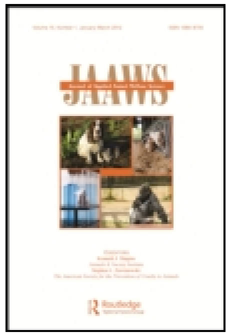

Ordinal or visual analogue scales for assessing aspects of broiler chicken welfare?

Journal:	*Journal of Applied Animal Welfare Science*
Manuscript ID	HAAW-2022-0032.R2
Manuscript Type:	Research Article
Keywords:	animal welfare, animal-based measures, categorical scale, continuous scale, poultry

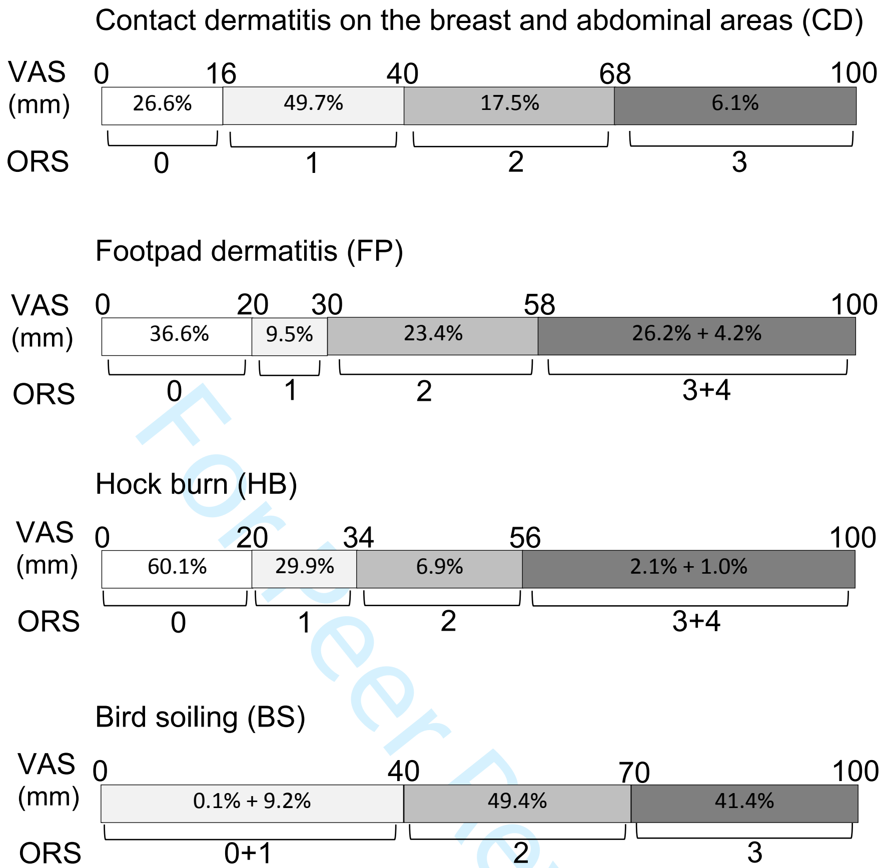
Contact dermatitis on the breast and abdominal areas (CD)
VAS (mm)
0
16
40
68
100
26.6%
49.7%
17.5%
6.1%
ORS
0
1
2
3
Footpad dermatitis (FP)
VAS (mm)
0
20
30
58
100
36.6%
9.5%
23.4%
26.2% + 4.2%
ORS
0
1
2
3+4
Hock burn (HB)
VAS (mm)
0
20
34
56
100
60.1%
29.9%
6.9%
2.1% + 1.0%
ORS
0
1
2
3+4
Bird soiling (BS)
VAS (mm)
0
40
70
100
0.1% + 9.2%
49.4%
41.4%
ORS
0+1
2
3

Table 1. Estimates of inter-rater reliability and confidence interval, 5,000 bootstrap samples, for animal welfare indicators from 1,303 broiler chickens assessed on farm by three raters using both ordinal scale (ORS) and visual analogue scale (VAS).

Welfare indicator	Scale	Intraclass correlation	Confidence interval (95%)	P-value*
Contact dermatitis on the breast and abdominal areas	ORS	0.68	(0.58 - 0.77)	<0.001
	VAS	0.77	(0.67 - 0.85)	
Footpad dermatitis	ORS	0.91	(0.87 - 0.93)	<0.001
	VAS	0.88	(0.83 - 0.92)	
Hock burns	ORS	0.67	(0.55 - 0.76)	<0.001
	VAS	0.72	(0.60 - 0.80)	
Bird soiling	ORS	0.61	(0.46 - 0.73)	0.447
	VAS	0.54	(0.36 - 0.69)	

Table 2a. Correlation of ordinal scale (ORS) and visual analogue scale (VAS) for the mean of values given by the three raters and for the individual values of each rater.

Table 2b. Correlation of broiler chicken welfare indicators measured on farm using ORS and VAS, 1,303 birds.

Table 2a

Indicator	Spearman rank correlation between ORS and VAS*	
	Mean	Individual
Contact dermatitis on the breast and abdominal areas (CD)	0.96	0.89
Footpad dermatitis (FP)	0.97	0.95
Hock burn (HB)	0.90	0.77
Bird soiling (BS)	0.94	0.81

*P < 0.0001

Table 2b

Correlation between indicators* (ORS, Spearman correlation; VAS, Pearson correlation)				
Indicator	Scale	FP	HB	BS
CD	ORS	0.06	0.24	0.34
	VAS	0.09	0.35	0.34
FP	ORS		0.17	0.08
	VAS		0.26	0.12
HB	ORS			0.25
	VAS			0.24

Ordinal or visual analogue scales for assessing aspects of broiler chicken welfare?

