion: score ≤ 20, 20 ≤ score ≤ 55, score ≥ 55. For each simplification strategy, only the principles and criteria that were affected by the simplification were considered. The tables with the results of the analysis show the %equal, %se and %sp as explained in the previous paragraph for each group.
Each table also shows the spearman rank correlation coefficient between the result of the golden standard and the simplified strategy. The relationship between the golden standard and the simplified strategy is also illustrated with graphs. These graphs show the number of flocks in the different groups (score ≤ 20, 20 ≤ score ≤ 55, score ≥ 55) and thus provide an impression of the distribution of flocks over the different groups.

The 90% confidence interval is shown as illustration of the inaccuracy of the estimation of the summarising measures. If only a few flocks are in a certain category the estimated sensitivity or specificity will have a large confidence interval.

3 Results

3.1 Variation for individual measures

Appendix 3 contains graphs representing the variation for individual measures . In general, large variation was found between flocks for individual measures. Exceptions showing less variation were touch test scores and dust scores. For dust scores, extreme levels were scarce. For touch test scores, a large number of flocks showed either a very low or a very high score.

3.2 Effect of rearing system on the animal-based measures

3.2.1 Farm measures

In general, there was a large effect of rearing system on the animal-based measures. Figure 3 shows differences in bird cleanliness between standard and alternative systems. Slower growing birds in alternative systems were in general cleaner than fast growing ones in standard systems.

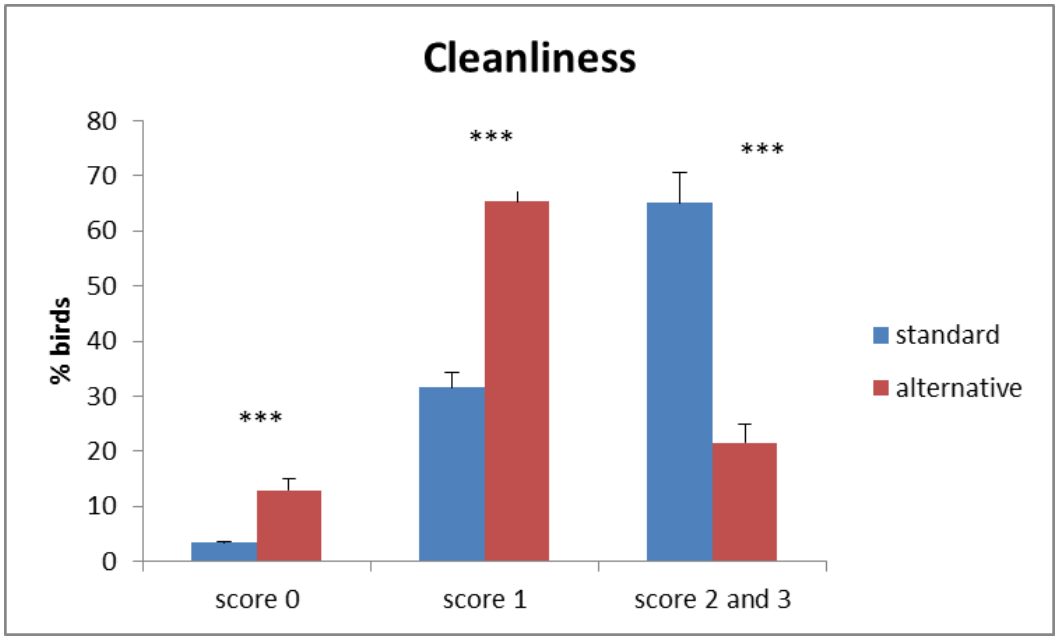


Figure 2. Differences in cleanliness scores between birds in standard and alternative systems (*** P<0.001). Score 1: clean birds; score 2: slightly dirty birds; score 2: moderately dirty; score 3: very dirty birds.

Figure 3 shows that more birds displayed panting behaviour in standard systems than those on alternative systems. Huddling was only very occasionally observed, thus no analysis was done on these data.

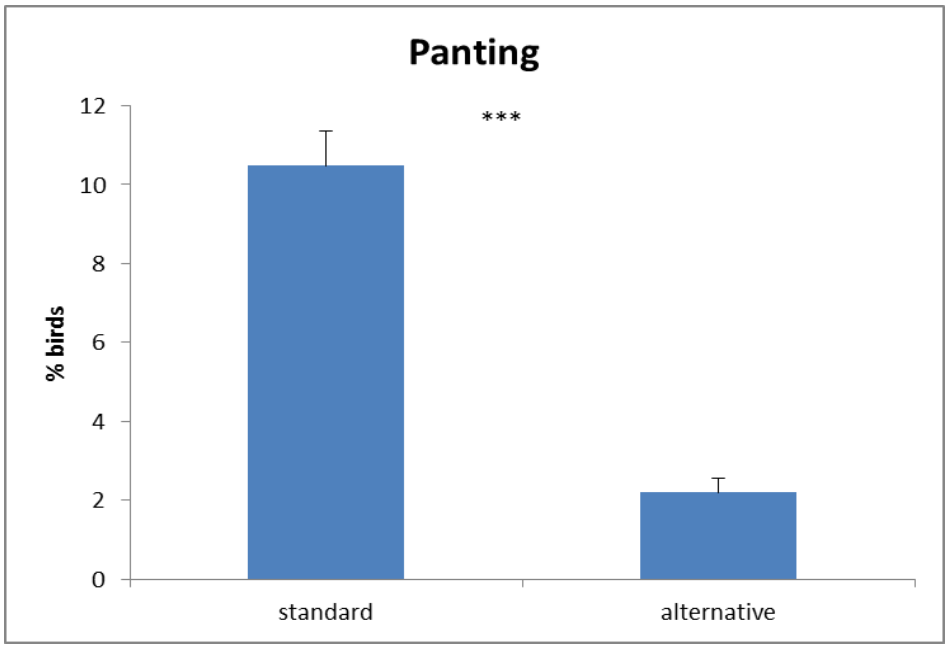


Figure 3. Percentage of birds showing panting for standard and alternative rearing systems (***P<0.001).

There were also large differences in gait scores between standard rearing systems with fast growing birds and alternative rearing systems with slower growing birds. Fewer birds in alternative rearing systems were lame (scores 3 and higher) (Figure 4). More than half of the broilers in standard rearing systems were lame.

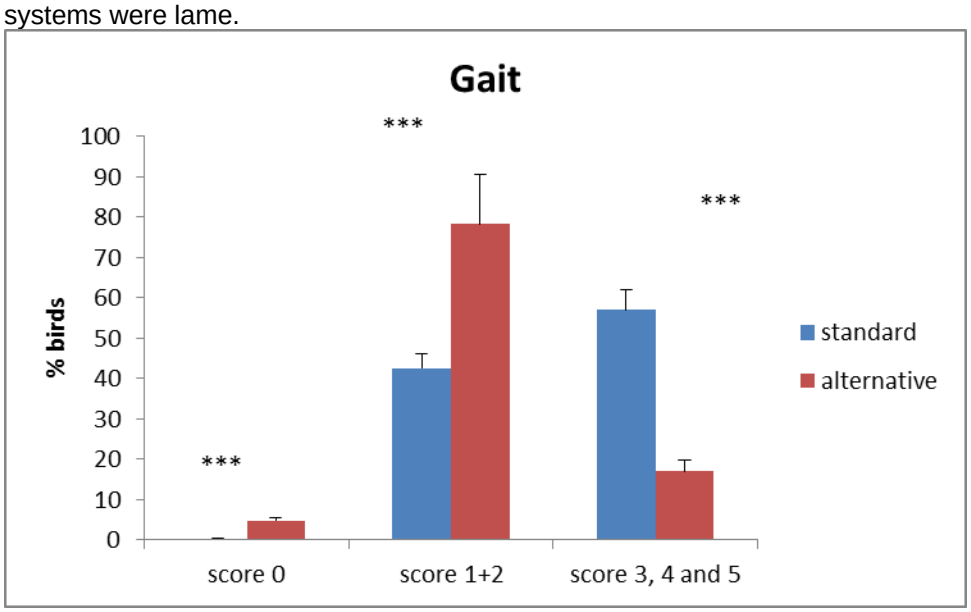


Figure 4. Percentage of birds with sound gait (score 0), a little or moderate walking deficiency (scores 1 and 2) or lame (score 3 and higher) in standard and alternative rearing systems (***P<0.001).

Fast growing birds in standard rearing systems had significantly more hock burn (Figure 5) and foot pad dermatitis (Figure 6) compared to slow growing birds in alternative systems.

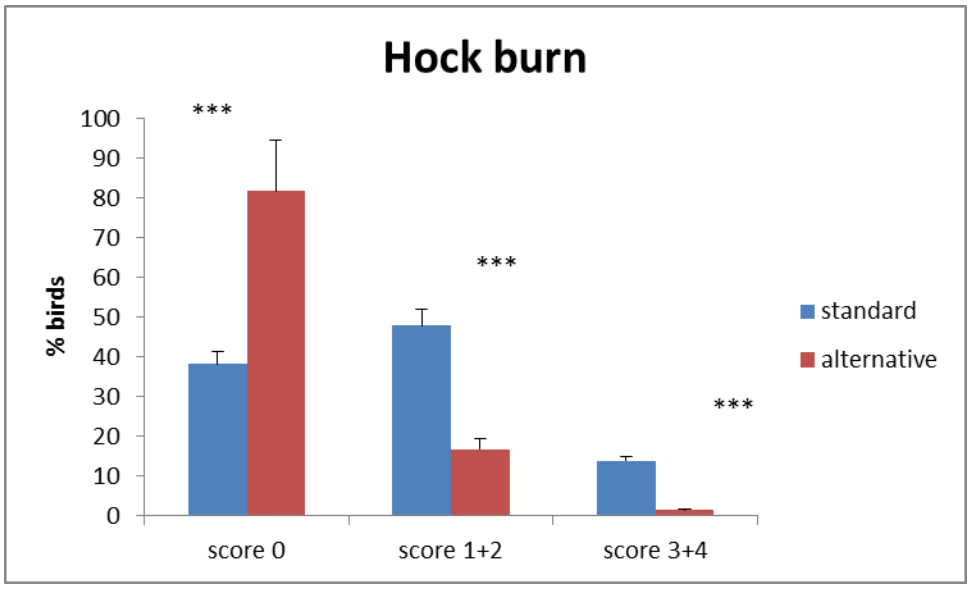


Figure 5. Percentage of birds with no hock burn (score 0), minimal evidence of hock burn (score 1 and 2) or evidence of hock burn (score 3 and 4) in standard and alternative rearing systems (***P<0.001).

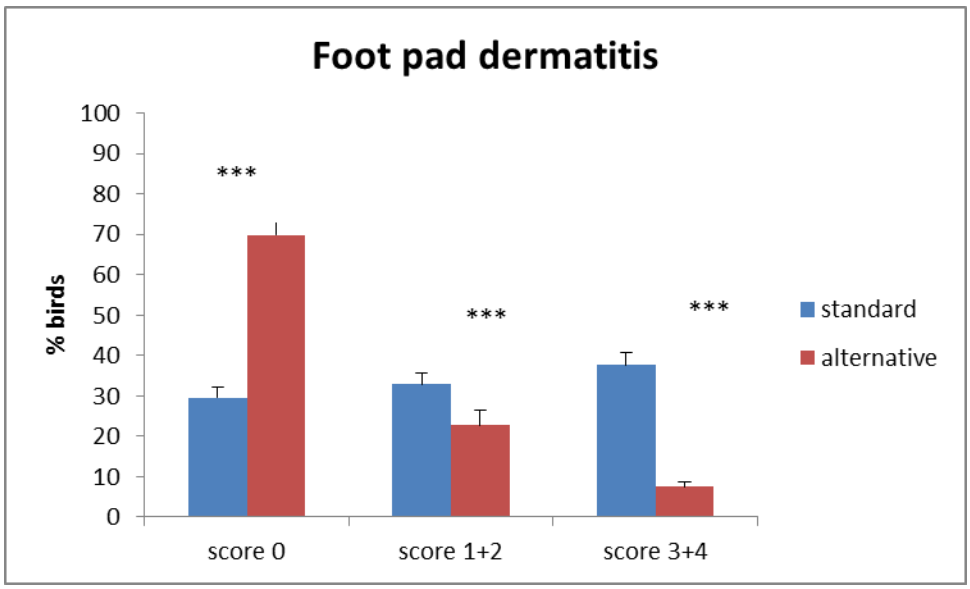


Figure 6. Percentage of birds with no foot pad dermatitis (score 0), minimal evidence of foot pad dermatitis (score 1 and 2) or clear evidence of foot pad dermatitis (score 3 and 4) in flocks from standard and alternative rearing systems (***P<0.001).

Mortality figures did not differ significantly between flocks from standard and alternative rearing systems. It should be noted that mortality was recorded on the day of visit, which was on average at 38 days of age for fast growing birds in standard systems and 54 days of age for the slow growing birds in alternative systems.

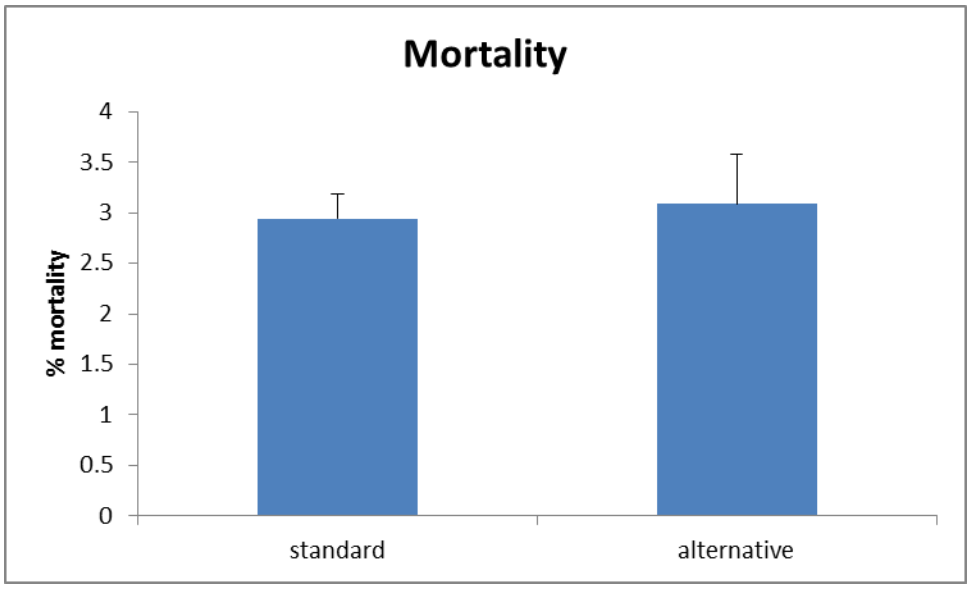


Figure 7. Average mortality in flocks from standard and alternative rearing systems.

Birds in standard and alternative rearing systems also differed significantly in their behaviour in the touch test as well as their behaviour in general (scored with the QBA). Touch test scores were on average lower for flocks in alternative systems than for flocks in standard systems. A lower score means that less birds could be touched by the assessor. Flocks with slow growing birds in alternative systems had on average higher scores in the QBA, which means that in general they could be characterised as having more positive than negative behaviours.

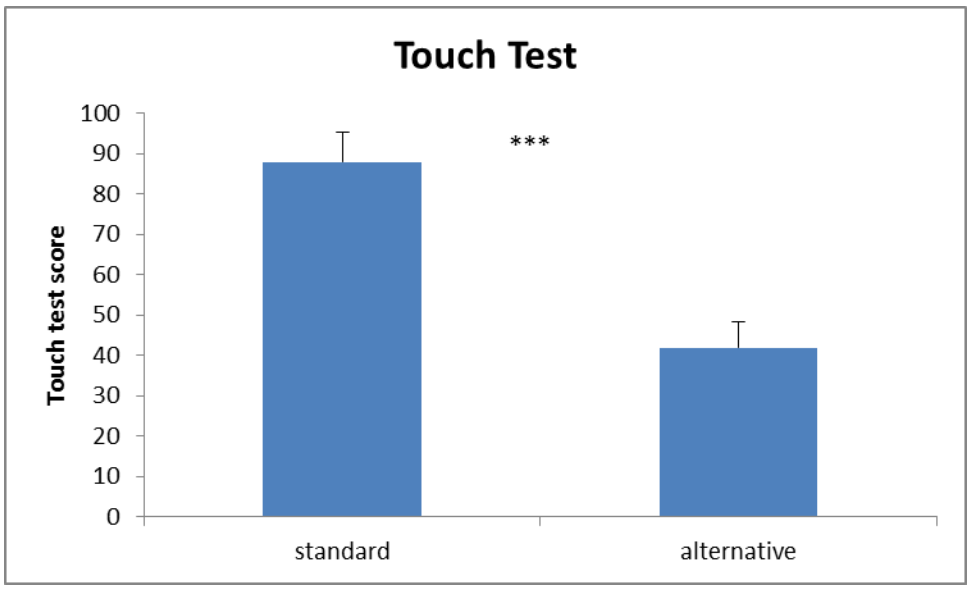


Figure 8. Average touch test scores for flocks in standard and alternative rearing systems (***P<0.001).

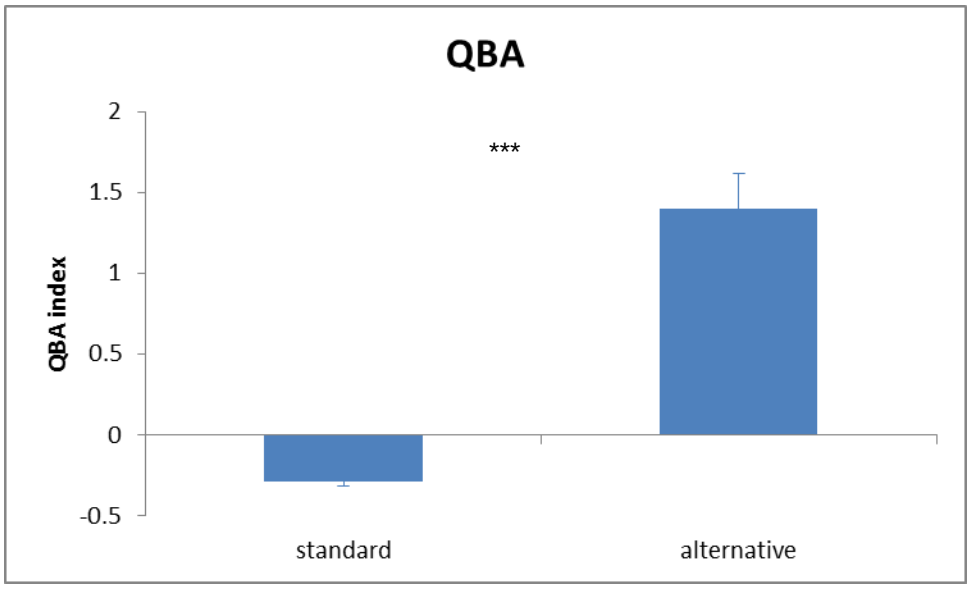


Figure 9. Average QBA index scores for flocks in standard and alternative rearing systems (***P<0.001).

3.2.2 Slaughter plant measures

Figures 10, 11, 12, and 13 show average values of the animal-based measures scored at the slaughter plant. Slow growing birds in alternative systems had less breast blisters, and less foot pad dermatitis and hock burn, as was also found for the on-farm measures. No differences were found for the percentage of rejections.

