Animal (2014), **8:4**, pp 643–649 © The Animal Consortium 2014
doi:10.1017/S1751731113002450

animal

Statistical tools to improve assessing agreement between several observers

I. Ruddat[1†], B. Scholz[2], S. Bergmann[3], A.-L. Buehring[2], S. Fischer[4], A. Manton[5], D. Prengel[3], E. Rauch[3], S. Steiner[3], S. Wiedmann[6], L. Kreienbrock[1] and A. Campe[1]

[1]*Department of Biometry, Epidemiology and Information Processing, WHO Collaborating Centre for Research and Training in Veterinary Public Health, University of Veterinary Medicine, Hannover, Germany;* [2]*Friedrich-Loeffler-Institut, Institute of Animal Welfare and Animal Husbandry, Celle, Germany;* [3]*Department of Veterinary Science, Faculty of Veterinary Medicine, Chair of Animal Welfare, Ethology, Animal Hygiene and Animal Housing, Ludwig-Maximilians-University, Munich, Germany;* [4]*Institute for Animal Breeding and Genetics, University of Veterinary Medicine, Hannover, Germany;* [5]*Department of Farm Animal Ethology and Poultry Science, University of Hohenheim, Stuttgart, Germany;* [6]*Bavarian State Research Center for Agriculture, Kitzingen, Germany*

(Received 4 June 2013; Accepted 6 December 2013; First published online 24 January 2014)

In the context of assessing the impact of management and environmental factors on animal health, behaviour or performance it has become increasingly important to conduct (epidemiological) studies in the field. Hence, the number of investigated farms per study is considerably high so that numerous observers are needed for investigation. In order to maintain the quality and validity of study results calibration meetings where observers are trained and the current level of agreement is assessed have to be conducted to minimise the observer effect. When study animals were rated independently by the same observers by a categorical variable the exclusion test can be performed to identify disagreeing observers. This statistical test compares for each variable and each observer the observer-specific agreement with the overall agreement among all observers based on kappa coefficients. It accounts for two major challenges, namely the absence of a gold-standard observer and different data type comprising ordinal, nominal and binary data. The presented methods are applied on a reliability study to assess the agreement among eight observers rating welfare parameters of laying hens. The degree to which the observers agreed depended on the investigated item (global weighted kappa coefficients: 0.37 to 0.94). The proposed method and graphical description served to assess the direction and degree to which an observer deviates from the others. It is suggested to further improve studies with numerous observers by conducting calibration meetings and accounting for observer bias.

Keywords: inter-rater reliability, observer bias, scoring system, welfare parameters, plumage condition

Implications

Observer reliability is an essential requirement to prevent observer bias in studies where different persons evaluate conditions of livestock husbandry. Despite of standardized evaluation tools observers may differ systematically and substantially from each other. Therefore, it is strongly encouraged to imply calibration meetings as an additional standard tool where observers are trained, the current level of overall agreement is assessed and the direction and degree of (single) observer deviance are identified. When a gold-standard observer cannot be determined and the data type is ordinal, nominal or binary an exclusion test should be applied, which is based on commonly used kappa coefficients.

† Present address: Department of Biometry, Epidemiology and Information Processing, University of Veterinary Medicine, Hannover, Bünteweg 2, D-30559 Hannover, Germany. E-mail: Inga.Ruddat@tiho-hannover.de

Introduction

Animal health and animal behaviour are often measured by observer ratings using scoring systems (Meagher, 2009). Systems exist, for example, to score welfare of laying hens in different housing systems (Blokhuis *et al.*, 2007), to assess lameness of dairy cows (Winckler and Willen, 2001) or to measure behavioural traits in dogs (Svartberg, 2005) containing binary, nominal or ordinal outcomes. In multi-personnel study settings measurements are individually influenced by the rater itself. Therefore, reliability studies are necessary to measure inter-observer agreement and to reduce this measurement bias. When observer trainings are conducted to improve reliability, commonly, the degrees of agreement among all observers are compared between different time points (e.g. after several training sessions: Brenninkmeyer *et al.*, 2007, or before and after training: Thomsen *et al.*, 2008). Additionally, one might want to