Abstract

Information may be lost when the gradation of animal welfare is scored through ordinal scales. Therefore, some advocate the use of continuous scales, which may be tagged with internal anchors. Equidistant tags are used; however, studies have demonstrated that empirical data for the space between tags tend to be non-equidistant. Ordinal rate scales (ORS) and visual analogue scales (VAS) were tested for the assessment of contact dermatitis on the breast and abdominal areas (CD), footpad dermatitis (FP), hock burns (HB) and bird soiling (BS) in broiler chickens. Calculations regarding the inter-rater reliability, the correlation between VAS and ORS and amongst the welfare indicators measured with both scales, as well as the equidistance of ORS categories in relation to values measured using VAS, were made. A total of 1,303 broiler chickens from 10 flocks was assessed on-farm by three trained raters using an ORS and a VAS anchored only with the minimum and the maximum scores at each end. Inter-rater reliabilities of CD (0.68 vs 0.77, $P<0.001$) and HB (0.67 vs 0.72, $P<0.001$) were higher when using VAS compared with ORS, but that of FP (0.91 vs 0.88, $P<0.001$) was lower. Correlations between ORS and VAS varied between 0.90-0.97 and 0.77-0.95 ($P<0.001$) respectively, considering mean and individual values of the three raters. Low to moderate correlations were observed between the four indicators using ORS and VAS. Tags on VAS that best represented ORS were non-equidistant. Results suggest both scales were reliable to assess the selected broiler chicken welfare indicators.

Keywords: animal welfare, animal-based measures, categorical scale, continuous scale, poultry

1. **Introduction**

The development and application of protocols to assess animal welfare (AW) has increased worldwide. In addition to registering absence and presence of AW issues, it is often useful and informative to score gradations of these issues. Assuming equal reliability, the more refined these gradations are scored, the more sensitive becomes the detection of relevant AW aspects, such as AW progress over time, differences between the welfare of groups of animals or effects interventions have on the lives of animals. Scientific research has encouraged the development of new techniques to assess AW in field conditions. Reliability between raters is an important criterion in the selection of AW indicators, since there is high probability of single person assessments due to manpower costs of animal-based monitoring schemes (Tuyttens et al., 2014). There are some initiatives for assessing the welfare of broiler chickens, like the Welfare Quality® (2009), the AssureWel (2014) and the Global Animal Partnership® (2018). These protocols include measures, predominantly presented as ordinal rating scales (ORS) ranging from 2- to 6-point scales. Raters can be trained to score reliably using ORS, and much of advances in knowledge of broiler chicken welfare are due to the application of ORS in the assessment of welfare in experimental and commercial flocks.

Descriptors, photos and videos may be used for illustrating, and practicing the recognition of stepwise increases in severity, thereby increasing consistency within and between observers. This also implies that data from different studies can be compared if the same ORS are used. However, assessing continuous welfare traits by using discontinuous scales may be disadvantageous (Tuyttens et al., 2009). The use of ORS may result in reduced sensitivity when raters are able to discriminate more levels of the assessed indicator than the number of categories allow for and are forced to group gradations they perceive as different into the same category.

A different type of scale, the visual analogue scale (VAS), is largely used to assess pain in humans and non-human animals (de Grauw and van Loon, 2016; Hjermstad et al., 2011). In AW assessment, VAS has also been applied to assess qualitative behavior defined by Wemelsfelder et al. (2001) as one of the "whole animal" measures that aim to assess the overall subjective experience or mood of an animal (Fleming et al., 2016; Grosso et al., 2016; Minero et al., 2016), and lameness (Flower and Weary, 2006; Nalon et al., 2014; Tuyttens et al., 2009; Vieira et al., 2015) in different species. VAS is a continuous scoring system that consists of a line, which varies usually from 100 to 125 mm in length, anchored by the minimum and the maximum score at each end. Thus, VAS removes the constraint of grouping information into discrete units and enables raters to achieve greater sensitivity in their scoring for aspects that vary along a continuum. In general, continuous variables present more statistical power as compared to ordinal or categorical data, and this is likely the case with VAS as compared to ORS. The downside of the conventional VAS is the difficulty to train raters to score different gradations consistently, and as observed by de Grauw and van Loon (2016), the inter-rater reliability may be low. In this case, the tagged VAS (tVAS), which is a VAS with internal anchors, has been investigated as a tool to combine the advantages of both ORS and VAS (Nalon et al., 2014; Tuyttens et al., 2009). The tags add information to guide raters through different gradations thereby increasing reliability and facilitating the training of raters (Tuyttens et al., 2009).