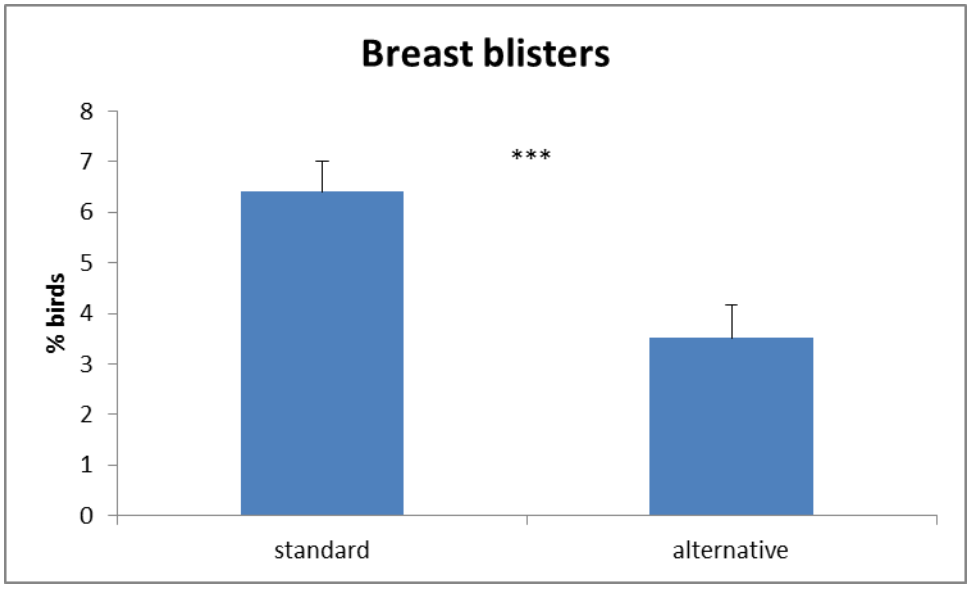


Figure 10. Percentage of birds in a flock with breast blisters (blisters and breast irritation) for flocks in standard and alternative rearing systems (***P<0.001).

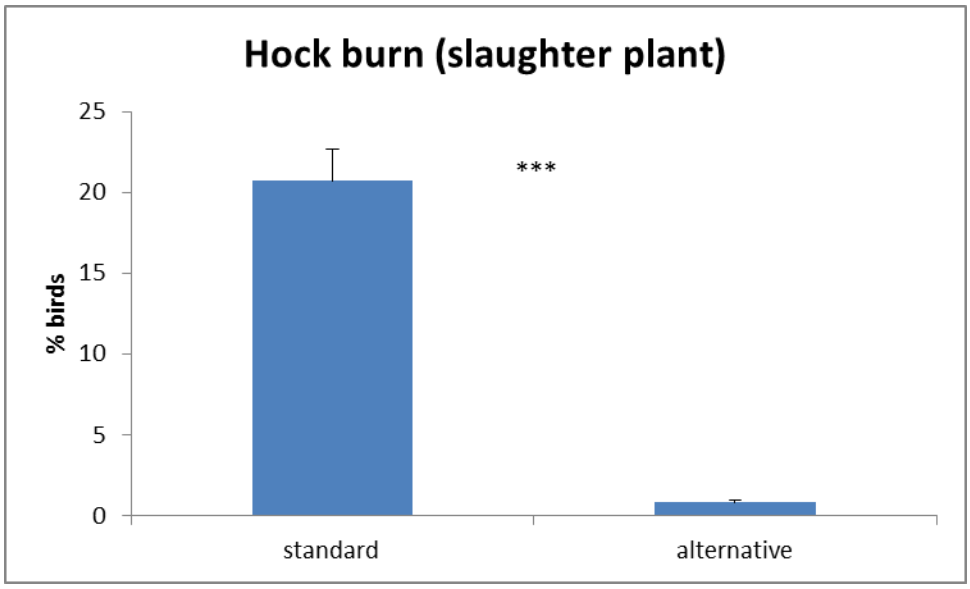


Figure 11. Average percentage of birds with evidence of hock burn in standard and alternative rearing systems, measured at the slaughter plant (***P<0.001).

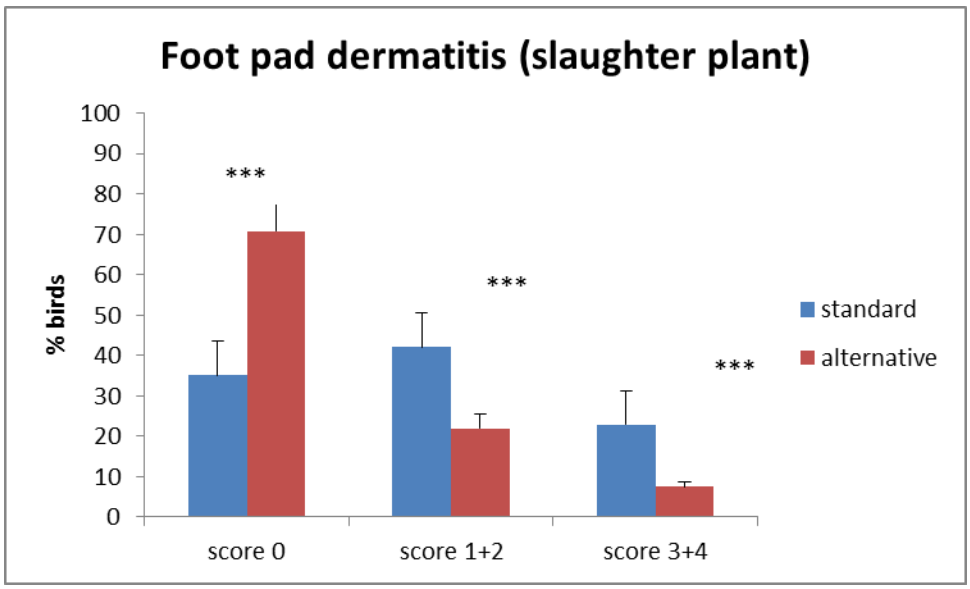


Figure 12. Percentage of birds with no foot pad dermatitis (score 0), minimal evidence of foot pad dermatitis (score 1 and 2) or clear evidence of foot pad dermatitis (score 3 and 4) in flocks in standard and alternative rearing systems, measured at the slaughter plant (***P<0.001).

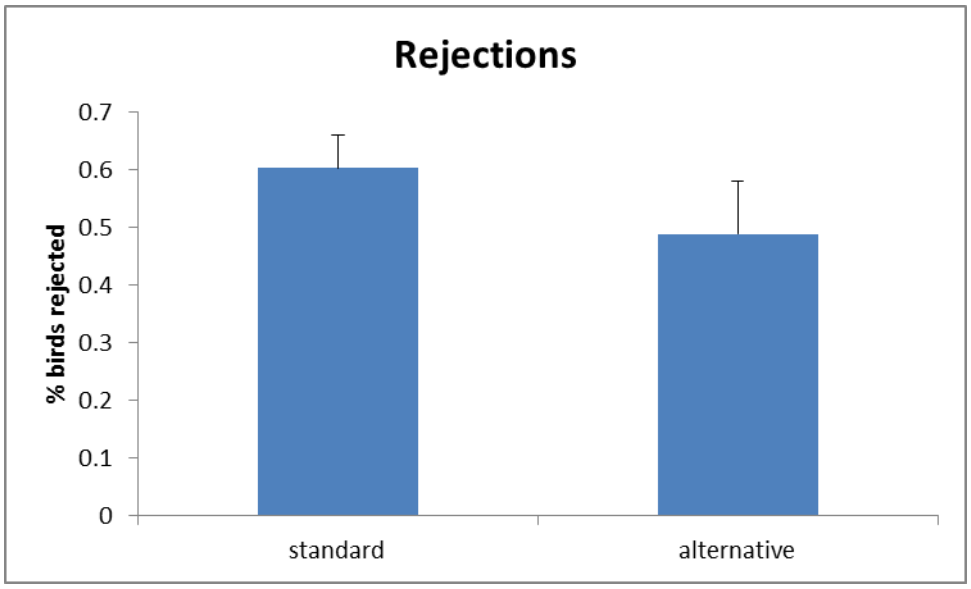


Figure 13. Average percentage of rejections at the slaughter plant for flocks from standard and alternative rearing systems.

3.3 Relationship between animal-based measurements

For an indication of possibilities for simplification, relationships between scores for moderate and severe classes of animal-based measures were most interesting, and when such correlations were found at the overall level as well as at the level of standard and alternative flocks separately. These results are presented below. Only moderate correlations ($r \geq 0.3$; $r<0.7$) are presented for the on-farm measures as no high correlation was found between on-farm measures. Where no correlation is shown, no significant correlation and/or a very low correlation ($r<0.3$) was found. Strong correlations ($r \geq 0.7$) were only found between on-farm measures and slaughter plant measures.

3.3.1 Relationships between on-farm measurements

Tables 3, 4, 5 and 6 show significant and moderate correlations between animal-based measures scored on-farm. There was a moderate positive correlation of the percentage of birds panting with severe gait score, but only for the birds in standard rearing systems. There was also a moderate correlation between the percentage of birds with moderate foot pad dermatitis scores and hock burn scores and the percentage of birds panting, but these correlations were not found for all types of systems (Table 3).

Table 3. Correlation coefficient (r) for relations between the percentage of birds panting and gait scores, scores for hock burn (HB) and scores for foot pad dermatitis (FPD), at an overall level and for birds in standard and alternative systems separately. Correlations shown in the table were significant ($P<0.05$ at least).

Correlation of percentage of birds panting with:		Overall r	Standard r	Alternative r
Gait score	Moderate (score 1+2)	-0.517	-0.445	
	Severe (score 3+4)	0.540	0.447	
FPD score	Moderate			-0.321
HB score	Moderate	0.347		

Tables 4 and 5 show the correlations between moderate and severe gait scores respectively, and other animal-based measures. The highest correlations were found between moderate gait scores and moderate hock burn scores, and severe gait scores and severe hock burn scores. These correlations were not only found at the overall level, but also within the different rearing systems (Table 4 and 5).

Table 4. Correlation coefficient (r) for relations between the percentage of birds with moderate gait scores, and scores for cleanliness, scores for hock burn (HB) and scores for foot pad dermatitis (FPD), at an overall level and for birds in standard and alternative rearing systems separately. Correlations shown in the table were significant (P<0.05 at least).

Correlation of % birds with moderate gait with:		**Overall r**	**Standard r**	**Alternative r**
Cleanliness	Moderate (score 1)	0.406		
	Soiled (score 2+3)	-0.473	-0.327	
FPD	Severe	-0.341		
HB	Severe	-0.597	-0.448	-0.375

Table 5. Correlation coefficient (r) for relations between the percentage of birds with severe gait scores, and scores for cleanliness, hock burn (HB) and foot pad dermatitis (FPD), overall and for birds in standard and alternative systems separately. Correlations shown in the table were significant (P<0.05 at least).

Correlation of % birds with severe gait with:		**Overall r**	**Standard r**	**Alternative r**
Cleanliness	Moderate (score 1)	-0.422		
	Soiled (score 2 and 3)	0.492	0.329	
FPD	Severe (score 3 and 4)	0.370		
HB	Moderate (score 1 and 2)	0.305		0.317
	Severe (score 3 and 4)	0.615	0.448	0.443

Correlations between the percentage of birds with severe foot pad dermatitis and scores for cleanliness and hock burn were lower in comparison to correlations with gait scores. In addition, these correlations were only for soiled birds found at the overall level and flocks in standard and alternative systems. For the other measures correlations were only found for one system (Table 6).

Table 6. Correlation coefficient (r) for relations between the percentage of birds with severe foot pad dermatitis, and scores for cleanliness and hock burn (HB), overall and for birds in standard and alternative systems separately. Correlations shown in the table were significant (P<0.05 at least).

Correlation of severe foot pad dermatitis with:		**Overall r**	**Standard r**	**Alternative r**
Cleanliness	Moderate (score 1)	-0.422		-0.477
	Soiled (score 2 and 3)	0.449	0.152	0.380
HB	Moderate (score 1 and 2)	0.389		0.432
	Severe (score 3 and 4)	0.458	0.308	

3.3.2 Relations between slaughter plant measures

The only significant and moderate correlation on the overall level between slaughter plant measures was found for the percentage of birds with severe FPD and the percentage of birds with hock burn measured at the slaughter plant (r=0.544, P<0.001). Also for the birds in standard rearing systems a moderate correlation was found between these two measures (r=0.351, P<0.001). This correlation was higher for birds in alternative rearing systems (r=0.507, P<0.01).

3.3.3 Relations between slaughter plant measures and farm measures

In order to determine the possibilities for simplification of assessment protocol, the relationship between clinical scores measured at the slaughter plant (foot pad dermatitis, hock burn and breast blisters) and clinical scores measured on-farm (foot pad dermatitis, hock burn and cleanliness) was calculated, as well as correlations of slaughter plant measures with gait score. Overall, a moderate correlation was found for severe gait score at the farm and hock burn scored at the slaughter plant (r=0.518, P<0.001), but this correlation could not be found for birds in standard and alternative systems separately.
Table 7 summarizes the results of the analysis of correlation between clinical scores at the slaughter plant and clinical scores at the farm. Strong correlations were found between FPD scores at farm and slaughter plant, overall and for birds in standard and alternative systems. For birds in standard systems this correlation was moderate, but still relatively high. Correlations between cleanliness and hock burn at farm level, and slaughter plant measures were found at an overall level (for severe foot pad dermatitis and hock burn) and for birds in alternative systems but not for birds in standard rearing systems.

Table 7. Correlation coefficients (r) for correlations between farm measures and slaughter plant measures, on an overall level and for birds in standard and alternative systems separately. All correlations shown were significant (P<0.05 at least).

Correlation between	**Severe FPD (farm)**	**Severe Hock burn (farm)**	**Cleanliness ≥ 2 (soiled) (farm)**
Overall			
Severe FPD (slaughter)	0.732	0.344	0.396
Hock burn (slaughter)	0.452	0.527	0.591
Breast burn (slaughter)			
Standard			
Severe FPD (slaughter)	0.609		
Hock burn (slaughter)		0.346	
Breast burn (slaughter)			
Alternative			
Severe FPD (slaughter)	0.723		0.493
Hock burn (slaughter)	0.399		0.525
Breast burn (slaughter)			

3.4 Calculation of end scores

3.4.1 Result of expert consultation, new spline functions and Choquet Integral for criterion 7: absence of disease

Appendix 4 presents a modified calculation model for criterion 7, based on measures for rejections and total mortality. For further explanation of these calculations, we refer to Botreau et al., 2009.

3.4.1 Spline function and Choquet Integral for criterion 1: absence of hunger

The modified spline function and Choquet Integral for criterion 1 can be found in appendix 5.

3.4.2 Choquet Integral for criterion 6: absence of injuries

The modified Choquet Integral for criterion 6 can be found in appendix 6.

3.4.3 End scores based on the full assessment protocol

Results of the calculations of end scores per flock are presented in Table 8. For 53 flocks, no end score could be calculated due to the absence of a score for breast blisters. This was due to the fact that for these flocks no slaughter plant assessments could be performed or breast blister scores were lacking (7 flocks of the Dutch farms visited in 2011, as well as the Belgian flocks, the UK flocks, and some of the Dutch and Italian flocks visited in 2008). The majority of flocks received the score acceptable. Only a few farms were scored as enhanced, no farms scored excellent and also a few farms were not classified.

Table 8. Number and percentage of flocks in each category, based on the full broiler welfare assessment protocol (Welfare Quality®, 2009). NA=not scored due to missing principle score.

WQ category		**Total number per category for each farming system**		**Percentage per category for each farming system**	
	Total number in category	*Standard*	*Alternative*	*standard*	*alternative*
Excellent	**0**	0	0	0	0
Enhanced	**7**	2	5	1.5	12.2
Acceptable	**104**	86	18	61.9	43.9
Not classified	**16**	10	6	7.2	14.6
NA	**53**	41	12	29.4	29.3
Total	**180**	139	41		

In order to provide more insight as to why the majority of farms were scored in the same category, we also analysed the scores per principle. We provide some examples of principle scores for flocks not classified and flocks that were acceptable. Table 9 shows the scores per principle for farms that were scored as not classified. This table illustrates that the reason for being categorized as not classified may differ between standard and alternative rearing systems. The majority of flocks in alternative systems in this category received a low score on principle 4 (appropriate behaviour). The majority of the flocks in standard rearing systems received a low score for principle 2 (good housing).

Table 9. Principle scores for flocks that were not classified. Flocks that did not score at least 10 on all principles and at least 20 on three of the principles, were not classified. Lowest scores per flock are presented in bold type.

Farming system	Principle 1 (good feeding)	Principle 2 (good housing)	Principle 3 (good health)	Principle 4 (appropriate behaviour)
Standard	42	43	27	**2**
Standard	**13**	**16**	23	24
Standard	36	**16**	27	**14**
Standard	**15**	**14**	26	28
Standard	59	**14**	22	**18**
Standard	**5**	**16**	23	25
Standard	41	**16**	28	19
Standard	41	**17**	25	20
Standard	70	**10**	31	19
Standard	**2**	34	33	**11**
Alternative	51	62	38	**8**
Alternative	**7**	53	42	30
Alternative	66	62	44	**10**
Alternative	55	55	38	**9**
Alternative	21	68	42	**5**
Alternative	29	61	28	**3**

Table 10 shows some examples of flocks that scored acceptable. It illustrates that in most cases for flocks with slow growing birds in alternative systems the lowest score was for principle 4, appropriate behaviour, whereas for flocks with fast growing birds in standard systems the lowest score was for principle 2, good housing.
Flocks that were classified as enhanced were flocks with slow growing birds in alternative systems, and only two Italian flocks with fast growing birds in standard systems were classified in this category.

Table 10. Principle scores for examples of flocks classified as acceptable. Flocks received a classification acceptable with a score of at least 10 on all principles and at least 20 on three of them. The lowest score for a flock is presented in bold type.

Bird type	Principle 1 (good feeding)	Principle 2 (good housing)	Principle 3 (good health)	Principle 4 (appropriate behaviour)
Standard	64	**18**	28	29
Standard	67	**17**	31	28
Standard	36	**21**	27	**12**
Standard	62	**12**	33	27
standard	44	**11**	37	30
alternative	96	57	51	**12**
alternative	97	57	52	**13**
alternative	84	49	37	**25**
alternative	41	61	40	**15**
alternative	81	51	45	**28**

End scores for each flock were also calculated with dataset 2, in which breast blister scores were replaced with scores for hock burn measured at the slaughter plant. This did not result in a shift in classification of the flocks. For the results we refer the reader to appendix 7, table 7.1.

3.5 Strategies for simplification of the broiler welfare assessment protocol

3.5.1 Results of simplification at the level of the end score of flocks

3.5.1.1 Distribution of flocks over categories for GS1 and GS2, dataset 1

Table 11 shows the distribution of flocks over the final categories for both GS1 and GS2. With GS1 more farms are classified as 'acceptable' and less farms are 'not classified' as compared to GS2. For both golden standards, no farms are classified 'excellent' and the majority of the farms are classified 'acceptable'.

Table 11. Number of flocks in each category, according to the full assessment protocol (Welfare Quality®, 2009) for Golden Standard 1 (GS1) and Golden Standard 2 (GS2) (GS1: criterion8=100; GS2: criterion 8=min(criterion6,criterion7). NA=not scored due to missing principle score.

WQ category	**GS1**	**GS2**
Excellent	0	0
Enhanced	7	7
Acceptable	104	99
Not classified	16	21
NA	53	53
Total number	180	180

3.5.1.2 Distribution of flocks over classifications (flock scores) for the full model and two simplified models, with the original dataset (dataset 1)

Tables 12 and 13 show the comparison of strategy 1 of simplification (replacing gait scores with hock burn scores on-farm) for GS1 and GS2. Results are equal for the full model (GS1) and the simplified model (table 12). Agreement is lower, but still high, for the full model (GS2) and the simplified model (table 13). A large confidence interval is found for specificity which is influenced by the low number of flocks in the categories enhanced and excellent (Table 12 and 13).

Table 12. Comparison of a simplified model (strategy 1, prediction of gait score from hock burn on-farm) with the full model for GS1. For explanation of terms, see 2.8.3.2. The table shows the distribution of flocks over the categories, with a summary of the measures at 90% confidence intervals. Est=estimated agreement; low=lower limit of confidence interval; upp=upper limit of confidence interval.

Strategy 1 → Full model ↓	Excell	Enhanc	Accept	Not cl	NA	Margin			90% conf. interval	
Excellent	0	0	0	0	0	0		Est	Low	upp
Enhanced	0	7	0	0	0	7	%equal	100.0	97.1	100.0
Acceptable	0	0	104	0	0	104	%sp	100.0	59.0	100.0
Not classified	0	0	0	16	0	16	%se	100.0	97.0	100.0
NA	0	0	0	0	53	53	%fn	0.0	0.0	41.0
Margin	0	7	104	16	53	180	%fp	0.0	0.0	3.0

Table 13. Comparison of a simplified model (strategy 1, prediction of gait score from hock burn on-farm) with the full model for GS2. For explanation of terms, see 2.8.3.2. The table shows the distribution of flocks over the categories, with a summary of the measures at 90% confidence intervals. Est=estimated agreement; low=lower limit of confidence interval; upp=upper limit of confidence interval.