Previous studies assumed equidistant tags to VAS to assess specific indicators of animal welfare based on existing categories used in ORS (Meeremans et al., 2017; Nalon et al., 2014; Rufener et al., 2018; Tuyttens et al., 2009). However, Vieira et al. (2015) challenged this rationale by presenting a non-equidistant characteristic of tags in VAS as with lameness in dairy goats. In this case, tags that are based on existing categories from ORS are expected to be checked with respect to what their correct positions are on the VAS and whether these are spaced equidistantly or not. As with lameness, many other relevant welfare problems vary continuously and could be assessed by a continuous scale rather than an ORS. For broiler chickens, contact dermatitis and related measures are considered important animal welfare indicators. They have been systematically scored using ORS in a variety of scoring scales: contact dermatitis (Allain et al., 2009; de Jong et al., 2014; Ekstrand et al., 1998; Haslam et al., 2007; Martland, 1985; Souza et al., 2018; Welfare Quality®, 2009) and bird soiling (Dawkins et al., 2004; Elwinger, 1995; Weeks et al., 1994; Welfare Quality®, 2009; Wilkins et al., 2003), for example. Potential improvement in the use of VAS to assess these indicators seems to warrant further studies, especially testing for reliability.

Recent studies have compared ORS and VAS, including tVAS, in animal welfare assessment. For example, Vogt et al. (2017) considered VAS reliable to assess the temperament of animals, and both VAS and ORS were considered reliable scales to assess lameness in dairy cattle (Flower and Weary, 2006). Considering the use of tags in VAS, tVAS and 5-point ORS presented similarly high interobserver reliability for the assessment of lameness in sows, but both were better than for 2-point ORS (Nalon et al., 2014). However interobserver reliability was better for the tVAS than for the ORS (Tuyttens et al., 2009) when assessing lameness in dairy cattle. In contrast, Meeremans et al. (2017) observed that use of tVAS did not improve the reliability of the assessment of fish vitality as compared to categorical scoring.

Regarding the decision on the best type of scale, the determinant seems to rely on how observers are able to discriminate between the levels of the indicator (Engel et al., 2003). Based on this, we aimed to test the application of ORS and VAS for four broiler chicken welfare indicators. The indicators were contact dermatitis on the breast and abdominal areas (CD), footpad dermatitis (FP), hock burns (HB) and bird soiling (BS). We studied inter-rater reliability, the correlation between the VAS and ORS and amongst the welfare indicators measured with VAS and ORS. We also tested the equidistance of ORS categories in relation to values measured using the VAS.

2. Material and Methods

2.1 Ethical statement

This project was approved by the Animal Use Ethics Committee of the Agricultural Campus (n. 079/2015; November 12th, 2015) of the Federal University of Paraná.

2.2 Animals, housing and data collection

A total of 1,303 broiler chickens, randomly selected from 10 flocks, was assessed in the State of Paraná, Southern Brazil, from January 9th to 13th 2017. The sampling size of 1,300 birds was calculated considering a maximum error of 5% and 95% confidence interval. The sample was not selected to be representative of bird welfare in Brazilian industrial broiler chicken units. The poultry barns had sidewalls with wire mesh covered by blackout curtains working as dark house (n = 1) or covered by yellow curtains, with natural lighting (n = 9). The farms were selected as a convenience sample according to our objective, which was to test the ordinal and analogue scales. All units had automatic feeders, nipple drinkers, sprinklers, exhaust fans and wood shaving litter, and nine units maintained evaporative cooling systems. Indoor mean temperature in the units at time of the visit was 27.7 ± 1.4 °C. Average broiler house area was 1,540 ± 187 m^2 and the number of birds per house was 18,904 ± 2,604, with a stocking density of 36.4 ± 0.9 kg/m^2. Birds were male and female Cobb 500®, assessed at 41.3 ± 2.0 days of age. The raters were one animal scientist and two veterinarians, one of them experienced in auditing poultry farms. The non-experienced raters underwent a 4 h classroom instruction about the indicators via picture observation, followed by a 4 h training session at the Federal University of Paraná farm. Scales used on the training sessions were obtained from Souza et al. (2018) and Welfare Quality® (2009). One month after the training, the non-experienced raters were asked to score 13 pictures for FP and 15 pictures for CD and BS to check concordance among them and solve any doubts before the experiment. Kendall's coefficient of concordance corrected for ties among raters were 0.89 (P=0.002), 0.79 (P=0.004) and 0.93 (P=0.001) for FP, CD and BS, respectively, and were considered adequate (Landis and Koch, 1977).

Raters scored each bird simultaneously but independently. They performed a visual inspection of a total 130 birds from five locations in each poultry house. The feet of the birds were cleaned by gently rubbing with the tip of observer's fingers. All assessors scored each bird simultaneously, so that any lesions were seen by all immediately after the cleaning procedure. Following the regular Welfare Quality procedures, birds were not individually identified. As for CD and BS, the original ORS by Souza et al. (2018) were applied, which included a colour picture and a description of each level of the scale. For FP and HB, the scales by the Welfare Quality® (2009) were used, including a colour picture representative of each level of the scale (Fig. 1). To collect data, a questionnaire was developed at the QuickTapSurvey® website to be used as a mobile phone application. Raters scored each bird using both the ORS and the VAS for each indicator. The application presented the ORS followed by VAS, thus the raters usually scored ORS first.. In the ORS, the raters had to select a score on a 4- or 5-point scale. The VAS consisted of a line initially designed with 10 cm, and proportional to this length depending on the screen size, and anchored only with the minimum and the maximum score at each end (absence or severe CD, FP, HB and BS). The raters could move a marker along the line to register the level of severity observed in the bird for each indicator. Data from QuickTapSurvey® were downloaded into an Excel file and checked for errors before use.

[Insert Fig. 1]

Fig. 1. Ordinal scales for the assessment of four broiler chicken welfare indicators; [1] (Souza et al., 2018), [2] (Welfare Quality®, 2009).

2.3 Statistical analysis

Linear mixed models were fitted to evaluate the inter-rater reliability. The intraclass correlation coefficient (ICC) was estimated as a measure of inter-rater reliability, and bootstrap confidence intervals were derived based on 5,000 simulations. The total data variability (TDV) was decomposed into variability attributed or not attributed to the raters (VNA). ICC values were calculated based on the VNA:TDV ratio, adjusted for the variability between poultry farms.Poultry farm was included as a factor in the model, as measurements in individual animals within the same farm tended to be related to each other. Thus, for any animal welfare indicator, indicated as Y, the following linear mixed model was defined:

$$Y_{ijkl} = n + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma_k + \epsilon_{ijkl}$$, where

Y_{ijkl} is the *l-th* assessment of the rater *j* in the animal *k* of the poultry farm *i;*

$\alpha_i \sim Normal(0, \sigma_\alpha^2)$ is the random effect of poultry farm;

$\beta_j \sim Normal(0, \sigma_\beta^2)$ is the random effect of rater;

$\gamma_k \sim Normal(0, \sigma_\gamma^2)$ is the random effect of animal;

$(\alpha\beta)_{ij} \sim Normal(0, \sigma_{\alpha\beta}^2)$ is the random effect of the interaction between rater and poultry house;

η is the model intercept;

$\epsilon_{ijkl} \sim Normal(0, \sigma^2)$ is the random error.

Based on this, the ICC was calculated as:

$$ICC = \frac{\sigma_\alpha^2 + \sigma_\gamma^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_{\alpha\beta}^2 + \sigma_\gamma^2 + \sigma^2}$$

ICC values were estimated under both scales (ORS and VAS), and the difference $ICC_{ORS} - ICC_{VAS}$ was calculated. To evaluate the statistical significance of this difference, the null hypothesis of equality was tested through additional simulations, and the simulated p-values are presented.

Spearman's rank correlation coefficients for the mean of values given by the three raters and for the individual values of each rater was used to test correlations between ORS and VAS for all indicators, as well as correlations amongst all indicators measured using the ORS. Pearson Correlation Coefficient was used to test correlations amongst all indicators measured using the VAS. Correlations from 0.3 to 0.6 were considered moderate, and values above 0.6 were considered high (de Jong et al., 2015).

Linear mixed models were also fitted to test the assumption of equidistance of ORS categories according to values measured using the VAS. For each indicator, the VAS values were considered as the response variable, and the ORS values as the predictor (independent variables). Random effects of animal, rater, poultry house and interaction between rater and poultry house were also included in the models. Two linear mixed models were fitted for each indicator, assuming (Model 1) or not assuming equidistance (Model 2) between the scores. In Model 1, ORS was included as a numerical variable defined by the p+1 different values. In Model 2, ORS was included as a categorical variable, not assuming a fixed increment across scores. So, the following models were considered:

Model 1: $Y_{ijkl} = n + NRS_{ijkl} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma_k + \epsilon_{ijkl}$, where Y_{ijkl} and NRS_{ijkl} correspond to the rater *j* in the animal *k* of the poultry farm *i* for the *l-th* time in the scales VAS and ORS, respectively;

Model 2:

$$Y_{ijkl} = n + \tau_1 \times I(NRS_{ijkl} = 1) + \tau_2 \times I(NRS_{ijkl} = 2) + \tau_3 \times I(NRS_{ijkl} = 3) + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma_k + \epsilon_{ijkl}$$

, where $I(NRS_{ijkl} = x)$ is the indicator function, assuming value zero when ORS score is different of an x value, and assuming value one when ORS score is equal to x; τ_1, τ_2 and τ_3 are the effects that reflect the association between ORS and VAS.

To evaluate the equidistance hypothesis, the fitted linear mixed models were compared using the Akaike Information Criterion (AIC), following the method of Burnham and Anderson (2002), based on the evidence ratio, defined by:

$$ER = \frac{1}{\exp\left(-0.5(AIC_{mod1} - AIC_{mod2})\right)}$$

ER will be equal to 1 if the evidence for both models is the same. The greater ER, the greater the evidence for model 2 (non-equidistant), whereas ER moves toward zero if model 1 (equidistance) has the highest evidence.

For each indicator for which the ER analysis confirmed non-equidistance, the best values for the non-equidistant tags of VAS were obtained through classification tree analysis, with VAS as the predictor and ORS the response. The classification tree method proposed by Breiman et al. (1984) employs successive partitions of a sample to constitute subsamples that are homogeneous in relation to response values, in our case ORS. The rules for the partitions were as $VAS < x$ versus $VAS \geq x$, being a VAS value, so that the observations were allocated to different subsamples (nodes) according to the partition rules. The final number of nodes was defined based on a cross validated procedure. In addition, the number of tags for the ordinal scale was also considered to determine the number of final nodes in a proper way. Analyses were performed using R Statistical Computing Environment software version 3.3.1 (R Core Team, 2016), through the packages boot (Angelo Canty and Brian Ripley, 2016), lme4 (Bates et al., 2015) and rpart (Therneau et al., 2015).

3. **Results**

Estimated inter-rater reliability was higher for CD and HB using VAS, and higher for FP using ORS (Table 1).

Table 1. Estimates of inter-rater reliability and confidence interval, 5,000 bootstrap samples, for animal welfare indicators from 1,303 broiler chickens assessed on farm by three raters using both ordinal scale (ORS) and visual analogue scale (VAS).