Strategy 1 → Full model ↓	Excell	Enhanc	Accept	Not cl	NA	Margin			90% conf. interval	
Excellent	0	0	0	0	0	0		est	low	upp
Enhanced	0	6	1	0	0	7	%equal	97.6	93.2	99.5
Acceptable	0	0	98	1	0	99	%sp	85.7	42.1	99.6
Not classified	0	0	1	20	0	21	%se	100.0	97.0	100.0
NA	0	0	0	0	53	53	%fn	14.3	0.4	57.9
Margin	0	6	100	21	53	180	%fp	0.0	0.0	3.0

Table 14 and table 15 show the comparison of strategy 2 of simplification (predicting on-farm measures from slaughter plant measures) for GS1 and GS2. Results are almost equal for the full model and the simplified model for GS1 and GS2. For the specificity, a large confidence interval is found which is due to the low number of flocks in the categories enhanced and excellent (table 14 and 15). This also explains the relatively high percentage of false negatives (%fn).

Table 14. Comparison of a simplified model (strategy 2, predicting on-farm measures from slaughter plant measures) with the full model for GS1. For explanation of terms, see 2.8.3.2. The table shows the distribution of flocks over the categories, with a summary of the measures at 90% confidence intervals. Est=estimated agreement; low=lower limit of confidence interval; upp=upper limit of confidence interval.

Strategy 2 → Full model ↓	Excell	Enhanc	Accept	Not cl	NA	Margin			90% conf. interval	
Excellent	0	0	0	0	0	0		est	low	upp
Enhanced	0	5	2	0	0	7	%equal	97.6	93.2	99.5
Acceptable	0	1	103	0	0	104	%sp	71.4	29.0	96.3
Not classified	0	0	0	16	0	16	%se	99.2	95.4	100.0
NA	0	0	0	0	53	53	%fn	28.6	3.7	71.0
Margin	0	6	105	16	53	180	%fp	0.8	0.0	4.6

Table 15. Comparison of a simplified model (strategy 2, predicting on-farm measures from slaughter plant measures) with the full model for GS2. For explanation of terms, see 2.8.3.2. The table shows the distribution of flocks over the categories, and a summary of measures at 90% confidence interval. Est=estimated agreement; low=lower limit of confidence interval; upp=upper limit of confidence interval.

Strategy 2 → Full model ↓	Excell	Enhanc	Accept	Not cl	NA	Margin			90% Conf. interval	
Excellent	0	0	0	0	0	0		est	low	Upp
Enhanced	0	5	2	0	0	7	%equal	95.2	90.0	98.2
Acceptable	0	1	96	2	0	99	%sp	71.4	29.0	96.3
Not classified	0	0	1	20	0	21	%se	99.2	95.4	100.0
NA	0	0	0	0	53	53	%fn	28.6	3.7	71.0
Margin	0	6	99	22	53	180	%fp	0.8	0.0	4.6

3.5.1.3 Distribution of flocks over classifications (flock scores) for the full model and two simplified models for dataset 2

The results of the simplification, based on dataset 2 (breast blister measures replaced by hock burn measures at the slaughter plant), do not differ from the results of simplification using the original dataset. These results are presented in appendix 7.

3.5.2 Results of simplification at the level of individual principle scores for dataset 1

We only considered principles that were affected by the individual simplification strategies. For strategy 1, it concerns principle 3 (good health). For strategy 2 (predicting on-farm measures from slaughter plant measures) it concerns principle 2 (good housing) and principle 3 (good health). Scores for principle 3 will be affected by the choice of the golden standard (GS1 and GS1). This did not affect scores for principle 2.

3.5.2.1 Comparing golden standards with strategies for simplification: principle level, dataset 1

Table 16 shows the comparison of the full model, GS1 and GS2, with strategy 1 for simplification (calculation of gait scores from hock burn scores on-farm) for principle 3 (good health). Results are equal (GS1) or almost equal (GS2) for the full and the simplified model. Figure 14 shows a graphic representation of the score for principle 3 for the golden standard against strategy 1 for simplification. In general, correlation between the golden standard and the simplified model is high. The large confidence interval for the sensitivity when comparing GS2 with the full model for principle 3 is caused by the low number of farms in this category ≤ 20.

Table 16. Comparison of a simplified model (strategy 1, replacing gait score with hock burn scores on-farm) for GS1 (upper part of the table) and GS2 (lower part of the table). For explanation of terms, see 2.8.3.2. Empty cells indicate that there are no values for this category. Rsp: Spearman rank correlation coefficient. Est=estimated agreement; low=lower limit of confidence interval; upp=upper limit of confidence interval.

Principle 3 (good health)		20.0			55.0			80.0		
GS1:C8=100		90% Conf.interval			90% Conf.interval			90% Conf.interval		
		est.	lower	upper	est.	lower	upper	est.	lower	upper
Prediction of gait score from hock burn	%equal	100.0	97.7	100.0	100.0	97.7	100.0	100.0	97.7	100.0
	%se				100.0	97.7	100.0	100.0	97.7	100.0
Rsp=0.79	%sp	100.0	97.7	100.0						

Principle 3 (good health) GS2:C8=min(C6,C7)		20.0			55.0			80.0		
		90% Conf.interval			90% Conf.interval			90% Conf.interval		
		est.	lower	upper	est.	lower	upper	est.	Lower	upper
Prediction of gait score from hock burn	%equal	96.9	93.1	98.9	100.0	97.7	100.0	100.0	97.7	100.0
	%se	81.8	53.0	96.7	100.0	97.7	100.0	100.0	97.7	100.0
Rsp=0.86	%sp	98.3	94.8	99.7						

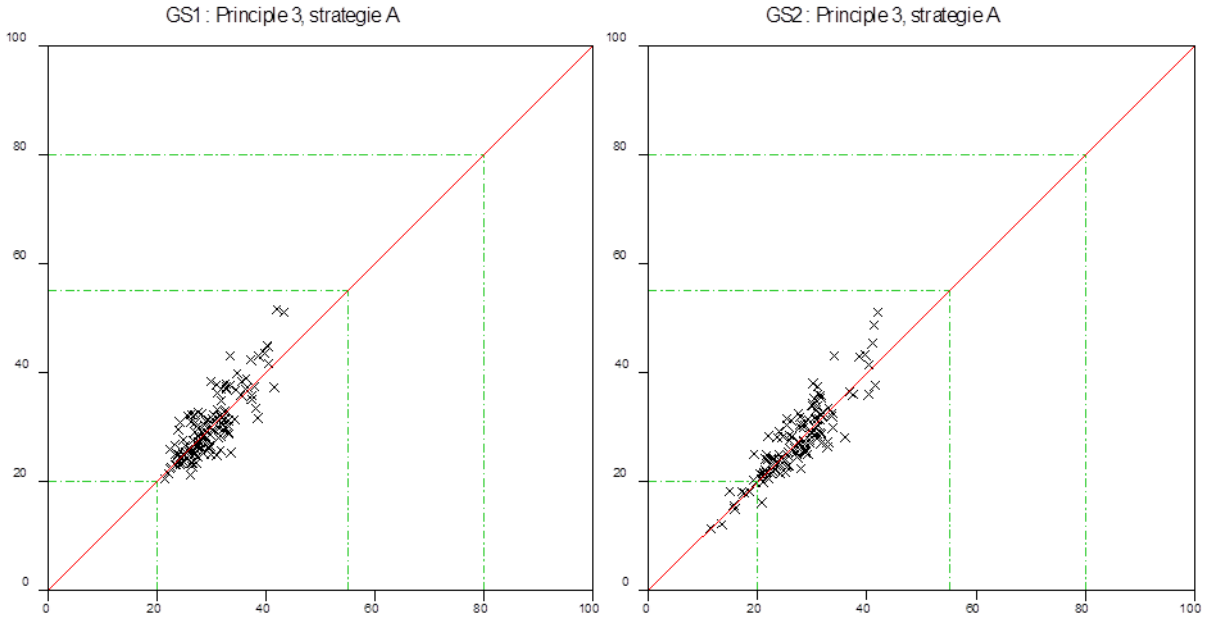


Figure 14. Score for principle 3 (good health) for the golden standard (GS) on the Y-axis against the simplified model (strategy 1, gait score replaced by hock burn on-farm) on the X-axis. The left figure shows the results for GS1, the right figure for GS2.

Table 17 shows the comparison of the full model, GS1 and GS2, with strategy 2 for simplification (prediction of on-farm measures from slaughter plant measures) for principles 2 and 3. Results are for the full model and the simplified model are either equal or close to equality. For GS2 the confidence interval for specificity is large, which is due to the few farms in the category ≤ 20. Figure 15 shows a graphic representation of the scores for principle 3 for the golden standard against strategy 2 for simplification. In general, correlation between the golden standard and the simplified model is moderate (GS1, principle 3) to high. Figure 16 shows a graphic representation of the scores for principle 2 for the golden standard against strategy 2 for simplification. Correlation between the golden standard and the simplified model is very high for principle 2.

Table 17. Comparison of a simplified model (strategy 2, replacing on-farm measures with slaughter plant measures) for principle 3, GS1 (upper part of the table) and GS2 (middle part of the table), and principle 2. For explanation of terms, see 2.8.3.2. Empty cells mean that there are no values for this category. Rsp: Spearman rank correlation coefficient. Est=estimated agreement; low=lower limit of confidence interval; upp=upper limit of confidence interval.

Principle 3 (good health)		20.0			55.0			80.0		
GS1:C8=100		90% Conf.interval			90% Conf.interval			90% Conf.interval		
		est.	lower	upper	est.	lower	upper	est.	Lower	upper
Prediction of on-farm measures from slaughter plant measures	%equal	100.0	97.7	100.0	100.0	97.7	100.0	100.0	97.7	100.0
	%se				100.0	97.7	100.0	100.0	97.7	100.0
Rsp=0.61	%sp	100.0	97.7	100.0						

Principle 3 (good health)		20.0			55.0			80.0		
GS2:C8=min(C6,C7)		90% Conf.interval			90% Conf.interval			90% Conf.interval		
		est.	lower	upper	est.	lower	upper	est.	lower	upper
Prediction of on-farm measures from slaughter plant measures	%equal	96.2	92.1	98.5	100.0	97.7	100.0	100.0	97.7	100.0
	%se	72.7	43.6	92.1	100.0	97.7	100.0	100.0	97.7	100.0
Rsp=0.73	%sp	98.3	94.8	99.7						

Principle 2 (good housing)		20.0			55.0			80.0		
Same for GS1 and GS2		90% Conf.interval			90% Conf.interval			90% Conf.interval		
		est.	lower	upper	est.	lower	upper	est.	lower	upper
Prediction of on-farm measures from slaughter plant measures	%equal	98.5	95.5	99.7	96.4	92.5	98.6	100.0	97.8	100.0
	%se	97.8	89.9	99.9	98.2	94.5	99.7	100.0	97.8	100.0
Rsp=0.99	%sp	98.9	95.0	99.9	87.5	70.8	96.5			

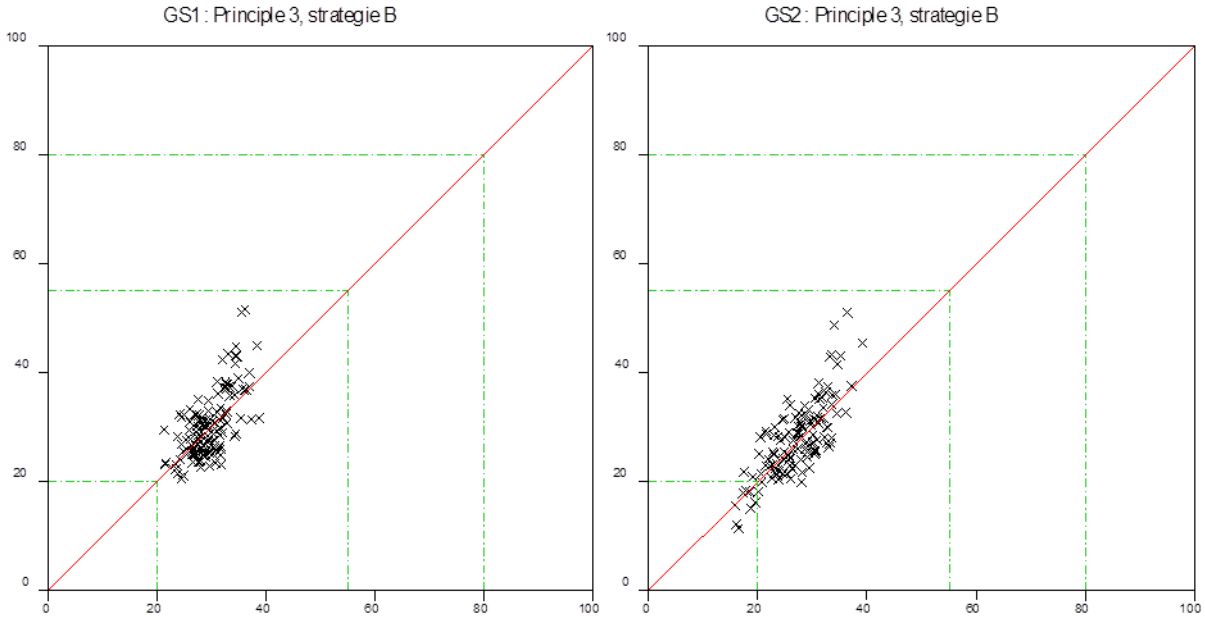


Figure 15. Score for principle 3 (good health) for the golden standard (GS) on the Y-axis against the simplified model (strategy 2, prediction of on-farm measures from slaughter plant measures) on the X-axis. The left figure shows the results for GS1, the right figure for GS2.

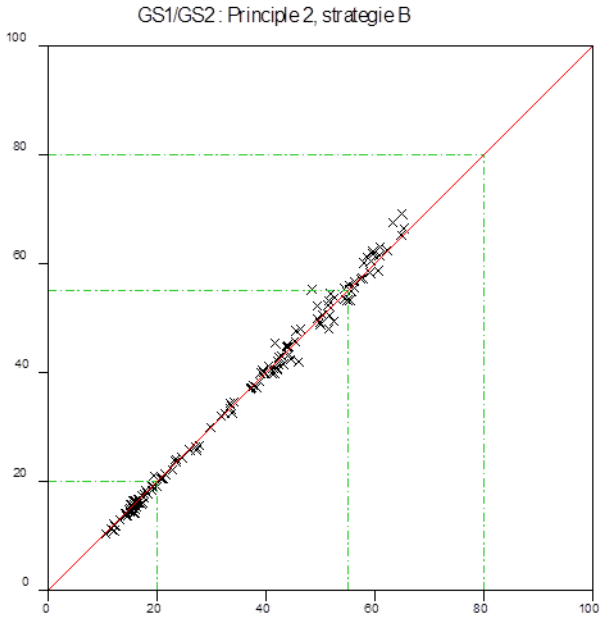


Figure 16. Score for principle 2 (good housing) for the golden standard (GS) on the Y-axis against the simplified model (prediction of on-farm scores from slaughter plant measures) on the X-axis. Scores for principle 2 are the same for GS1 and GS2.

3.5.3 Results of simplification at the level of individual principle scores for dataset 2

For results of simplification at the level of individual principle scores for dataset 2, breast burn measures replaced by hock burn measures at the slaughter plant, we refer the reader to appendix 8. Results with dataset 2 do not differ from results with dataset 1.

3.5.4 Results of simplification at the level of individual criterion scores for dataset 1

We only considered criteria affected by the individual simplification strategies. For strategy 1, it concerns criterion 6 (absence of injuries). For strategy 2 (predicting on-farm measures from slaughter plant measures) it concerns criterion 3 (comfort around resting) and criterion 6 (good health). Scores for criteria will not differ between GS1 and GS2.

3.5.4.1 Comparing golden standard with strategies for simplification: criterion level, dataset 1

Table 18 shows the results of the comparison between the golden standard and simplification strategy 1 (predicting gait scores from hock burns on-farm). This table shows that results approximate equality between the golden standard and the simplified model. The correlation between the criterion score calculated according to the golden standard and the simplified model is high, as is shown in figure 17.

Table 18. Comparison of a simplified model (strategy 1, predicting gait scores from hock burn on-farm) for criterion 6 (absence of injuries). For explanation of terms, see 2.8.3.2. Empty cells indicate that that there are no values for this category. Rsp: Spearman rank correlation coefficient. Est=estimated agreement; low=lower limit of confidence interval; upp=upper limit of confidence interval.

Criterion 6		20.0			55.0			80.0		
		90% Conf. interval			90% Conf. interval			90% Conf. interval		
		est.	lower	upper	est.	lower	upper	est.	lower	Upper
	%equal	88.2	82.8	92.3	100.0	97.9	100.0	100.0	97.9	100.0
	%se	94.1	88.6	97.4	100.0	97.9	100.0	100.0	97.9	100.0
Rsp=0.81	%sp	74.4	61.2	84.9						

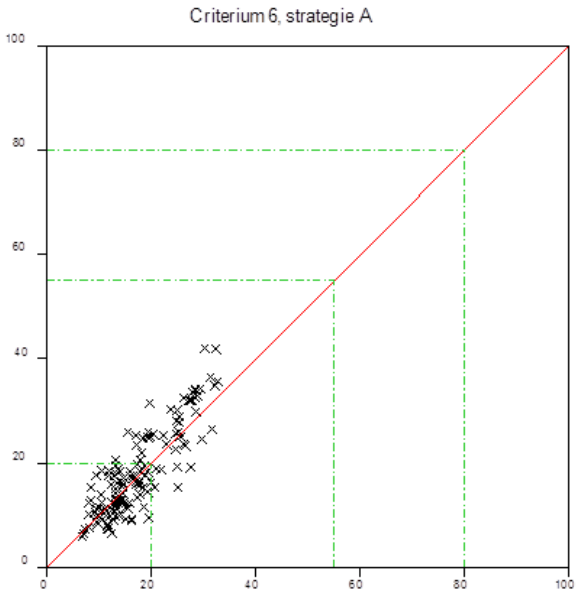


Figure 17. Score for criterion 6 (absence of injuries) for the golden standard (GS) on the Y-axis against the simplified model (strategy 1, prediction of gait scores from hock burn on-farm) on the X-axis.

Table 19 shows the results of the comparison of the golden standard with simplification strategy 2 (predicting on-farm measures from slaughter plant measures) for criterion 6 (absence of injuries). The table shows that the criterion scores for the full model and the simplified model are almost equal, although the specificity is not very high and the correlation is moderate (see figure 18).

Table 19. Comparison of a simplified model (strategy 2, prediction of on-farm measures from slaughter plant measures) for criterion 6. For explanation of terms, see 2.8.3.2. Empty cells indicate that that there are no values for this category. Rsp: Spearman rank correlation coefficient. Est=estimated agreement; low=lower limit of confidence interval; upp=upper limit of confidence interval

Criterion 6		20.0			55.0			80.0		
		90% Conf. interval			90% Conf. interval			90% Conf. interval		
		est.	lower	upper	est.	lower	upper	est.	lower	Upper
	%equal	86.1	80.5	90.6	100.0	97.9	100.0	100.0	97.9	100.0
	%se	95.0	89.9	98.0	100.0	97.9	100.0	100.0	97.9	100.0
Rsp=0.68	%sp	65.1	51.5	77.1						

.

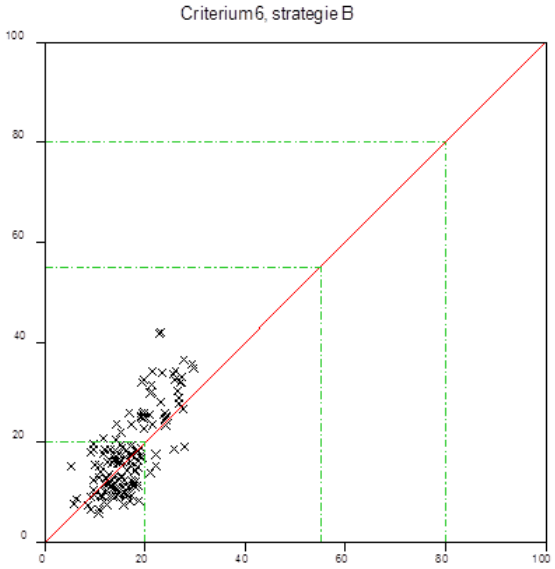


Figure 18. Score for criterion 6 (absence of injuries) for the golden standard (GS) on the Y-axis against the simplified model (strategy 2, predicting on-farm measures from slaughter plant measures) on the X-axis.