Welfare indicator	Scale	Intraclass correlation	Confidence interval (95%)	P-value*
Contact dermatitis on the breast and abdominal areas	ORS	0.68	(0.58 - 0.77)	<0.001
	VAS	0.77	(0.67 - 0.85)	
Footpad dermatitis	ORS	0.91	(0.87 - 0.93)	<0.001
	VAS	0.88	(0.83 - 0.92)	
Hock burns	ORS	0.67	(0.55 - 0.76)	<0.001
	VAS	0.72	(0.60 - 0.80)	
Bird soiling	ORS	0.61	(0.46 - 0.73)	0.447
	VAS	0.54	(0.36 - 0.69)	

*ORS x VAS intraclass correlation

High correlations were observed between ORS and VAS for each welfare indicator, considering mean and individual values (Table 2). When indicators were correlated amongst them, within each scale, we observed similar level of correlation of data using ORS and VAS (Table 2).

Table 2a. Correlation of ordinal scale (ORS) and visual analogue scale (VAS) for the mean of values given by the three raters and for the individual values of each rater.

Table 2b. Correlation of broiler chicken welfare indicators measured on farm using ORS and VAS, 1,303 birds.

Table 2a			Table 2b				
Indicator	Spearman rank correlation between ORS and VAS*		Correlation between indicators* (ORS, Spearman correlation; VAS, Pearson correlation)				
	Mean	Individual	Indicator	Scale	FP	HB	BS

Contact dermatitis on the breast and abdominal areas (CD)	0.96	0.89
Footpad dermatitis (FP)	0.97	0.95
Hock burn (HB)	0.90	0.77
Bird soiling (BS)	0.94	0.81

CD	ORS	0.06	0.24	0.34
	VAS	0.09	0.35	0.34
FP	ORS		0.17	0.08
	VAS		0.26	0.12
HB	ORS			0.25
	VAS			0.24

*P < 0.0001

For all indicators, the strength of evidence for the Model 2, which does not assume equidistance between tags, was higher than 0.99. Thus, the tags on VAS that better represent ORS are not evenly spaced. The calculated tags for each indicator are shown in Fig. 2. The prevalence of absence of soiling (score 0) among the broiler chickens assessed in our study was 0.1 %, while severe HB and FP (score 4) was observed in 1.0% and 4.2%, respectively. Since these frequencies did not allow an adequate tag calculation, scores 0 and 1 were aggregated for BS, as well as scores 3 and 4 for HB and FP (Fig. 2).

[Insert Fig. 2}

Fig. 2. Tags for ordinal scale (ORS) for broiler chicken welfare indicators calculated by the classification tree considering visual analogue scale (VAS) as predictor. Percentages refer to the number of birds classified in each ORS category, data from 1,303 birds assessed on farm by three raters.

4. **Discussion**

Higher ICC for CD and HB using VAS, and for FP using ORS were observed; however, common ICC interpretation suggest that both scales were reliable to assess the animal-based indicators proposed in this study. This warrants further research comparing a greater number of raters. Direct comparison across studies using ORS and VAS is not possible due to different methods employed to estimate reliabilities. For those studies in which the reliability was given by a value between 0 and 1, the range of reported values considered to be reliable was similar to the range observed in our study (Flower and Weary, 2006; Meeremans et al., 2017; Nalon et al., 2014). As a general guideline, ICC reliability as measured with ICC is considered good when between 0.60 and 0.74, and excellent when higher than 0.75 (Cicchetti, 1994). In the case of BS, lack of difference between ORS and VAS seems related to high data variability. FP is observed as a clearer indicator, perhaps as consequence of a simpler scale. In the case of CD and BS, pictures needed to expose other animal parts, like skin, foot, and feathers, which may induce raters to reflect more about animal condition. In this case, data obtained may be influenced by something else, like experience or personal views (Meagher, 2009).

Other factors may have affected inter-rater reliability, such as place of assessment, training, quality of the descriptive textual and photographic material to support the assessment, and the limited number of raters. Studies comparing ORS and VAS for animal welfare purposes frequently combine video recordings and a large group of raters (e.g. Tuyttens et al., 2009; Nalon et al., 2014). In our study, on-farm assessments may have improved inter-rater reliability, even with three raters, since they could have chosen the best angle and touched the birds during the physical assessment. Touching the birds was important to remove dirt to confirm the presence and size of FP and HB. Since only one rater was experienced in broiler chicken welfare assessment, training, rather than experience, may have played an important role in helping raters to discriminate between the levels of each indicator (Meeremans et al., 2017). In addition, successful learning depends on a scoring system with clear definitions and photographs (Gibbons et al., 2012). In our case, training was done with the available scientifically validated scales to score the four proposed indicators. These materials were related to the use of ORS, which means that raters were trained to recognize four or five different levels of severity, depending on the indicator. Nevertheless, raters were able to coherently score birds using the VAS. The quality of the scoring system is important to provide all information required by the raters before and during the assessment, and clear definitions are essential to make scoring systems less dependent on personal experience or any factor that reduces inter-rater reliability (Meagher, 2009). In this regard, it is expected that more comprehensive training material, with pictures of various gradations in severity along the VAS, will increase inter-rater reliability.