Table 20 shows the results of the comparison of the golden standard with simplification strategy 2 (predicting on-farm measures from slaughter plant measures) for criterion 3 (comfort around resting). This table shows that the criterion scores for the full model and the simplified model are equal, although the confidence interval for sensitivity for flocks in the category ≤ 20 is large, probably due to the small number of farms in this category. The correlation between the golden standard and the simplified model is high (see also figure 19).

Table 20. Comparison of a simplified model (strategy 2, prediction of on-farm measures from slaughter plant measures) for criterion 3 (comfort around resting). For explanation of terms, see 2.8.3.2. Empty cells indicate that there are no values for this category. Rsp: Spearman rank correlation coefficient. Est=estimated agreement; low=lower limit of confidence interval; upp=upper limit of confidence interval

Criterion 3		20.0			55.0			80.0		
		90% Conf. interval			90% Conf. interval			90% Conf. interval		
		est.	lower	upper	est.	lower	upper	est.	lower	upper
	%equal	100.0	97.9	100.0	97.2	93.6	99.0	100.0	97.9	100.0
	%se	100.0	36.8	100.0	100.0	97.6	100.0	100.0	97.9	100.0
Rsp=0.91	%sp	100.0	97.8	100.0	77.8	56.1	92.0			

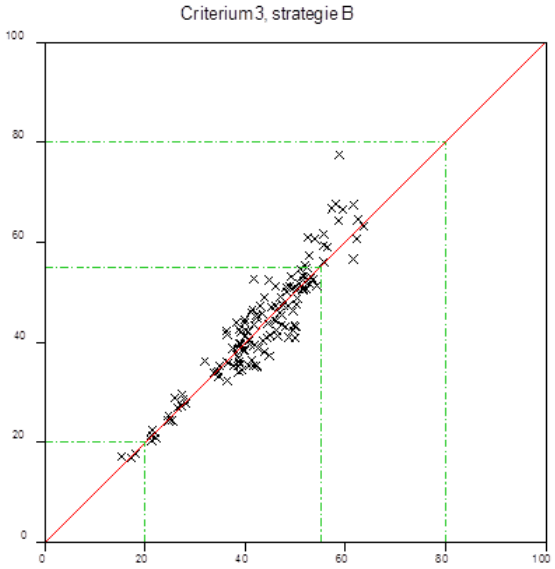


Figure 19. Score for criterion 3 (comfort around resting) for the golden standard (GS) on the Y-axis against the simplified model (strategy 2, predicting on-farm measures from slaughter plant measures) on the X-axis.

3.5.5 Comparing golden standard with strategies for simplification: criterion level, dataset 2

For results of simplification at the level of individual criterion scores for dataset 2, breast burn measures replaced by hock burn measures at the slaughter plant, we refer the reader to appendix 9. Results with dataset 2 do not differ from results with dataset 1.

4 Discussion and conclusions

4.1 Differences between flocks in standard and alternative systems

Standard and alternative rearing systems not only differ in environment and management, but also in the type of bird used. Both environment and breed may affect bird welfare, but it has also been shown that there is an interaction between the environment and genetic traits that may affect bird welfare (EFSA, 2010). According to the EFSA (2010) there is a lack of robust scientific data for Europe on outcome of welfare indicators for commercial broilers. The current study on a large number of flocks provides important supply of additional information.

The data collected in this study show that there are large differences in almost all measures between flocks in standard and alternative rearing systems. In general, flocks in alternative systems had much better welfare estimates in terms of lower incidence of foot pad dermatitis, hock burns, breast blisters, a much better gait score, and better scores for cleanliness. Flocks with slow growing birds in alternative systems displayed less panting and had a more positive score for the qualitative behaviour assessment, compared to standard flocks. Only for rejections and mortality were there no significant differences to be found between alternative and standard systems, and touch test scores indicated a higher level of fearfulness in flocks in alternative systems.

Wet litter is considered to be an important influencing factor for the occurrence of contact dermatitis (foot pad dermatitis, hock burn and breast blisters) (Shepherd and Fairchild, 2010). On average, litter quality was worse for standard systems than for alternative systems (average litter score 2.45 versus 1.66 respectively), which may explain the differences in incidence of contact dermatitis between the different rearing systems. Wet litter can be prevented by management, such as feed type, water management, ventilation, environmental temperature and litter type (see Shepherd and Fairchild, 2010 for a review). Monitoring programmes for foot pad dermatitis in combination with management advice, as introduced in Sweden en Denmark, have substantially decreased the incidence of foot pad dermatitis in broilers in these countries (Algers and Berg, 2001; Berg and Algers, 2004). It has been shown that flocks with access to an outdoor range in the UK had a higher incidence of foot pad dermatitis than those kept indoors (also using slow growing types of broilers) (Pagazaurtundua and Warriss, 2006). Because we only had a few systems with outdoor range included in the assessment, we were unable to provide a separate analysis of the results of these systems. The few Dutch flocks with outdoor range had very high values for severe FPD (on average 83%), but the UK outdoor flocks assessed in 2008 had much lower FPD (on average 5.9%) levels. Probably the quality of the surface of the outdoor range plays an important role in this.

Results on incidence of foot pad dermatitis and hock burn in standard and alternative systems are in line with results from earlier studies (Arnould, unpublished data in EFSA, 2010; Cooper, unpublished data in EFSA, 2010). Studies that reared different genotypes under the same conditions showed that slow growing birds had less FPD and hock burns compared to fast growing birds (Van Middelkoop et al., 2002), and differences between different fast growing genotypes have also been shown (Allain et al., 2009; Sanotra et al., 2003; Van Harn, unpublished data), indicating that not only environmental conditions but also genetic background may affect levels of contact dermatitis.

The effect of growth rate and environmental factors on leg disorders have been reviewed by EFSA (2010). In general, a higher growth rate is associated with a higher incidence of lameness (Knowles et al., 2008). We have showed here that the percentage of lame birds (having a gait score of three and more) was much higher in standard systems with fast growing birds than those from alternative systems with slower growing birds. Compared to data from earlier studies, as reviewed in EFSA (2010), the incidence of lame birds is high in Dutch flocks in standard systems .

The better litter scores of alternative systems are reflected in better cleanliness scores compared to the standard rearing systems. Dirty feathers are mainly caused by wet and sticky litter.

It is generally known that fast growing broiler strains are more susceptible to heat stress as compared to slow growing broiler strains (EFSA, 2010). But management factors such as ventilation and stocking density may also induce heat stress in broilers. It has been shown that panting behaviour increased with increasing stocking density (McLean et al., 2002). It is highly probable that both genetic background

together with environmental factors such as stocking density may explain the higher incidence of panting in standard systems compared to alternative systems.

Although flocks in alternative systems had on average a more positive score in the qualitative behaviour assessment (QBA), they had a lower score in the touch test, indicating more fearfulness for humans. Although these findings are not necessarily contradictory , we did not have the impression that birds in alternative systems were more fearful than those in standard rearing systems. We had the impression that the touch test rather measured mobility of the birds and not fearfulness. This is discussed further in section 4.3.

4.2 Correlations between animal -based measures

Correlations between animal-based measures were the starting point for definition of strategies for simplification. For on-farm measures the highest correlation coefficient was found between severe gait scores and severe hock burn. Hock burn typically arises when broilers increase the time spent sitting on their hocks (Haslam et al., 2007). The incidence of hock burn increases with increasing age and weight of the broilers (Hepworth et al., 2010). Broilers decrease their activity levels with age and a strong decrease can be observed after three weeks of age (Newberry and Hall, 1990; Shields et al., 2005). This decreased activity is caused by the increased weight of the birds, which causes heat stress when birds become very active, but also increases the risk for leg problems resulting in a high gait score (Corr et al., 2003; Knowles et al., 2008). It can therefore be expected that birds with leg problems decrease their activity and spent more time sitting on their hocks, which increases the risk for hock burn when litter quality is poor.

In general, it can be expected that there is a relationship between scores for cleanliness, hock burn and foot pad dermatitis. Poor litter quality which is the most important factor causing contact dermatitis (Shepherd and Fairchild, 2010) is also thought to cause soiled birds.

A high correlation was found between foot pad dermatitis scores at the slaughter plant and foot pad dermatitis scores on-farm, despite the fact that assessment of foot pad dermatitis on- farm is more difficult than assessment at the slaughter plant due to dirty feet and low light levels on- farm (De Jong et al., 2011). In a previous study with a very small number of flocks no correlation was observed between foot pad dermatitis assessment on-farm and at the plant (De Jong et al., 2011), but in the current project with a larger number of broiler flocks the correlation was found to be high.

4.3 Calculation of flock scores

The high variation between flocks in scores for most of the measures was not reflected in variation in the end scores for flocks. The majority of flocks received the classification acceptable, while there were large differences in individual measures. Large differences were also found for measures between flocks from standard and alternative rearing systems. One aspect that may play a role in the very uniform flock scores is the fact the Welfare Quality® model for calculating flock scores does not allow compensation at criterion and principle level. Another possibility is that the process of calculation of end scores for flocks as described in the broiler assessment protocol, is unable to differentiate between flocks that potentially differ in the level of welfare. This was actually the first time the scoring was carried out on a large set of data of broiler flocks (with high variation in measure scores). A critical review of the calculation method for flock scores is advised prior to further consideration of the use of these end scores in practice.

A remarkable observation when comparing principle scores using the full protocol was that flocks in alternative systems received a low score for principle 4, appropriate behaviour. Scores for principle 4 are based on results of the qualitative behaviour assessment (QBA) and the touch test. Flocks in alternative systems generally received higher QBA scores than flocks in standard systems, but flocks with slow growing birds in alternative systems received low scores for the touch test. Probably, this low score for principle 4 was caused by a low score for the touch test, as no compensation between scores for these criteria is possible (Welfare Quality®, 2009). A low score for the touch test indicates that relatively fewer birds could be touched. We had the impression that the touch test, which was meant to measure fearfulness, measured mobility of the birds rather than fear of humans. It is therefore questionable whether the touch test is indeed a good measure of fearfulness and it is therefore advised

to perform a validation study for this test. Although the touch test has been performed in several flocks, it has not been validated whether or not it really measures fearfulness of humans (Butterworth, personal communication). On an overall level, and within standard systems, correlations have been found between touch test measures and gait score. A high gait score was significantly positively correlated with the number of birds that could be touched and the touch index (data not shown). This supports our impression that touch test scores might be related to walking ability of the birds and we therefore advise a validation of this test prior to large scale usage.

4.4 Simplification strategies

In general, both strategies for simplification, i.e. predicting gait scores from hock burn on-farm and predicting on-farm measures (gait score, cleanliness, foot pad dermatitis and hock burn) from slaughter plant measures (foot pad dermatitis and hock burn), showed close agreement between flock scores and close agreement in scores for individual principles and criteria. In addition, on principle and criteria level, a high correlation was found between the golden standard and the simplified strategies. Agreement was highest when simplification strategies were compared with GS1. For the less tolerant model, GS2, agreement was lower but still sufficiently high. As differences between the original dataset (dataset 1) and the dataset where breast burn scores were replaced by hock burn scores at the slaughter plant (dataset 2) were small, we can focus attention on the results with the original dataset.

The proximity of agreement between the golden standard and the simplification strategies seems to be promising for future use of a simplified assessment protocol for broilers. However, the results of simplification with the current dataset are calculated on flocks of which the majority ends up in the same classification (i.e. acceptable). At flock level, large confidence intervals were sometimes found for sensitivity and specificity, caused by the fact that only a few farms were in some categories (enhanced, not classified) and indicating that we should be careful with definite conclusions regarding simplification. On the other hand, close agreement and high correlations were found at principle as well as criterion level. In cases where flocks were more widely distributed over different categories, smaller confidence intervals were found. Taken together, the results indicate that simplification of the broiler assessment protocol, using either strategy 1 or 2, is encouraging , but requires further validation by testing in new flocks that are preferably better distributed over the different categories.

Both simplification strategies are promising in terms of considerably reducing the time of assessment. The strategy that predicts gait scoring by hock burn scoring on-farm reduces the time spent on the on-farm assessment by approximately one hour, which is equivalent to a 25-33% reduction in performance time. The strategy that predicts on-farm measures from slaughter plant measures is probably even more encouraging in terms of performance time reduction. One option is, for example, to assess the flocks at the slaughter plant on a regular basis (e.g., each flock delivered) and to do the additional assessments on farm (i.e. on behaviour) at a much lower frequency e.g. once per year. As regular assessment of hock burn and (in future) assessment of foot pad dermatitis will be performed for standard flocks housed at the maximum stocking density under the broiler welfare regulation, an assessment procedure will already be in place at slaughter plants which may facilitate the implementation of a broiler welfare assessment protocol. A definite choice for either one of the simplification strategies will also be dependent on practical issues, e.g. planning issues at slaughter plants or possibilities for assessors to score additional measures at the plant.

4.5 Practical experience with the broiler assessment protocol

In the current project we applied the broiler welfare assessment protocol for the first time to a large number of flocks. In general, a full flock on-farm assessment was possible within the time frame as indicated in the broiler assessment protocol, i.e. 3-4 hours (Welfare Quality®, 2009). Assessments at the slaughter plant were very efficient (only about 1 hour per flock) but due to abrupt changes in the slaughter plant planning assessors spent considerable periods waiting before scoring the birds. Thorough training of assessors as conducted during this project is considered essential for reliable assessment. Reduction in training time is not advocated as after the second day of training examination of the assessors revealed that further training was necessary (agreement with golden standard, a well-trained researcher, this was below 75% after two days of training for the majority of assessors but more than 75% after an additional half day of training). Data recording with pda's and uploading data in a

database was very efficient and is advised when applying the protocol in practice. It also facilitates feedback of results to the farmer.

4.6 Conclusions

With respect to possible simplification of the broiler assessment protocol, we conclude that

- Both strategies for simplification of the broiler assessment protocol, i.e. predicting gait scores from hock burn scores on-farm, and predicting on-farm measures from slaughter plant measures, are encouraging in terms of agreement with the golden standard score at flock level, principle level and criterion level;

- Both strategies for simplification of the broiler assessment protocol seem promising in terms of reducing time for assessment and may thus facilitate implementation in practice;

- The simplification strategies should be validated further, preferably in flocks that are more widely distributed over the different categories.

With respect to the outcomes of the measures and flock scores we conclude that

- Large variation between flocks was found for almost all animal-based measures;

- Large differences were found in outcomes between flocks with fast growing birds in standard rearing systems, and flocks with slower growing birds in alternative rearing systems. In general, measures indicate a better welfare level in flocks in alternative rearing systems;

- Most flocks were classified in the same category, i.e. acceptable and only a few flocks were classified enhanced;

- Validation of the touch test is advised.

Acknowledgements

Special thanks to the assessors who put so much effort into the farm and slaughter plant visits. We are very grateful to them for their enthusiasm, their efforts and their motivation to deliver a good job, in the weekends, early morning or evenings. Thanks to (in random order): Cindy Hoeks, Annemae Kremer, Judith Lammers, Erik Schuiling, Theo van Hattum, Jan Jochemsen, Jitske Westra, Sander Lourens, Henk Schilder, Guus Nijeboer, Ido Alferink and Henk Gunnink.

Thanks to Tomas Perez Moya, who did his MSc thesis on a part of this project and put much effort into the data analysis.

Dr. Andy Butterworth, dr. Paolo Ferrari and drs. Roselien Vanderhasselt kindly provided us with data from British, Italian and Belgian flocks included in the analysis. Also thanks to Dr. Andy Butterworth for training the assessors in applying the broiler welfare protocol.

We are very grateful to Ruud Dekker for assisting Monique and Vincent with the planning of the farm and slaughter plant visits in 2011. This was a difficult job, especially because slaughter plant time schedules could change abruptly, sometimes without warning.

Hans van de Heuvel provided us with software for pda's and downloading data to the database, and a format for communication to the farmers. This is essential when a large number of flocks is assessed.

A special thank you also to the statisticians of Biometris who put much effort into the statistical analysis.

Literature

Algers, B., Berg, C., 2001. Monitoring animal welfare on commercial broiler farms in Sweden. Acta Agric. Scand. Sect. A-Anim. Sci., 88-92.

Allain, V., Mirabito, L., Arnould, C., Colas, M., Le Bouquin, S., Lupo, C., Michel, V., 2009. Skin lesions in broiler chickens measured at the slaughterhouse: relationships between lesions and between their prevalence and rearing factors. Br. Poult. Sci. 50, 407-417.

Anonymus, 2009. Afsprakenkader vleeskuikenrichtlijn. www.minlnv.nl/txmpub/files/?p_file_id=43482Similar

Berg, C., Algers, B., 2004. Using welfare outcomes to control intensification: the Swedish model, in: Weeks, C.A., Butterworth, A. (Eds.), Measuring and auditing broiler welfare, CABI, Oxford, pp. 223-229.

Blokhuis, H.J., Veissier, I., Miele, M., Jones, B., 2010. The Welfare Quality® project and beyond: Safeguarding farm animal well-being. Acta Agric. Scand. Sect. A-Anim. Sci. 60, 129-140.

Botreau R., Buist, W., Butterworth, A., Perny, P., Veissier, I., 2009. Reports on the construction of welfare criteria for different livestock species. Part 3 – subcriteria construction for broilers on farm. Welfare Quality® Deliverable 2.8c.

Corr, S.A., Gentle, M.J., McCorquodale, C., Bennett, D., 2003. The effect of morphology on walking ability in the modern broiler: a gait analysis study. Anim. Welf. 12, 159-171.

De Jong, I.C., Reimert, H.G.M., Vanderhasselt, R., Gerritzen, M.A., Gunnink, H., Van Harn, J., Hindle, V.A., Lourens, A., 2011. Ontwikkelen van methoden voor het monitoren van voetzoollaesies bij vleeskuikens. Wageningen UR Livestock Research Rapport 463.

EFSA, 2010. Scientific opinion on the influence of genetic parameters on the welfare and the resistance to stress in commercial broilers. EFSA Journal 8: 1666. Doi: 10.2903/j.efsa.2010.1666.

Haslam, S.M., Knowles, T.G., Brown, S.N., Wilkins, L.J., Kestin, S.C., Warriss, P.D., Nicol, C.J., 2007. Factors affecting the prevalence of foot pad dermatitis, hock burn and breast burn in broiler chicken. Br. Poult. Sci. 48, 264-275.

Hepworth, P.J., Nefedov, A.V., Muchnik, I.B., Morgan, K.L., 2010. Early warning indicators for hock burn in broiler flocks. Avian Pathol. 39, 405-409.

Kestin, S.C., Knowles, T.G., Tinch, A.E., Gregory, N.G., 1992. Prevalence of leg weakness in broiler chickens and its relationship with genotype. The Veterinary Record 131, 190-194.

Knowles, T.G., Kestin, S.C., Haslam, S.M., Brown, S.N., Green, L.E., Butterworth, A., Pope, S.J., Pfeiffer, D., Nicol, C.J., 2008. Leg Disorders in Broiler Chickens: Prevalence, Risk Factors and Prevention. Plos One 3.

Manten, A., De Jong, I.C., 2011. Protocollen voor het meten van dierenwelzijn. V-focus special 5: 16-17.

McLean, J.A., Savory, C.J., Sparks, N.H.C., 2002. Welfare of male and female broiler chickens in relation to stocking density, as indicated by performance, health and behaviour. Anim. Welf. 11, 55-73.

Newberry, R.C., Hall, J.W., 1990. USE OF PEN SPACE BY BROILER-CHICKENS - EFFECTS OF AGE AND PEN SIZE. Applied Animal Behaviour Science 25, 125-136.

Pagazaurtundua, A., Warriss, P.D., 2006. Levels of foot pad dermatitis in broiler chickens reared in 5 different systems. Br. Poult. Sci. 47, 529-532.

Sanotra, G.S., Berg, C., Lund, J.D., 2003. A comparison between leg problems in Danish and Swedish broiler production. Anim. Welf. 12, 677-683.

Shepherd, E.M., Fairchild, B.D., 2010. Footpad dermatitis in poultry. Poultry Science 89, 2043-2051.

Shields, S.J., Garner, J.P., Mench, J.A., 2005. Effect of sand and wood-shavings bedding on the behavior of broiler chickens. Poultry Science 84, 1816-1824.

Van Middelkoop, K., Van Harn, J., Wiers, W.J., Van Horne, P., 2002. Slower growing broilers pose lower welfare risks. World Poultry 18: 20-21.

Welfare Quality®, 2009. Welfare Quality® assessment protocol for poultry (broilers, laying hens). Welfare Quality® Consortium, Lelystad, The Netherlands.

Appendices

Appendix 1. Copy of a form with feedback of the results of a visit to the farmer.

Welfare Quality - Welzijnsmeting vleeskuikenbedrijf 12-May-2011

Bedrijf	[redacted]
Datum bezoek	26-Apr-2011
Stalnummer	1
Waarnemer(s)	ingrid

			Score	Oordeel (1)
Welzijnsprincipe goede voeding				
Afwezigheid van langdurende honger	(Slachterijrapport nog niet binnen)			
Afwezigheid van langdurende dorst	Aantal kuikens per	1: Nipples	0.0	+
Welzijnsprincipe goede huisvesting				
Rust en comfort	Bevuiling	Niet/licht bevuild	86%	+
		Matig bevuild	14%	
		Ernstig bevuild	0%	
	Strooiselkwaliteit		1.2	+
	Stof		2: Little	+
Thermale omgeving	Hijgende kuikens		0.0%	+
Bewegingsgemak	Bezetting		24.1 kg/m2	+
Welzijnsprincipe goede gezondheid				
Geen verwondingen	Loopscore	% Score > 2	0.0	+
	Voetzoolleasies bdr	Geen	100.0	+
		Mild	0.0	
		Ernstig	0.0	
	Voetzoolleasies sh	Geen	98.0	+
		Mild	2.0	
		Ernstig	0.0	
	Brandhakken bdr	Ernstig	0.0	+
	Brandhakken sh	Ernstig	0.6	+
	Borstblaren/irritatie	Ernstig	2.4	0
Afwezigheid van ziekten	Mortaliteit		1.2	+
Welzijnsprincipe goed gedrag				
Goede mens-dier relatie	Angst voor mensen		24.5	-
Positieve emotionele status (2)			94.0	+

(1) Oordeel: + Uw score is goed / bovengemiddeld. 0 Uw score is gemiddeld. - Uw score is slechter dan gemiddeld. Scores worden vergeleken met wat uit het verleden voor deze parameter bekend is uit onderzoek op bedrijven in dezelfde categorie.

(2) Emotionele status: + Het koppel kan gekarakteriseerd worden als rustig, ontspannen, comfortabel, energiek en vriendelijk; - Het koppel kan gekarakteriseerd worden als onrustig, gespannen, verveeld en opgewonden

Pag 1 WUR - Livestock Research

Appendix 2. Datasets used for expert consultation for criterion 7.

Table 2.1: Virtual dataset – birds rejected at the slaughterhouse = rejections (% birds)

	Rejections (% of birds)	***Score 100=perfect***
Farm 1	0	
Farm 2	0.5	
Farm 3	1	
Farm 4	1.5	
Farm 5	3	
Farm 6	5	
Farm 7	8	
Farm 8	15	
Farm 9	40	
Farm 10	80	
Farm 11	100	

Table 2.2: Virtual dataset – total mortality (% of birds in flock)

	Mortality (% of birds)	***Score 100=perfect***
Farm 1	0	
Farm 2	1	
Farm 3	2	
Farm 4	3	
Farm 5	4	
Farm 6	5	
Farm 7	8	
Farm 8	15	
Farm 9	40	
Farm 10	80	
Farm 11	100	

Table 2.3: Dataset – criterion score “Absence of disease”

	Rejections (welfare score between 0 and 100)	**Mortality (welfare score between 0 and 100)**	**Resulting overall score for “Absence of disease”**
Farm 1	25	75	
Farm 2	40	60	
Farm 3	50	50	
Farm 4	60	40	
Farm 5	75	25	

Experts consulted for % rejection: A. Butterworth, F. Tuyttens, I.C. de Jong
Experts consulted for % mortality: F. Tuyttens, A. Lourens, T. van Niekerk, I.C. de Jong

Appendix 3. Graphs showing variation between flocks for individual measures.

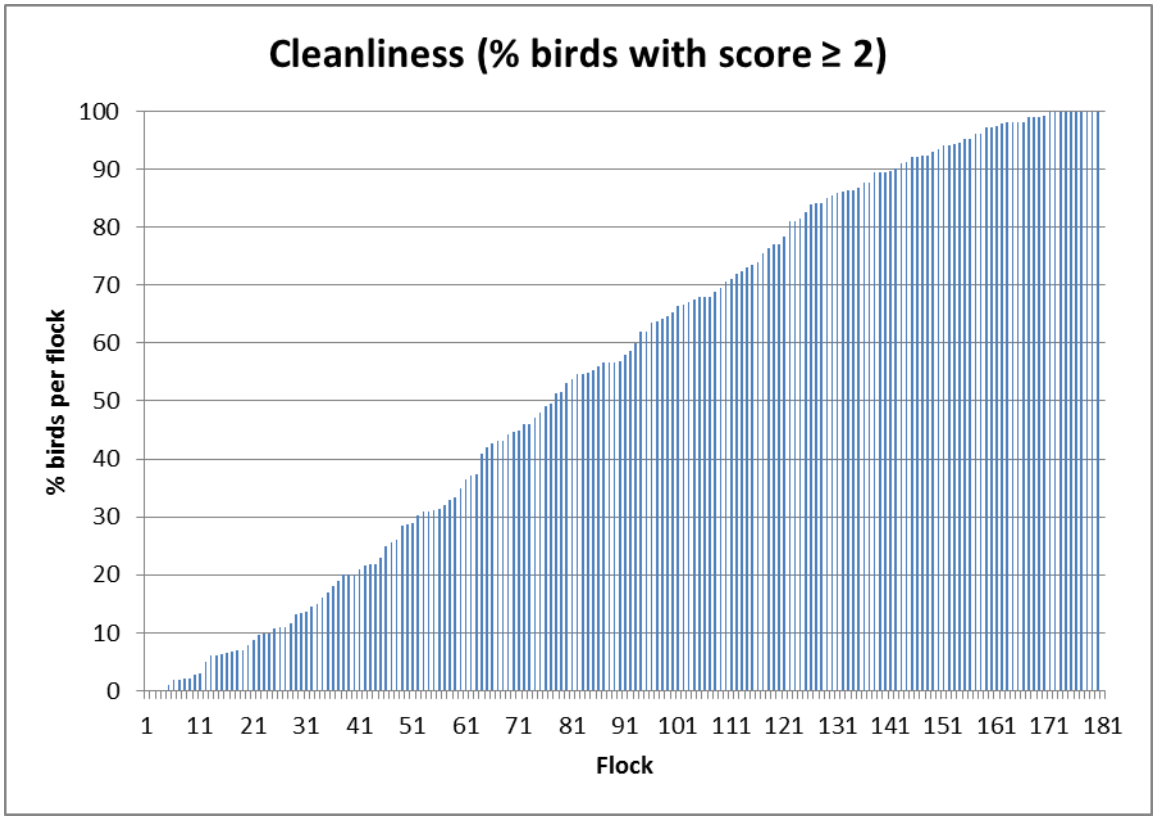


Figure 3.1. Variation between flocks in birds scored as dirty.

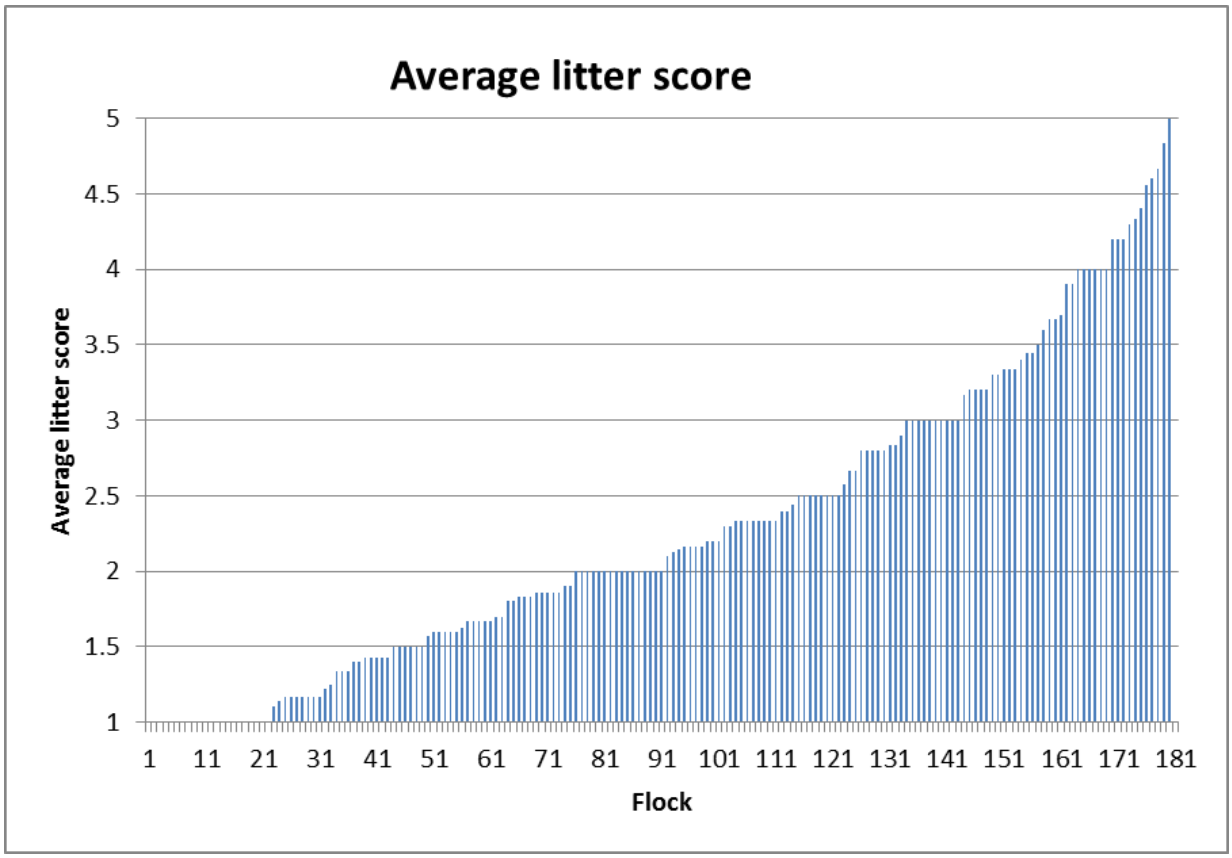


Figure 3.2. Variation between flocks in litter score (litter scores between 1-5).

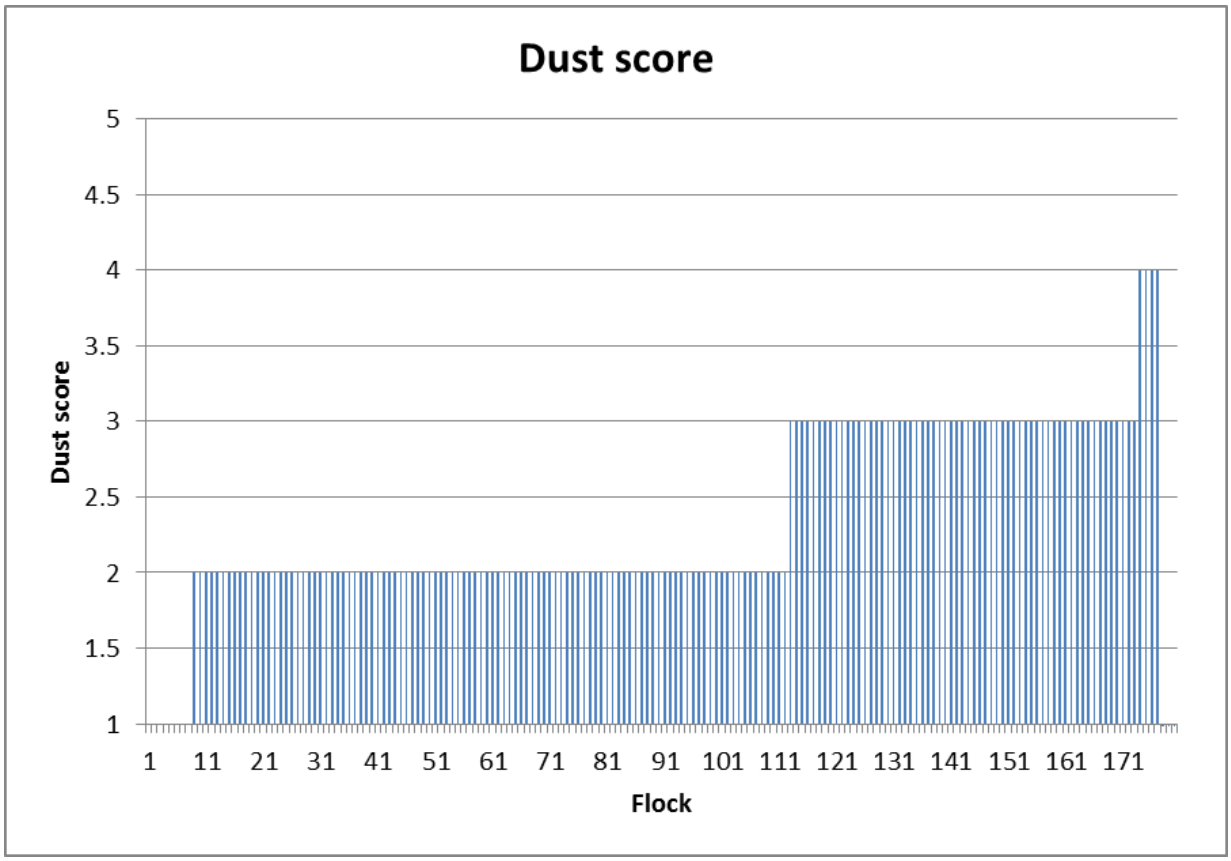


Figure 3.3. Variation between flocks in dust score (dust scores between 1 and 5).

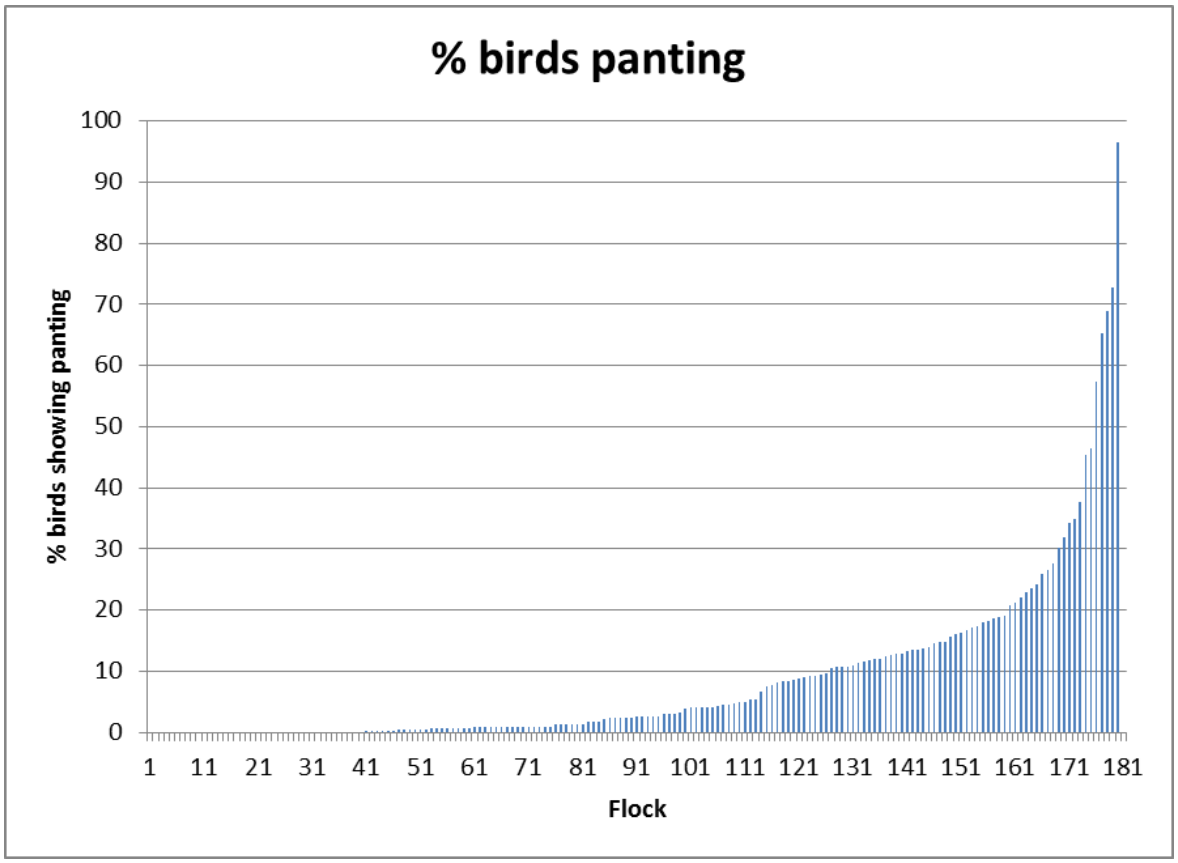


Figure 3.4. Variation between flocks in % of birds panting.

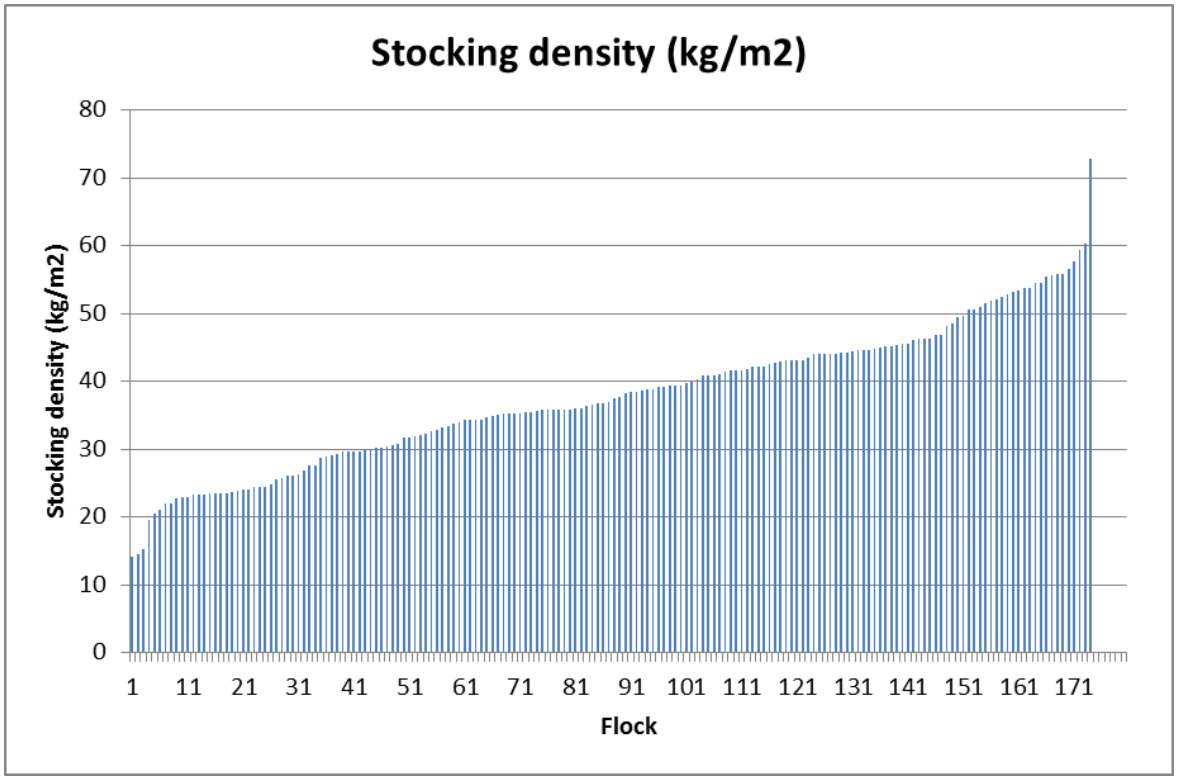


Figure 3.5. Variation between flocks in stocking density. Stocking density was calculated according to information provided by the farmer on number of birds/m^2 and the actual weight of the birds. For flocks measured in 2008, there was no maximum level of stocking density whereas the 2011 flocks were officially limited to a maximum stocking density of 42 kg/m^2.