Indicators showed the same level of correlation between them, regardless of the type of scale. The exception was the correlation between CD and HB, which was slightly higher when using the VAS compared with the ORS. Both CD and HB had higher inter-rater reliability using VAS, thus probably there was a refinement of the scoring using VAS, which impacted on the correlation between CD and HB. We expected higher correlation between CD, FP and HB, since contact dermatitis has been reported as to develop sequentially on different part of the body, starting with feet and followed by hocks and breast, as bird activity decreases (de Jong et al., 2014; Greene et al., 1985). Other factors, such as early age of modern fast-growing broiler chickens at slaughter and litter quality, may challenge the correlation between different types of skin lesion (Souza et al., 2018). Despite low to moderate correlation between indicators, the type of scale did not affect data interpretation for the selected outcomes in this study, suggesting both scales could be used to assess birds.

High correlation between ORS and VAS for all indicators may suggest applicability of both scales and is in line with results of comparisons between ORS and VAS for pain assessment (Hjermstad et al., 2011). Similar to Flower and Weary (2006), raters were able to coherently transpose ordinal scores into continuous scores even in the absence of internal tags on VAS. One possible limitation of this study was the application of both scales concomitantly, which may have motivated raters to virtually divide the VAS according to the ORS. Equidistant data would support this rationale, as observed by Engel et al. (2003).

However, data obtained in our study were not equidistant. The lack of equidistance has been observed in other studies using VAS to assess lameness (Thomsen et al., 2008; Vieira et al., 2015; Welsh et al., 1993) and in a study to determine cut-off points in a VAS for pain in patients with chronic musculoskeletal pain (Boonstra et al., 2014). Our results show that the decision regarding the location of tags had direct implication on the number of animals classified in each level of severity. As example, some birds who were scored as 0 using ORS, meaning absence of CD, FP and HB, received grades up to 16 or 20 mm using the VAS. These results are probably indicating that birds had less severe lesions than the ones described on level 1 of the ORS, and the rater had to choose between 0 and 1. In this case, the VAS was more sensitive to allow the rater to choose the best position between 0 and 1. In this example, the number of birds considered clinically absent of CD, FP and HB differed between ORS and VAS. If the different values of a specific welfare indicator are biased to one extreme, either more concentrated on the higher severity or the lower severity end, scales presenting more detailed assessment, as VAS, may offer higher accuracy.

According to Averbuch and Katzper (2004) and Nalon et al. (2014) inserting internal tags on VAS allows combining characteristics of the ORS, and improves uniformity of interpretation, with the flexibility of VAS to identify small changes between the tags. Although the VAS had a high reliability in this study, it is expected that the internal anchors of a tVAS will enable raters to score even more reliably. The position of the internal tags in a tVAS is important because it affects the number of animals in each level. As observed in Fig. 2, categories 0 and 1 were often narrower than the more severe categories. Perhaps the ORS over-emphasizes the milder cases, which were the most common for three indicators, while the VAS allows raters to better differentiate between the scores. Thus, to compare ORS and tVAS, it is important to have clear definitions about the position of different ORS categories along the continuous scale, and raters should be clearly instructed and trained on how to use the scale. This issue deserves more attention and seems especially relevant depending on the goal of the assessment, which may be to provide best practice recommendations or may be associated with sanctions (Main and Mullan, 2012) or bonuses for certification processes.

Many studies have been done to encourage the adoption of regular broiler chicken welfare assessment worldwide. This permanent monitoring of welfare may include the use of correlations, such as of contact dermatitis on farm and at the slaughterhouse (de Jong et al., 2015), as well as the use of technology to automate assessment on farm or at the slaughterhouse (Sassi et al., 2016). FP has been accepted as an important welfare indicator for surveillance purposes (European Commission, 2017), and automation of this assessment seems a priority. For automated assessment through image analysis, the ORS are commonly used, and in the case of FP they seem adequate. When both VAS and ORS work well, the choice of the scale will include a critical analysis of the conditions related to their use (Hjermstad et al., 2011). Adoption of an animal welfare indicator by organizations will depend on reliability, validity, sensitivity and power, but also feasibility and efficiency. VAS, including tVAS, presents potential to be considered for different animal welfare strategies, in addition to animal welfare assessment. As example, it may be used to validate automated monitoring of indicators showing higher inter-rater reliability using VAS or, since VAS is more sensitive (Welsh et al., 1993). Application may include its use during inspections for certification processes and as part of a verification procedure in an animal welfare management system (Souza and Molento, 2018), in studies in which high sensitivity is needed; or tVAS may be used as a silver standard for automated monitoring tools, since it is more likely to detect small differences and changes along time. In addition, future work studying the biological validity of both VAS and ORS with appropriate standards, as for instance histological assessments to check dermatitis severity, seem warranted to further understand accuracy of the measurements.

5. Conclusion

This is the first study to compare ORS and VAS for the selected broiler chicken welfare indicators. Both ORS and VAS were considered reliable to assess the broiler chicken welfare indicators CD, FP, HB and BS, despite some differences in inter-rater reliability. Although higher inter-rater reliability may lead to refined correlation studies, the interpretation of correlation did not differ between VAS and ORS. VAS, including tVAS, presents potential to add sensitivity on animal welfare assessment, and is a tool to be further explored in validation and certification protocols, especially in studies in which high sensitivity is needed. In this case, considering that results from animal welfare assessment may have direct implications to the animals and other stakeholders, the use of tVAS will demand clear specification about the position of tags on the continuous scale as well as the training of raters.