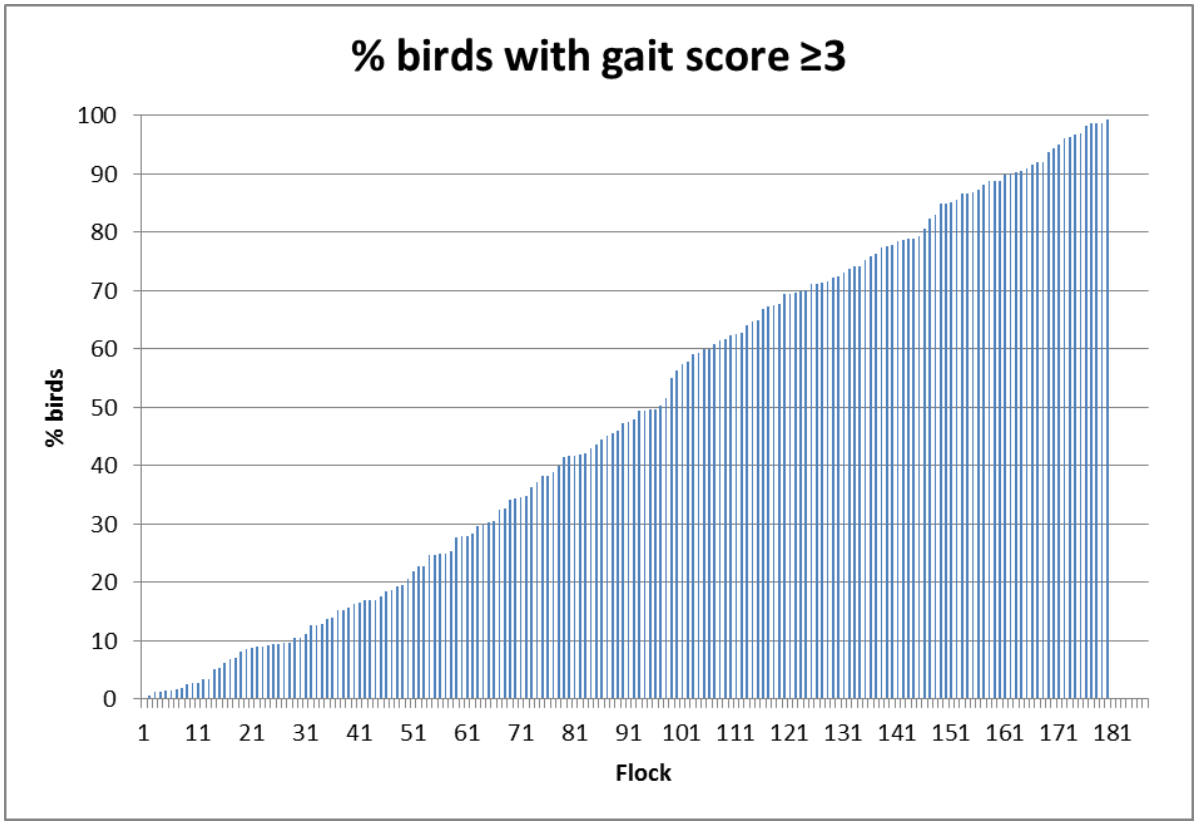


Figure 3.6. Variation between flocks in percentage of birds being severely lame.

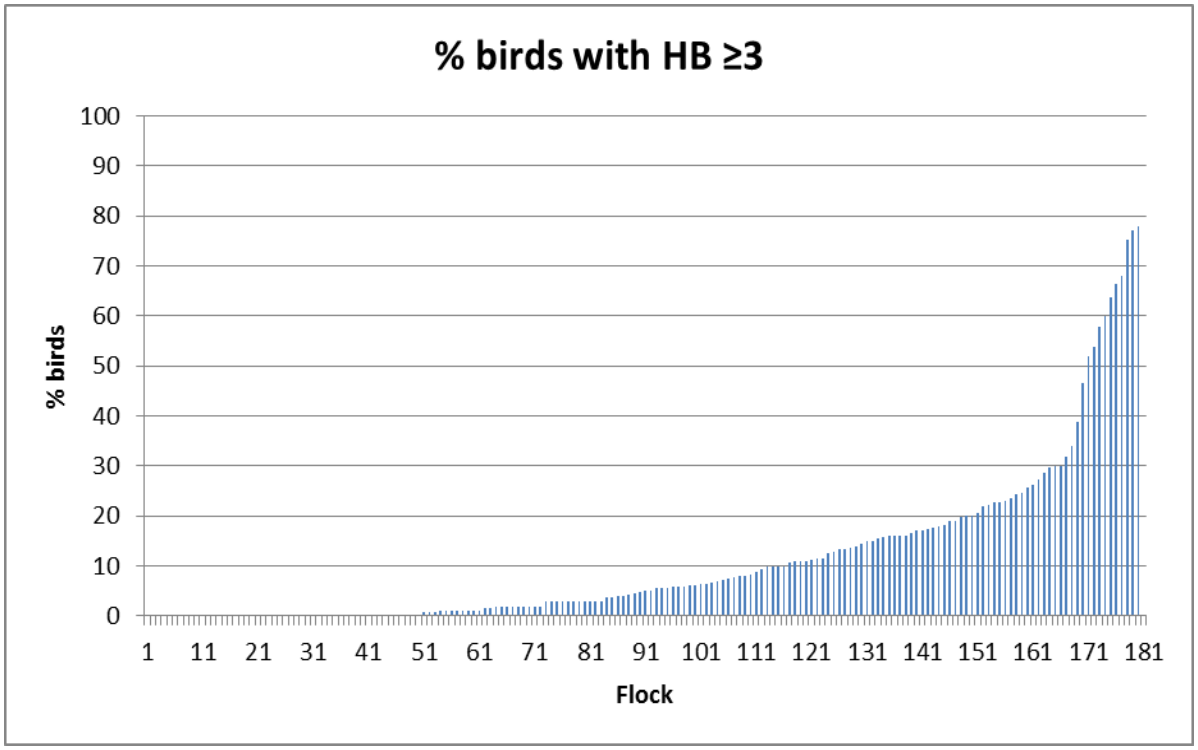


Figure 3.7. Variation between flocks in birds with severe hock burn.

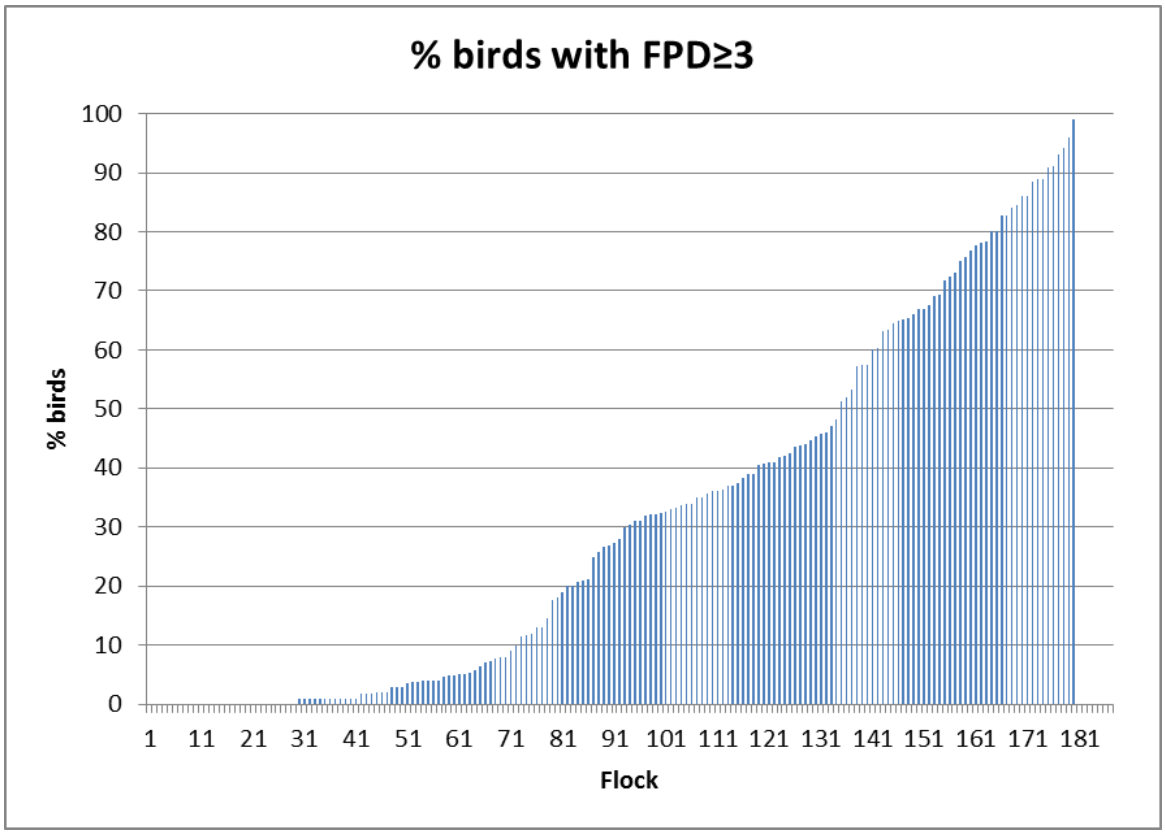


Figure 3.8. Variation between flocks in birds with severe foot pad dermatitis.

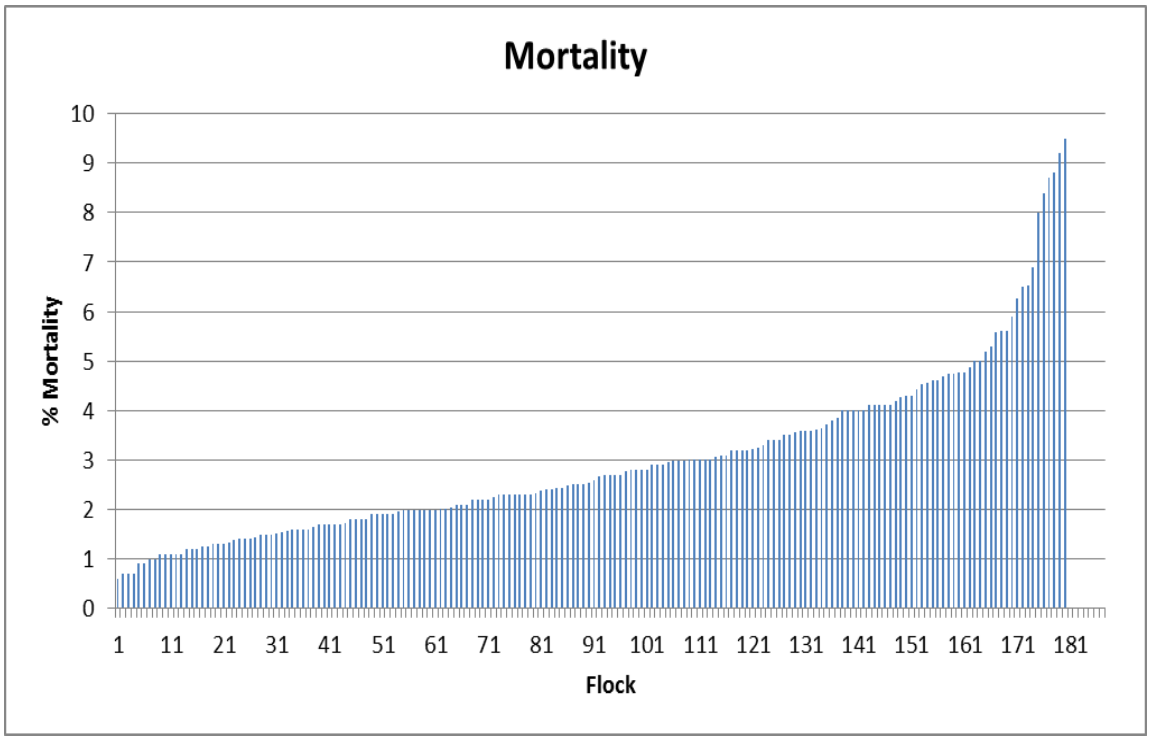


Figure 3.9. Variation between flocks in total mortality.

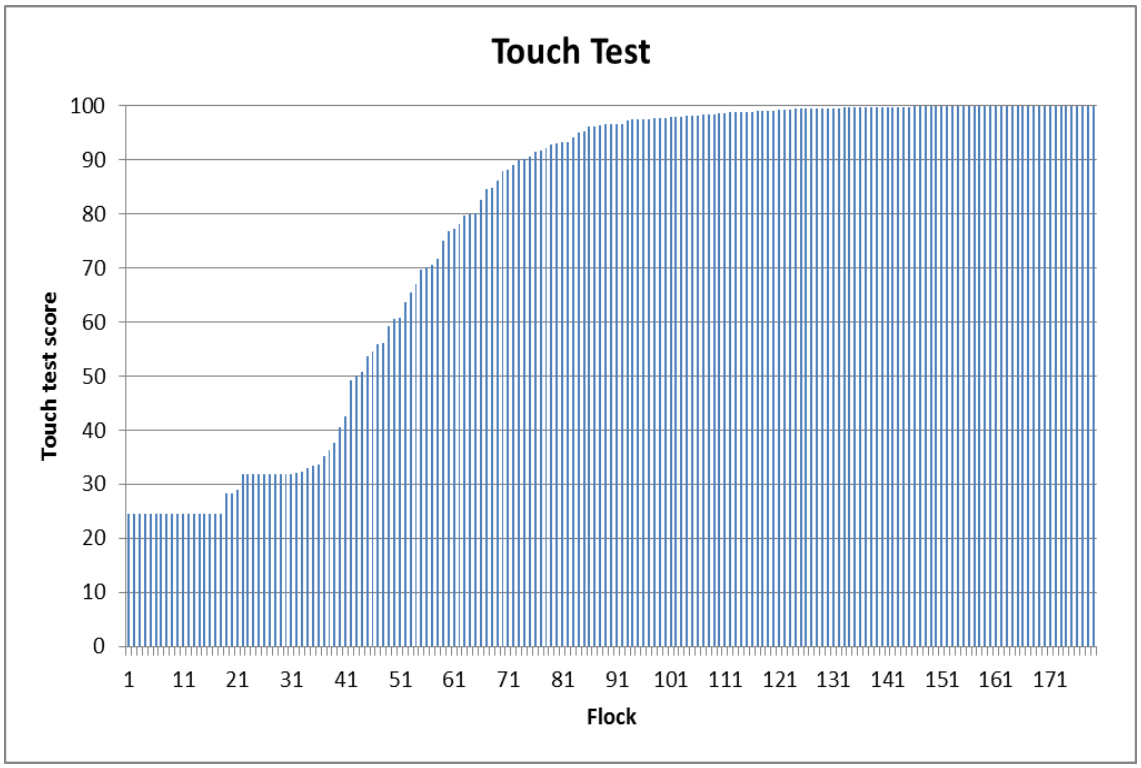


Figure 3.10. Variation between flocks in touch test scores.

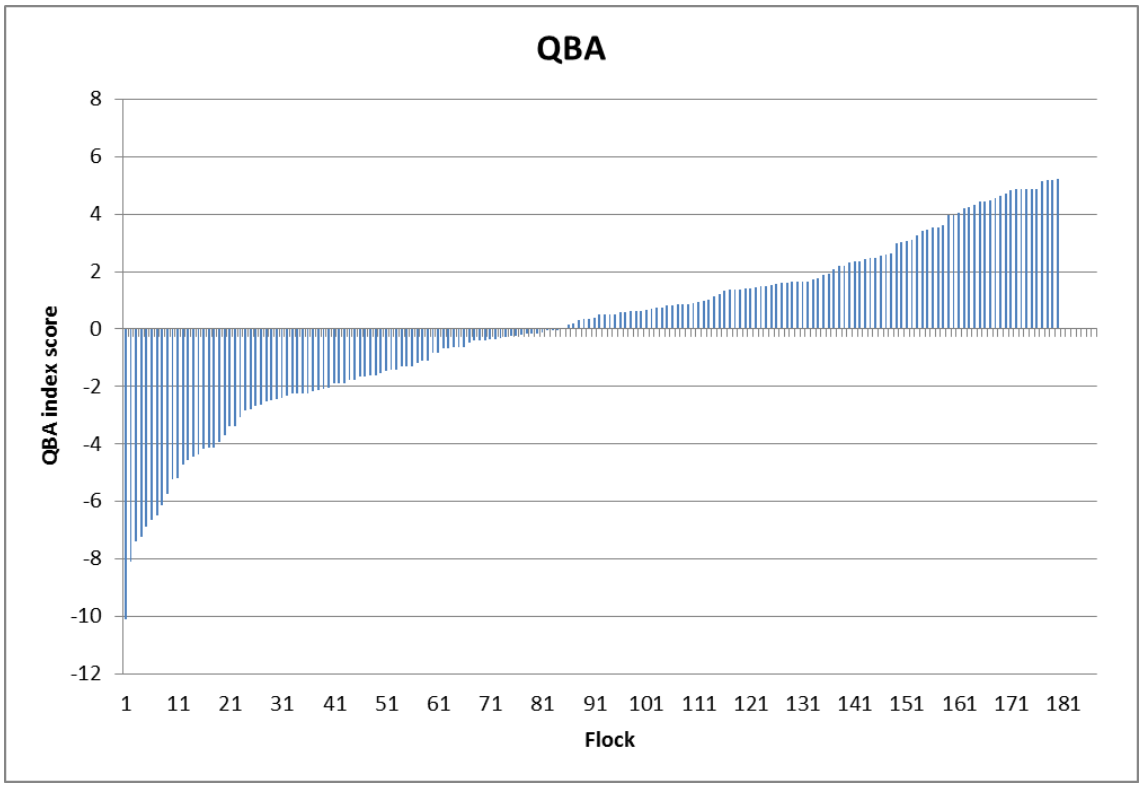


Figure 3.11. Variation between flocks in QBA index.

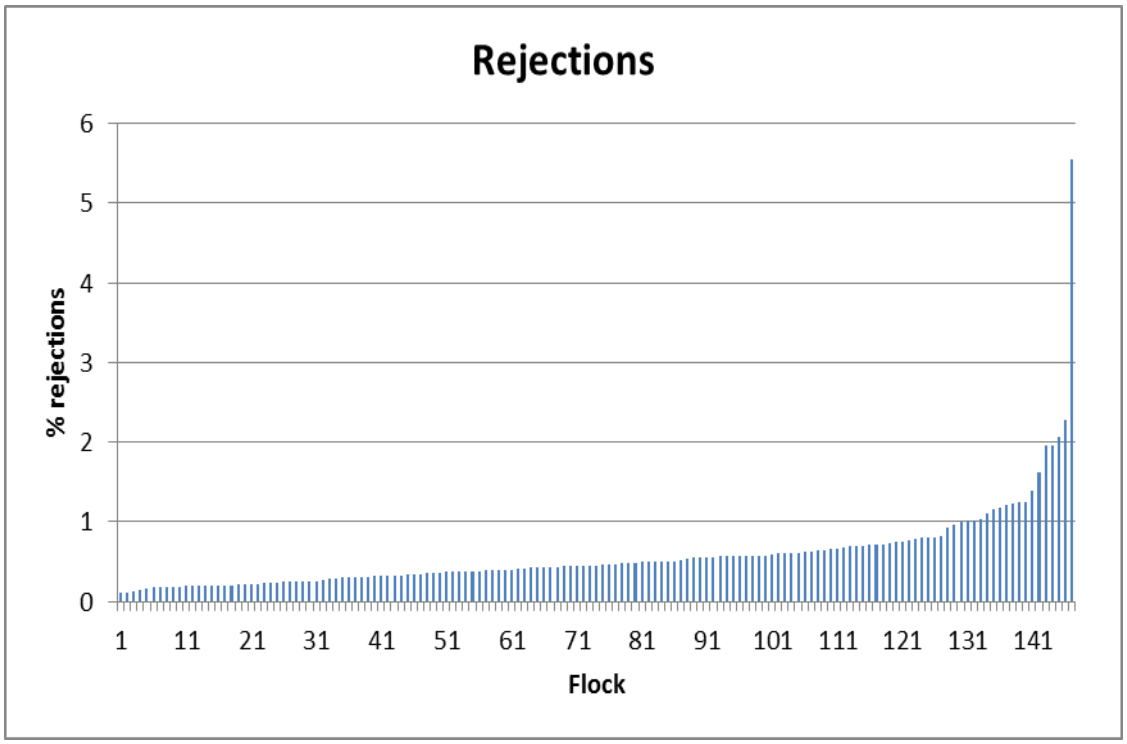


Figure 3.12. Variation between flocks in rejection percentage.

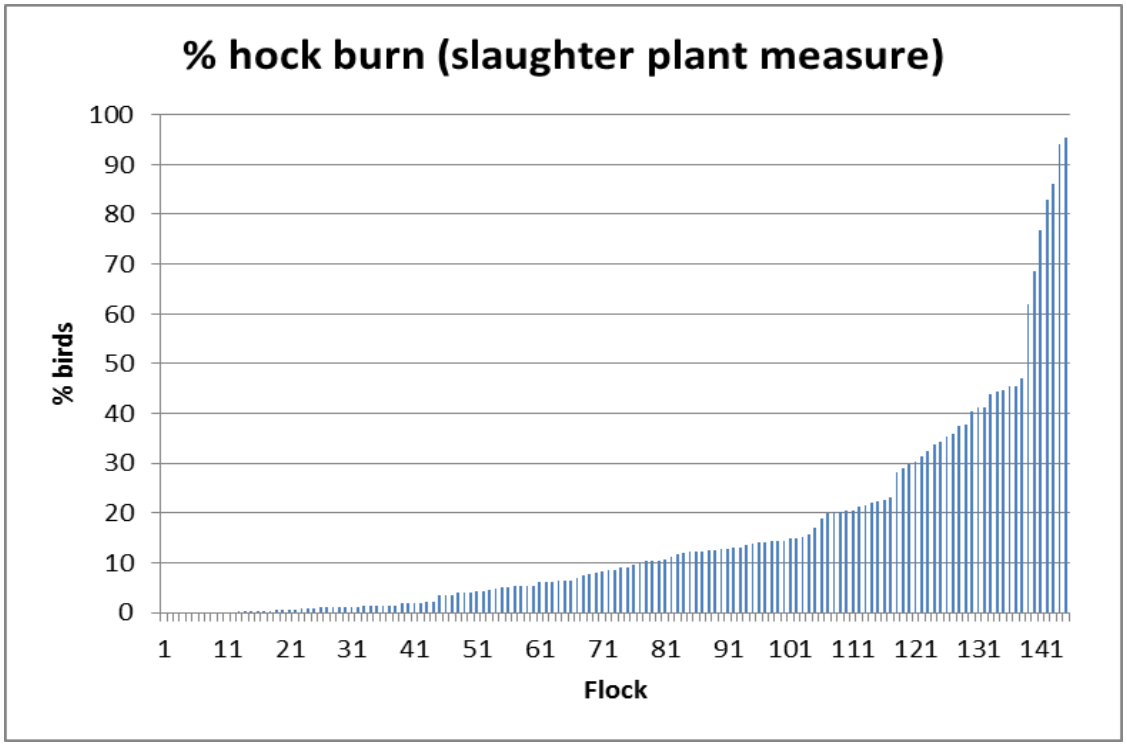


Figure 3.13. Variation between flocks in hock burns scored at the slaughter plant.

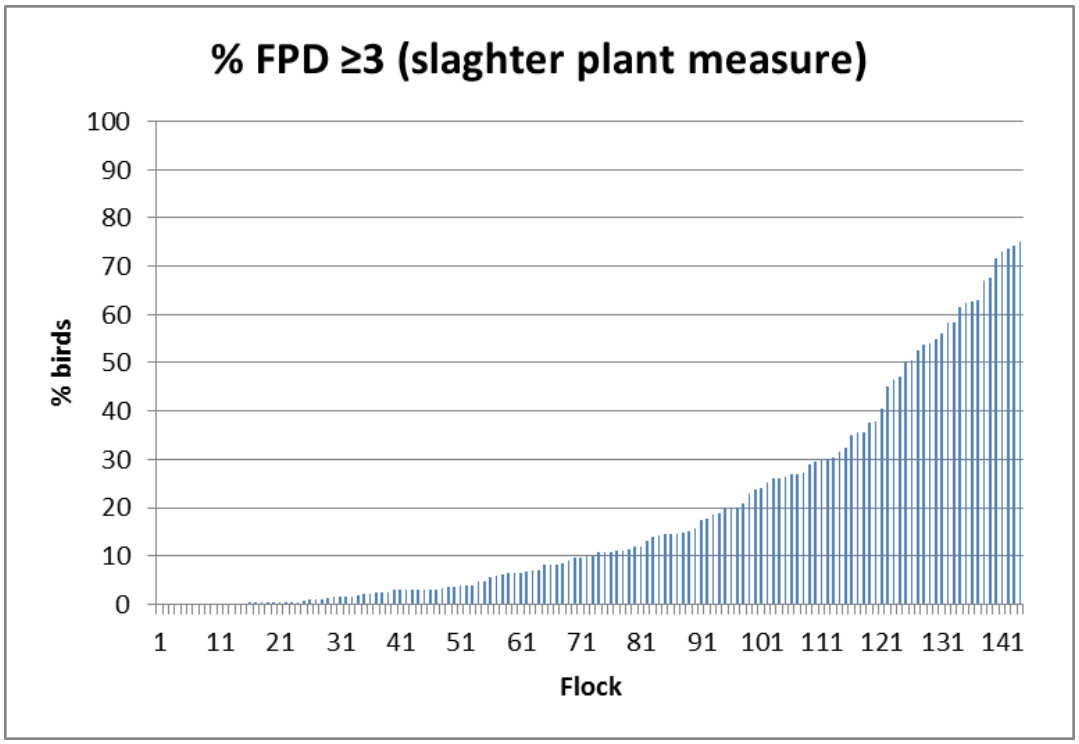


Figure 3.14. Variation between flocks in severe foot pad dermatitis scored at the slaughter plant.

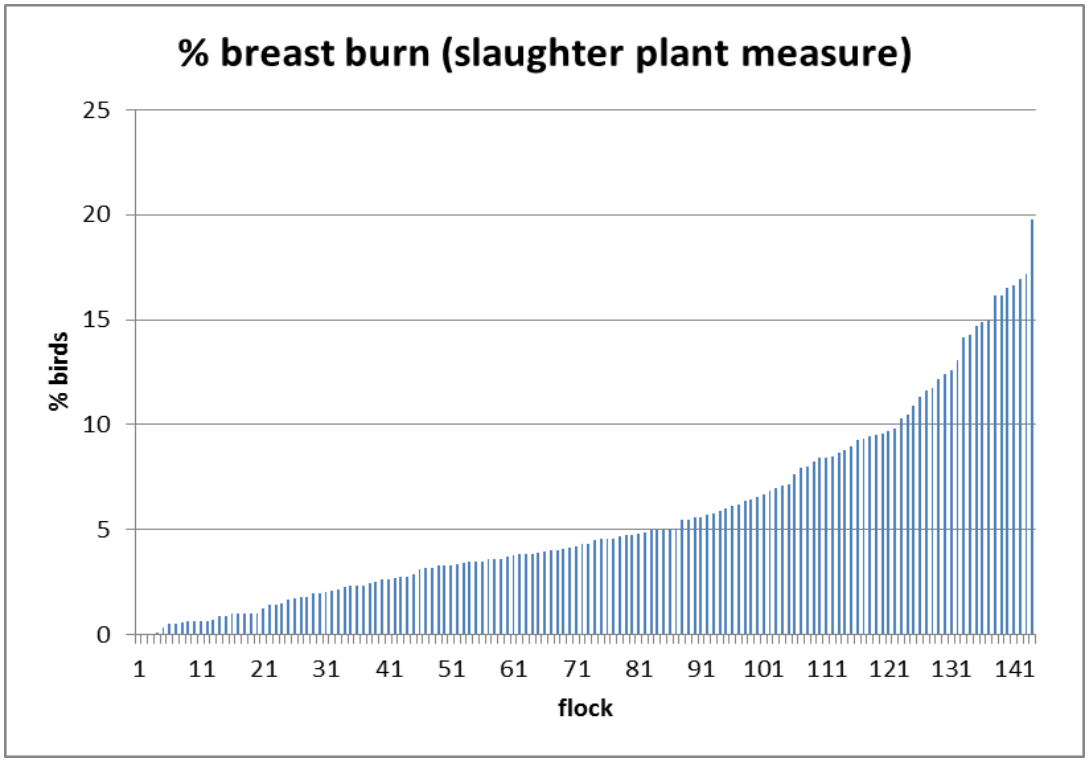


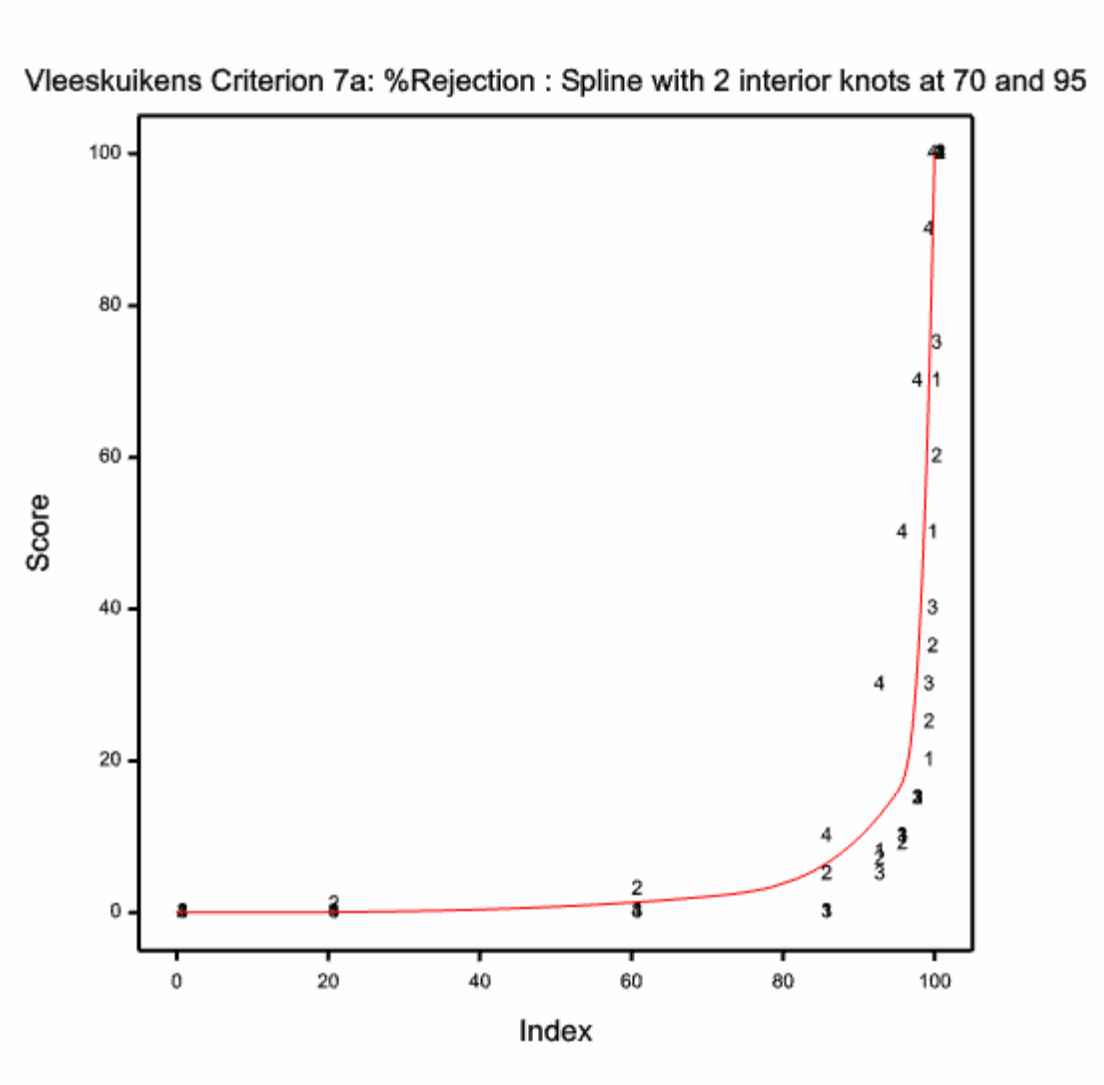
Figure 3.15. Variation between flocks in breast burns scored at the slaughter plant.

Appendix 4. Score for criterion 7, modified calculation based on measures of rejections and mortality only.

Rejection scores of four experts for different values of % of rejections:

```
read       %reject, Exp[1...4],           Mean
  0       100   100     100     100       100
  0.5      70    60      75     100        76.25
  1        50    35      40     100        56.25
  1.5      20    25      30      90        41.25
  3        15    15      15      70        28.75
  5        10     9      10      50        19.75
  8         8     7       5      30        12.5
 15         0     5       0      10         3.75
 40         0     3       0       0         0.75
 80         0     1       0       0         0.25
100         0     0       0       0         0      :

CALCULATE  Index = 100 - %reject
```



Coefficients:

COEF	VALUE
a1	0.0000000000
b1	0.0000000000
c1	0.0000000000
d1	0.0000061018
a2	-231.6675631157
b2	9.9286098477
c2	-0.1418372835
d2	0.0006815174
a3	-519197.8596370568
b3	16398.3343341204
c3	-172.6513677013
d3	0.6059781032

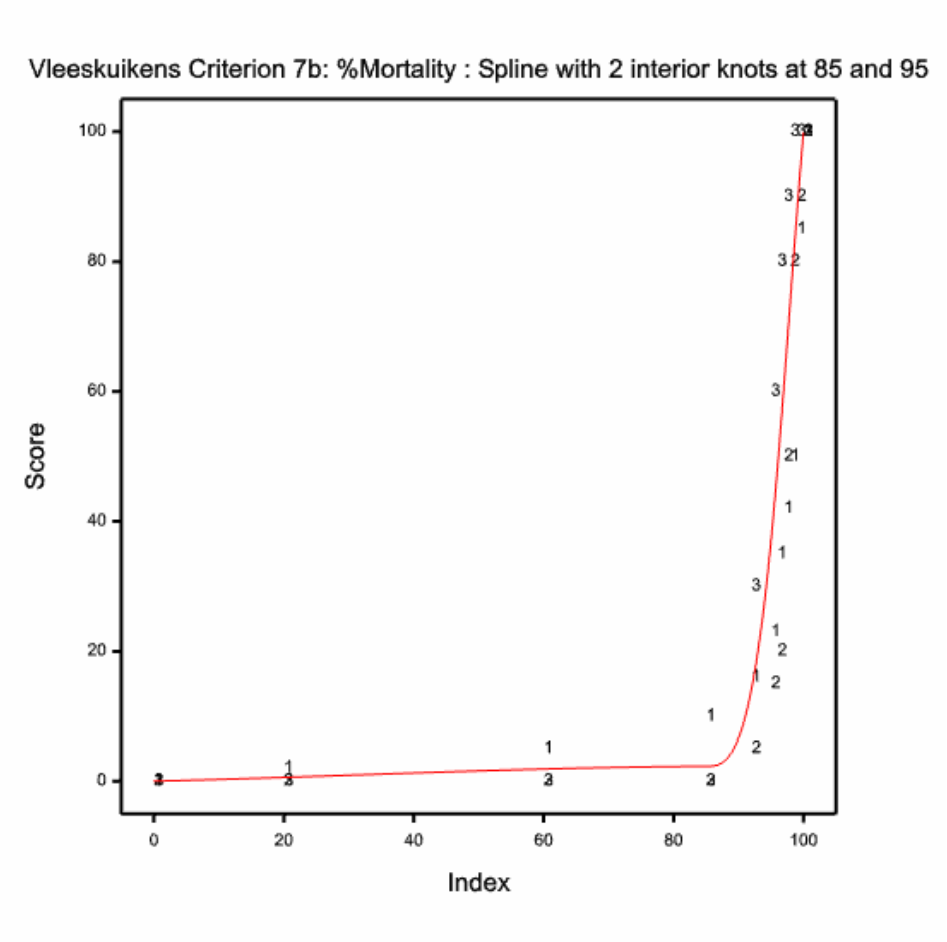
Index = I

When I≤70 then Score = $0.0000061018 \times I^3$
When I≥70 and I≤95 then Score = $-231.668 + (9.9286 \times I) - (0.14183728 \times I^2) + (0.0006815174 \times I^3)$
When I≥95 then Score = $-519197.86 + (16398.334 \times I) - (172.651368 \times I^2) + (0.6059781032 \times I^3)$

Mortality scores of three experts for different values of % mortality:

read	%mort,	Exp[1...3],		Mean
0	100	100	100	100.00
1	85	90	100	91.67
2	50	80	100	76.67
3	42	50	90	60.67
4	35	20	80	45.00
5	23	15	60	32.67
8	16	5	30	17.00
15	10	0	0	3.33
40	5	0	0	1.67
80	2	0	0	0.67
100	0	0	0	0.00 :

```
CALCULATE  Index = 100 - %mort
```



Coefficients:

COEF	VALUE
a1	0.0000000000
b1	0.0241473135
c1	0.0003195893
d1	-0.0000033715
a2	-21268.8183841396
b2	750.6883257716
c2	-8.8310236888
d2	0.0346293473
a3	109906.7696641756
b3	-3391.6987003969
c3	34.7730506336
d3	-0.1183674060