Acknowledgements

References

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. https://doi.org/10.1080/00071660903110901

Angelo Canty, Brian Ripley, 2016. boot: Bootstrap R (S-Plus) Functions.

AssureWel, 2014. AssureWel - Advancing Animal Welfare Assurance [WWW Document]. Broilers. URL http://www.assurewel.org/broilers (accessed 9.4.18).

Averbuch, M., Katzper, M., 2004. Assessment of Visual Analog versus Categorical Scale for Measurement of Osteoarthritis Pain. J. Clin. Pharmacol. 44, 368–372. https://doi.org/10.1177/0091270004263995

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67, 1–48. https://doi.org/10.18637/jss.v067.i01

Boonstra, A.M., Preuper, H.R.S., Balk, G.A., Stewart, R.E., 2014. Cut-off points for mild, moderate, and severe pain on the visual analogue scale for pain in patients with chronic musculoskeletal pain. Pain 155, 2545–2550. https://doi.org/10.1016/j.pain.2014.09.014

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and regression trees. CRC Press, Boca Raton.

Burnham, K.P., Anderson, D.R., 2002. Model selection and multimodel inference: a practical information-theoretic approach, 2nd ed. Springer-Verlag, New York.

Cicchetti, D. V., 1994. Guidlines, Criteria, and Rules of Thumb for Evalauting Normed and Standardized Assessment Instruments in Psychology. Psychol. Assess. 6, 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Dawkins, M.S., Donnely, . A. E., Jones, T.A., 2004. Chicken welfare is influenced more by housing conditions than by stocking density. Nature 427, 342–343.

de Grauw, J.C., van Loon, J.P.A.M., 2016. Systematic pain assessment in horses. Vet. J. 209, 14–22. https://doi.org/10.1016/j.tvjl.2015.07.030

de Jong, I.C., Gunnink, H., van Harn, J., 2014. Wet litter not only induces footpad dermatitis but also reduces overall welfare, technical performance, and carcass yield in broiler chickens. J. Appl. Poult. Res. 23, 51–58. https://doi.org/10.3382/japr.2013-00803

de Jong, I.C., Hindle, V.A., Butterworth, A., Engel, B., Ferrari, P., Gunnink, H., Perez Moya, T., Tuyttens, F.A.M., van Reenen, C.G., 2015. Simplifying the Welfare Quality® assessment protocol for broiler chicken welfare. Animal 10, 117–27.

https://doi.org/10.1017/S1751731115001706

Ekstrand, C., Carpenter, T.E., Andersson, I., Algers, B., 1998. Prevalence and control of foot-pad dermatitis in broilers in Sweden. Br. Poult. Sci. 39, 318–24. https://doi.org/10.1080/00071669888845

Elwinger, K., 1995. Broiler production under varying population densities - A field study . Arch. fur Geflugelkd. 59, 209–215.

Engel, B., Bruin, G., Andre, G., Buist, W., 2003. Assessment of observer performance in a subjective scoring system: Visual classification of the gait of cows. J. Agric. Sci. 140, 317–333. https://doi.org/10.1017/S0021859603002983

European Commission, 2017. Study on the application of the broilers directive (DIR 2007/43/EC) and development of welfare indicators. Brussels. https://doi.org/10.1149/1.3477934

Fleming, P.A., Clarke, T., Wickham, S.L., Stockman, C.A., Barnes, A.L., Collins, T., Miller, D.W., 2016. The contribution of qualitative behavioural assessment to appraisal of livestock welfare. Anim. Prod. Sci. 56, 1569–1578. https://doi.org/10.1071/AN15101

Flower, F.C., Weary, D.M., 2006. Effect of Hoof Pathologies on Subjective Assessments of Dairy Cow Gait. J. Dairy Sci. 89, 139–146. https://doi.org/10.3168/jds.S0022-0302(06)72077-X

Gibbons, J., Vasseur, E., Rushen, J., De Passillé, A.M., 2012. A training programme to ensure high repeatability of injury scoring of dairy cows. Anim. Welf. 21, 379–388. https://doi.org/10.7120/09627286.21.3.379

Global Animal Partnership's, 2018. Animal Welfare Rating Standard For Chickens Raised for Meat v3.1 [WWW Document]. v3.1. URL https://globalanimalpartnership.org/wp-content/uploads/2018/04/GAP-Standard-for-Meat-Chickens-v3.1-20180403.pdf (accessed 9.4.18).

Greene, J.A., McCracken, R.M., Evans, R.T., 1985. A contact dermatitis of broilers -clinical and pathological findings. Avian Pathol. 14, 23–38. https://doi.org/10.1080/03079458508436205

Grosso, L., Battini, M., Wemelsfelder, F., Barbieri, S., Minero, M., Dalla Costa, E., Mattiello, S., 2016. On-farm Qualitative Behaviour Assessment of dairy goats in different housing conditions. Appl. Anim. Behav. Sci. 180, 51–57. https://doi.org/10.1016/j.applanim.2016.04.013

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.