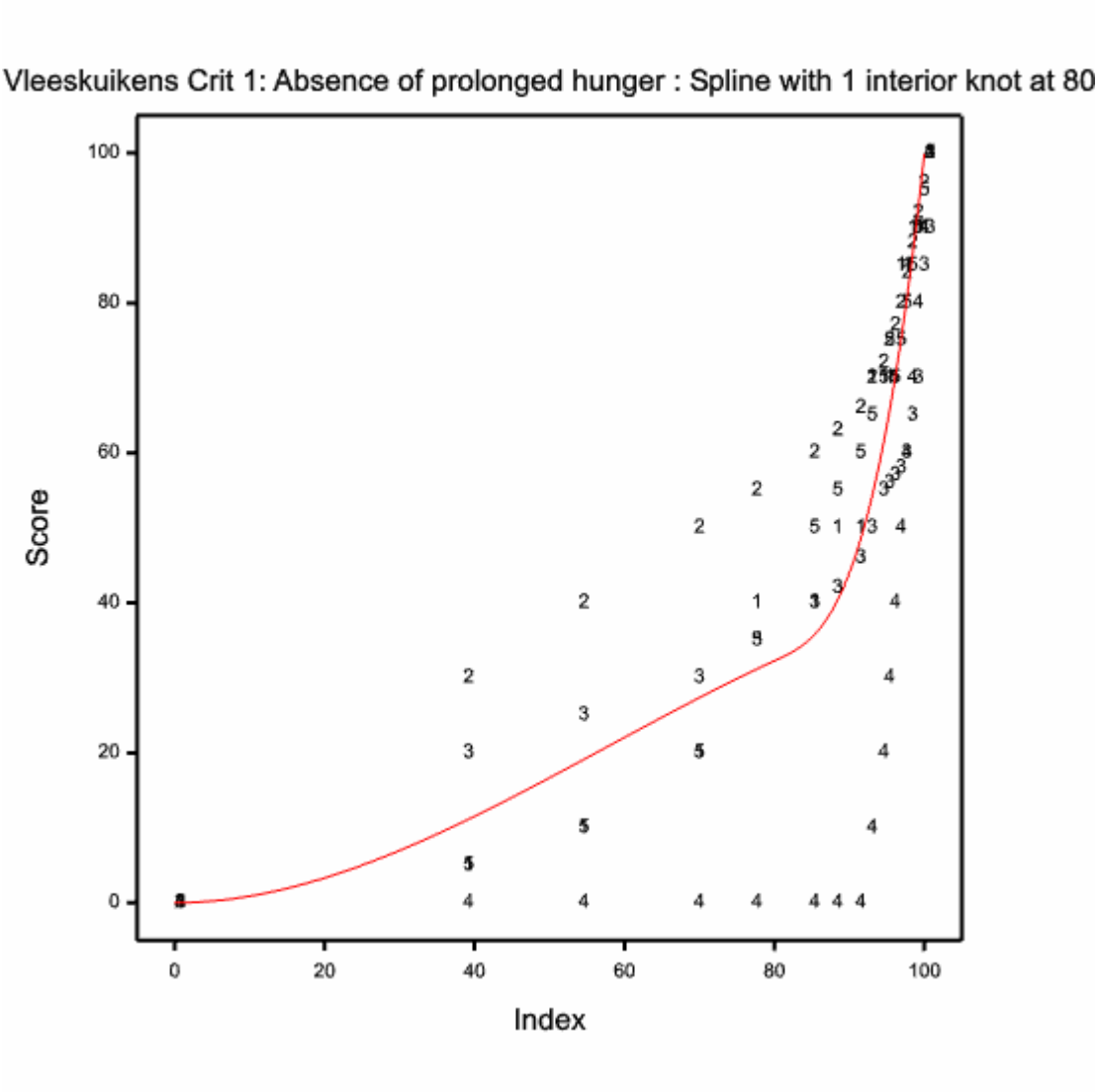
Index = I

When I≤85 then Score = $(0.024147 \times I) + (0.0003195893 \times I^2) - (0.0000033715 \times I^3)$
When I≥85 and I≤95 then Score = $-21268.82 + (750.6883 \times I) - (8.83102368 \times I^2) + (0.0346293473 \times I^3)$
When I≥95 then Score = $109906.77 - (3391.6987 \times I) + (34.77305 \times I^2) - (0.1183674060 \times I^3)$

Choquet integral:

Rejections: $\mu_a = 0.12$
Mortality: $\mu_b = 0.71$
Sum of capacities = 0.83

Appendix 5. New spline function for criterion 1: absence of hunger



Coëfficiënten:

	COEF	VALUE
a1		0.0000000000
b1		0.0000000000
c1		0.0093729150
d1		-0.0000541267
a2		-3865.4448840019
b2		144.9541766846
c2		-1.8025542202
d2		0.0074955694

Index = I

When I≤80 then Score = $0.0093729 \times I^2 - 0.0000541267 \times I^3$

When I≥80 then Score = $-3865.445 + (144.95418 \times I) - (1.8025542 \times I^2) + (0.0074955694 \times I^3)$

Appendix 6. New choquet integral for criterion 6: absence of injuries

CRITERION 6

Choquet integral:

	est	
m1	0.0599	Breast burns
m2	0.0557	Hock burns
m3	0.0057	Pododermatitis
m4	0.1682	Lameness
m12	0.1333	
m13	0.1667	
m14	0.2667	
m23	0.1667	
m24	0.2667	
m34	0.2833	
m123	0.1595	
m124	0.5387	
m134	0.6220	
m234	0.6679	

Appendix 7. Results of analysis of simplification methods, using dataset 2 (breast blister measures replaced by measures of hock burn at the slaughter plant).

For full explanation of the tables and terms, we refer to the methods section of the report and the results section on dataset 1.

Table 7.1. Number of flocks in each category (dataset 2), according to the full assessment protocol (Welfare Quality®, 2009) for Golden Standard 1 (GS1) and Golden Standard 2 (GS2) (GS1: criterion8=100; GS2: criterion 8=min(criterion6,criterion7). NA=not scored due to missing principle score.

	GS1	**GS2**
Excellent	0	0
Enhanced	7	7
Acceptable	104	94
Not classified	17	27
NA	52	52
Margin	180	180

Table 7.2. Comparison of a simplified model (strategy 1, prediction of gait score from hock burn on-farm) with the full model for GS1. For explanation of terms, see 2.8.3.2. The table shows the distribution of flocks over the categories, with a summary of the measures at 90% confidence intervals. Est=estimated agreement; low=lower limit of confidence interval; upp=upper limit of confidence interval.

Strategy 1 → Full model ↓	Excell	Enhanc	Accept	Not cl	NA	Margin		90% Conf. Interval		
Excellent	0	0	0	0	0	0		est	low	upp
Enhanced	0	7	0	0	0	7	%equal	100.0	97.2	100.0
Acceptable	0	0	104	0	0	104	%sp	100.0	59.0	100.0
Not classified	0	0	0	17	0	17	%se	100.0	97.0	100.0
NA	0	0	0	0	52	52	%fn	0.0	0.0	41.0
Margin	0	7	104	17	52	180	%fp	0.0	0.0	3.0

Table 7.3. Comparison of a simplified model (strategy 1, prediction of gait score from hock burn on-farm) with the full model for GS2. For explanation of terms, see 2.8.3.2. The table shows the distribution of flocks over the categories, with a summary of the measures at 90% confidence intervals. Est=estimated agreement; low=lower limit of confidence interval; upp=upper limit of confidence interval.

Strategy 1 → Full model ↓	Excell	Enhanc	Accept	Not cl	NA	Margin		90% Conf. Interval		
Excellent	0	0	0	0	0	0		est	low	upp
Enhanced	0	7	1	0	0	7	%equal	96.9	92.2	99.1
Acceptable	0	0	93	1	0	94	%sp	100.0	59.0	100.0
Not classified	0	0	3	24	0	27	%se	100.0	97.0	100.0
NA	0	0	0	0	52	52	%fn	0.0	0.0	41.0
Margin	0	7	96	25	52	180	%fp	0.0	0.0	3.0

Table 7.4. Comparison of a simplified model (strategy 2, prediction on-farm measures from slaughter plant measures) with the full model for GS1. For explanation of terms, see 2.8.3.2. The table shows the distribution of flocks over the categories, with a summary of the measures at 90% confidence intervals. Est=estimated agreement; low=lower limit of confidence interval; upp=upper limit of confidence interval.

Strategy 2 → Full model ↓	Excell	Enhanc	Accept	Not cl	NA	Margin			90% Conf. Interval	
Excellent	0	0	0	0	0	0		est	low	upp
Enhanced	0	5	2	0	0	7	%equal	96.8	92.1	99.1
Acceptable	0	1	102	1	0	104	%sp	71.4	29.0	96.3
Not classified	0	0	0	16	1	17	%se	99.2	95.4	100.0
NA	0	0	0	0	52	52	%fn	28.6	3.7	71.0
Margin	0	6	104	17	53	180	%fp	0.8	0.0	4.6

Table 7.5. Comparison of a simplified model (strategy 2, prediction on-farm measures from slaughter plant measures) with the full model for GS2. For explanation of terms, see 2.8.3.2. The table shows the distribution of flocks over the categories, with a summary of the measures at 90% confidence intervals. Est=estimated agreement; low=lower limit of confidence interval; upp=upper limit of confidence interval

Strategy 2 → Full model ↓	Excell	Enhanc	Accept	Not cl	NA	Margin			90% Conf. Interval	
Excellent	0	0	0	0	0	0		est	low	upp
Enhanced	0	5	2	0	0	7	%equal	92.1	86.0	96.2
Acceptable	0	1	90	3	0	94	%sp	71.4	29.0	96.3
Not classified	0	0	4	22	1	27	%se	99.2	95.4	100.0
NA	0	0	0	0	52	52	%fn	28.6	3.7	71.0
Margin	0	6	96	25	53	180	%fp	0.8	0.0	4.6

Appendix 8. Results of analysis of simplification methods at the level of individual principle scores, using dataset 2 (breast blister measures replaced by measures of hock burn at the slaughter plant).

For full explanation of the tables, figures and terms, we refer to the methods section of the report and the results section on dataset 1. The calculations only differ from dataset 1 for principle 3.

Table 8.1. Comparison of a simplified model (strategy 1, replacing gait score with hock burn scores on-farm) for GS1 (upper part of the table) and GS2 (lower part of the table). For explanation of terms, see 2.8.3.2. Empty cells indicate that there are no values for this category. Rsp: Spearman rank correlation coefficient. Est=estimated agreement; low=lower limit of confidence interval; upp=upper limit of confidence interval.

Principle 3 **GS1:C8=100**		20.0			55.0			80.0		
		90% Conf.interval			90% Conf.interval			90% Conf.interval		
		est.	lower	upper	est.	lower	upper	est.	lower	upper
Prediction of gait score from hock burn on-farm	%equal	98.5	95.3	99.7	100.0	97.7	100.0	100.0	97.7	100.0
	%se	33.3	1.7	86.5	100.0	97.7	100.0	100.0	97.7	100.0
Rsp=0.87	%sp	100.0	97.7	100.0						

Principle 3 **GS2:C8=min(C6,C7)**		20.0			55.0			80.0		
		90% Conf.interval			90% Conf.interval			90% Conf.interval		
		est.	lower	upper	est.	lower	upper	est.	lower	upper
Prediction of gait score from hock burn on-farm	%equal	95.4	91.2	98.0	100.0	97.7	100.0	100.0	97.7	100.0
	%se	72.2	50.2	88.4	100.0	97.7	100.0	100.0	97.7	100.0
Rsp=0.90	%sp	99.1	95.9	100.0						

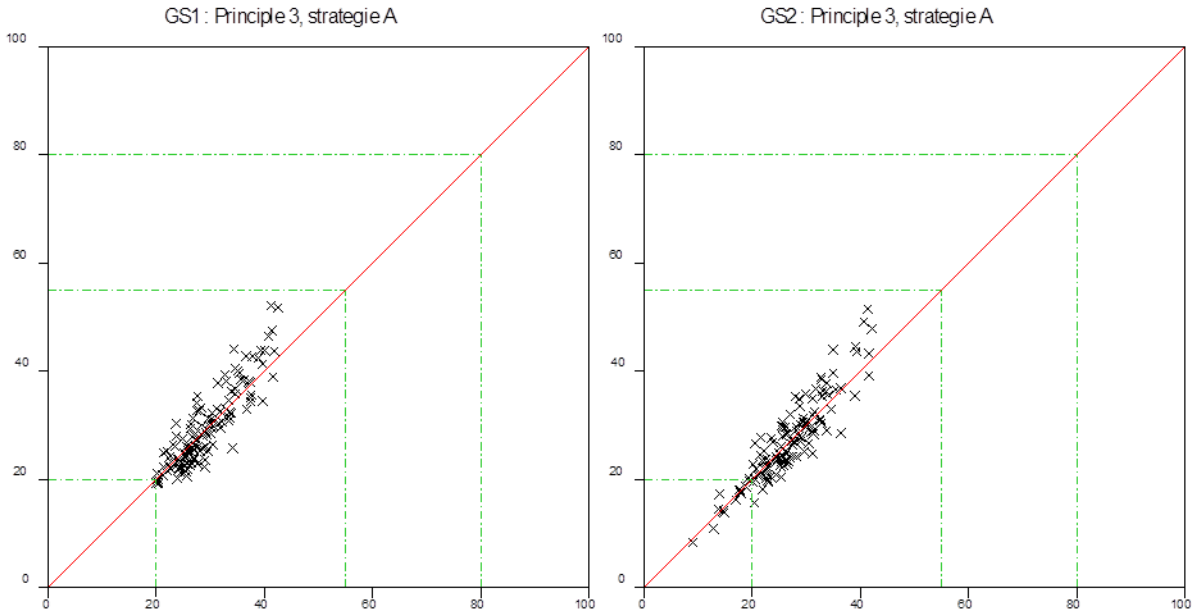


Figure 8.1. Score for principle 3 (good health) for the golden standard (GS) on the Y-axis against the simplified model (strategy 1, gait score replaced by hock burn on-farm) on the X-axis. The left figure shows the results for GS1, the right figure for GS2.

Table 8.2. Comparison of a simplified model (strategy 2, replacing on-farm measures with slaughter plant measures) for principle 3, GS1 (upper part of the table) and GS2 (middle part of the table), and principle 2. For explanation of terms, see 2.8.3.2. Empty cells indicate that there are no values for this category. Rsp: Spearman rank correlation coefficient. Est=estimated agreement; low=lower limit of confidence interval; upp=upper limit of confidence interval.

Principle 3 GS1:C8=100		20.0			55.0			80.0		
		90% Conf.interval			90% Conf.interval			90% Conf.interval		
		est.	lower	upper	est.	lower	upper	est.	lower	upper
Prediction of on-farm measures from slaughter plant measures	%equal	96.9	93.1	98.9	100.0	97.7	100.0	100.0	97.7	100.0
	%se	0.0	0.0	63.2	100.0	97.7	100.0	100.0	97.7	100.0
Rsp=0.75	%sp	100.0	97.7	100.0						

Principle 3 GS2:C8=min(C6,C7)		20.0			55.0			80.0		
		90% Conf.interval			90% Conf.interval			90% Conf.interval		
		est.	lower	upper	est.	lower	upper	est.	lower	upper
Prediction of on-farm measures from slaughter plant measures	%equal	92.3	87.3	95.8	100.0	97.7	100.0	100.0	97.7	100.0
	%se	66.7	44.6	84.4	100.0	97.7	100.0	100.0	97.7	100.0
Rsp=0.80	%sp	96.4	92.0	98.8						

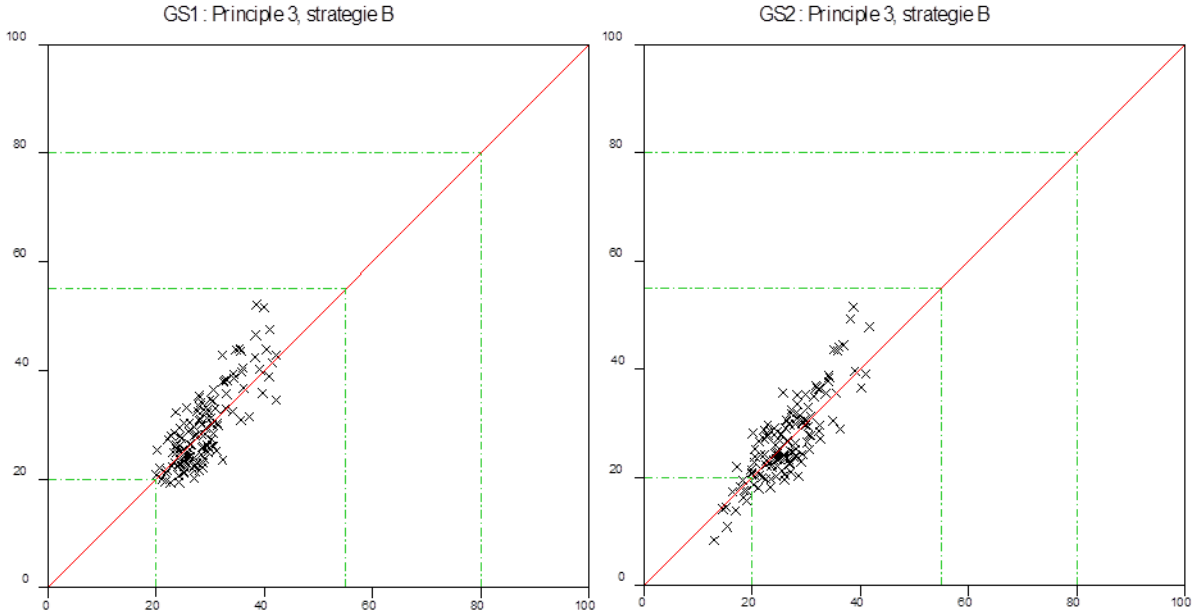


Figure 8.2. Score for principle 3 (good health) for the golden standard (GS) on the Y-axis against the simplified model (strategy 2, prediction of on-farm measures from slaughter plant measures) on the X-axis. The left figure shows the results for GS1, the right figure for GS2.

Appendix 9. Results of analysis of simplification methods at the level of individual criterion scores, using dataset 2 (breast blister measures replaced by measures of hock burn at the slaughter plant).

For full explanation of the tables, figures and terms, we refer to the methods section of the report and the results sections on dataset 1. The calculations only differ from dataset 1 for criterion 6.

Table 9.1. Comparison of a simplified model (strategy 1, predicting gait scores from hock burn on-farm) for criterion 6. For explanation of terms, see 2.8.3.2. Empty cells indicate that there are no values for this category. Rsp: Spearman rank correlation coefficient. Est=estimated agreement; low=lower limit of confidence interval; upp=upper limit of confidence interval.

Criterion 6.		20.0			55.0			80.0		
		90% Conf.. interval			90% Conf.. interval			90% Conf. interval		
		est.	lower	upper	est.	lower	upper	est.	lower	upper
	%equal	91.7	86.9	95.2	100.0	97.9	100.0	100.0	97.9	100.0
	%se	94.8	89.5	98.0	100.0	97.9	100.0	100.0	97.9	100.0
Rsp=0.88	%sp	85.4	74.4	93.0						

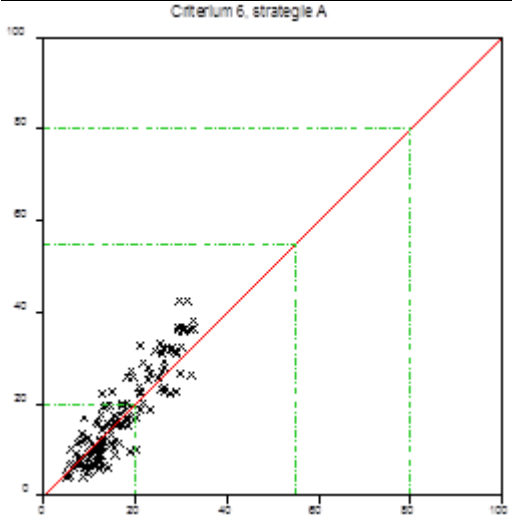


Figure 9.1. Score for criterion 6 for the golden standard (GS) on the Y-axis against the simplified model (strategy 1, predicting gait scores from hock burn on-farm) on the X-axis.

Table 9.2. Comparison of a simplified model (strategy 2, predicting on-farm measures from slaughter plant measures) for criterion 6. For explanation of terms, see 2.8.3.2. Empty cells indicate that there are no values for this category. Rsp: Spearman rank correlation coefficient. Est=estimated agreement; low=lower limit of confidence interval; upp=upper limit of confidence interval.

Criterion 6.		20.0			55.0			80.0		
		90% Conf. interval			90% Conf. interval			90% Conf. interval		
		est.	lower	upper	est.	lower	upper	est.	lower	upper
	%equal	92.4	87.7	95.7	100.0	97.9	100.0	100.0	97.9	100.0
	%se	97.9	93.6	99.6	100.0	97.9	100.0	100.0	97.9	100.0
Rsp=0.79	%sp	80.8	69.0	89.6						

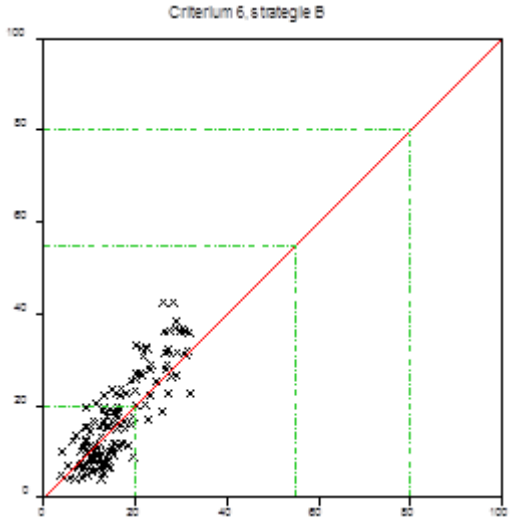


Figure 9.2. Score for criterion 6 for the golden standard (GS) on the Y-axis against the simplified model (strategy 2, predicting on-farm scores from slaughter plant scores) on the X-axis.

Wageningen UR Livestock Research

Edelhertweg 15, 8219 PH Lelystad T 0320 238238 F 0320 238050

E info.livestockresearch@wur.nl I www.livestockresearch.wur.nl