Hjermstad, M.J., Fayers, P.M., Haugen, D.F., Caraceni, A., Hanks, G.W., Loge, J.H., Fainsinger, R., Aass, N., Kaasa, S., 2011. Studies comparing numerical rating scales, verbal rating scales, and visual analogue scales for assessment of pain intensity in adults: A systematic literature review. J. Pain Symptom Manage. 41, 1073–1093. https://doi.org/10.1016/j.jpainsymman.2010.08.016

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174. https://doi.org/10.2307/2529310

Main, D.C.J., Mullan, S., 2012. Economic, education, encouragement and enforcement influences within farm assurance schemes. Anim. Welf. 21, 107–111. https://doi.org/10.7120/096272812X13345905673881

Martland, M.F., 1985. Ulcerative dermatitis dm broiler chickens: the effects of wet litter. Avian Pathol. 14, 353–364. https://doi.org/10.1080/03079458508436237

Meagher, R.K., 2009. Observer ratings: Validity and value as a tool for animal welfare research. Appl. Anim. Behav. Sci. 119, 1–14. https://doi.org/10.1016/j.applanim.2009.02.026

Meeremans, P., Yochum, N., Kochzius, M., Ampe, B., Tuyttens, F.A.M., Uhlmann, S.S., 2017. Inter-rater reliability of categorical versus continuous scoring of fish vitality: Does it affect the utility of the reflex action mortality predictor (RAMP) approach? PLoS One 12, 1–22. https://doi.org/10.1371/journal.pone.0179092

Minero, M., Dalla, E., Dai, F., Anne, L., Murray, M., Canali, E., Wemelsfelder, F., 2016. Use of Qualitative Behaviour Assessment as an indicator of welfare in donkeys. Appl. Anim. Behav. Sci. 174, 147–153.

Nalon, E., Maes, D., Van Dongen, S., van Riet, M.M.J., Janssens, G.P.J., Millet, S., Tuyttens, F.A.M., 2014. Comparison of the inter- and intra-observer repeatability of three gait-scoring scales for sows. Animal 8, 650–659. https://doi.org/10.1017/S1751731113002462

R Core Team, 2016. A language and environment for statistical computing.

Rufener, C., Baur, S., Stratmann, A., Toscano, M.J., 2018. A Reliable Method to Assess Keel Bone Fractures in Laying Hens From Radiographs Using a Tagged Visual Analogue Scale. Front. Vet. Sci. 5, 1–8. https://doi.org/10.3389/fvets.2018.00124

Sassi, N. Ben, Averós, X., Estevez, I., 2016. Technology and poultry welfare. Animals 6, 1–21. https://doi.org/10.3390/ani6100062

Souza, A., Soriano, V., Schnaider, M., Rucinque, D., Molento, C., 2018. Development and refinement of three animal-based broiler chicken welfare indicators. Anim. Welf. 27, 263–274.

Souza, A.P.O., Molento, C.F.M., 2018. Proposal of a management system to develop an animal welfare strategy for the animal food chain. CAB Rev. 13, 1–11. https://doi.org/10.1079/PAVSNNR201813001

Therneau, T., Atkinson, B., Brian Ripley, 2015. rpart: Recursive Partitioning and Regression Trees.

Thomsen, P.T., Munksgaard, L., Tøgersen, F.A., 2008. Evaluation of a Lameness Scoring System for Dairy Cows. J. Dairy Sci. 91, 119–126. https://doi.org/10.3168/jds.2007-0496

Tuyttens, F.A.M., de Graaf, S., Heerkens, J.L.T., Jacobs, L., Nalon, E., Ott, S., Stadig, L., Van Laer, E., Ampe, B., 2014. Observer bias in animal behaviour research: Can we believe what we score, if we score what we believe? Anim. Behav. 90, 273–280. https://doi.org/10.1016/j.anbehav.2014.02.007

Tuyttens, F.A.M., Sprenger, M., Van Nuffel, A., Maertens, W., Van Dongen, S., 2009. Reliability of categorical versus continuous scoring of welfare indicators: Lameness in cows as a case study. Anim. Welf. 18, 399–405. https://doi.org/10.1016/j.applanim.2010.05.003

Vieira, A., Oliveira, M.D., Nunes, T., Stilwell, G., 2015. Making the case for developing alternative lameness scoring systems for dairy goats. Appl. Anim. Behav. Sci. 171, 94–100. https://doi.org/10.1016/j.applanim.2015.08.015

Vogt, A., Aditia, E.L., Schlechter, I., Schütze, S., Geburt, K., Gauly, M., König von Borstel, U., 2017. Inter- and intra-observer reliability of different methods for recording temperament in beef and dairy calves. Appl. Anim. Behav. Sci. 195, 15–23. https://doi.org/10.1016/j.applanim.2017.06.008

Weeks, C.A., Nicol, C.J., Sherwin, C.M., Kestin, S.C., 1994. Comparison of the behaviour of broiler chickens in indoor and free-range environments. Anim. Welf. 3, 179–192.

Welfare Quality®, 2009. Welfare Quality ® Assessment protocol for poultry (broilers, laying hens). Welfare Quality Consortium, Lelystad, The Netherlands, p. 116.

Welsh, E.M., Gettinby, G., Nolan, A.M., 1993. Comparison of a visual analogue scale and a numerical rating scale for assessment of lameness, using sheep as a model. Am. J. Vet. Res. 54, 976–983.

Wemelsfelder, F., Hunter, T. E., Mendl, M., Lawrence, A. B., 2001. Assessing the ‘whole animal’: A free choice profiling approach. Anim. Behav. 62, 209–220.

Wilkins, L.J., Brown, S.N., Phillips, A.J., Warriss, P.D., 2003. Cleanliness of broilers when they arrive at poultry processing plants. Vet. Rec. 701–703.