analyse the existing data to assess the observer-specific agreement and to conduct an individual training. When physical conditions or behavioural traits are of interest, a gold-standard observer is often not available. Therefore, the reliability of one observer can only be assessed against the collectivity of all other participating raters. Literature comprises numerous methods to assess intra- and inter-observer agreement with most of the analyses based on kappa statistics (e.g. Elbers *et al.*, 2004; Kaler *et al.*, 2009; Pedersen *et al.*, 2011). As kappa statistics can be biased by the marginal distributions of ratings and the number of score levels, the PABAK analysis was developed and is used in several studies (Byrt *et al.*, 1993; Petersen *et al.*, 2004; Thomsen and Baadsgaard, 2006; Brenninkmeyer *et al.*, 2007; March *et al.*, 2007). However, this method does not account for ordinal data. The aim of this study was to describe an improved kappa-based statistical method to determine inter-observer agreement and to identify disagreeing observers in a situation where no gold-standard is available and the ratings can be of ordinal, nominal or binary nature. A reliability study on body condition of laying hens is used to illustrate the method.

Material and methods

Study data

The statistical methods are applied to a reliability study which is part of a network project to improve small group housing systems by assessing the effect of housing and management on laying hen welfare. One part of the investigation considered the body condition of hens. In order to provide comparable data when the body condition is evaluated in different study centres, a reliability study was conducted. Therefore, eight observers with a comparably little experience on evaluating body condition in laying hens were introduced to a scoring system to quantify plumage condition, skin lesions and other health characteristics of layers at the beginning of the network project. The used scoring system is an adapted version of the scoring system developed within the EU LayWel project (Blokhuis *et al.*, 2007). The introduction was performed by a team of well experienced observers in form of a workshop. The training was done before the rating, including all variables except for the overall impression of the hens, which had to be judged without attempting standardization. Afterwards, each observer independently rated the same 40 hens in a varying order. The hens were randomly selected within one experimental station. Hens were selected out of all cages of a housing system similar to those investigated in the network.

In total, 24 qualitative variables were observed (Table 1). The plumage condition was assessed using one binary variable for head condition (1: damaged feathers, 0: no damaged feathers) and six ordinal variables with four levels, rating the condition of neck, back, wing, breast and abdomen (4: <6 feathers damaged, 3: 6 to 10 feathers damaged, 2: 11 to 15 feathers damaged, 1: 16 or more feathers damaged) as well as tail (4: <6 feathers damaged, 3: 6 to 8 feathers damaged, 2: 9 to 12 feathers damaged, 1: 13 or more feathers damaged).

Nine nominal variables with three levels each were observed concerning lesions on comb, head, neck, back, wing, breast, abdomen, cloaca and feet (0: no lesion, 1: covered lesion, 2: fresh, bleeding lesion). Further variables of interest were associated with mites (1: yes, 0: no), keel bone status (ordinal with three levels each: 4: no deformity, 3: mild deformity, 2: moderate/severe deformity), hyperceratosis of foot pad and toe pads (1: moderate/severe, 0: no/mild), epithelial lesion of foot pad and toe pads (both ordinal with four levels each: 4: no lesion and no swelling, 3: superficial lesion and no swelling, 2: moderate-graded lesion and swelling, 1: severe lesion and swelling) as well as the occurrence of broken claws (1: yes, 0: no). Additionally, the overall condition of the hens was assessed reflecting the general impression of each observer (1: bad, 0: good).

Statistical analyses

To assess inter-observer agreement in a situation where a sample of n objects (here laying hens) was rated independently by the same m observers, global kappa coefficients were calculated and exclusion tests were performed. In general, kappa values can be interpreted as the proportion of agreement beyond chance with a possible range of -1 to 1. Kappa values >0 indicate that observers agree beyond chance. The larger the kappa value, the more evidence for inter-observer reliability can be assumed.

Assessing the global agreement. For each variable, a global weighted kappa coefficient $\kappa^{(global)}$ was calculated to quantify the inter-observer agreement among all m observers (see appendix, Krummenauer, 2005 and 2006). Different weight functions were chosen depending on the measurement scale. For dichotomous or nominal scales only agreeing observations (weight 1) and disagreeing observations (weight 0) were differentiated. For ordinal scales the quadratic weight function suggested by Fleiss and Cohen (1973) was used, which varies between 1 and 0 according to the strength of disagreement (see appendix). Asymptotical 95% confidence intervals for global kappa were calculated under normality using an asymptotical variance estimator (see appendix, Krummenauer, 2005).

Identifying disagreeing observers. To identify disagreeing observers an exclusion test was conducted for each observer A, $A = 1, \ldots, m$. For this purpose, the observed and expected agreement between observer A and the remaining $(m-1)$ observers was estimated. These were used to estimate an observer-specific weighted kappa coefficient $\kappa^{(A)}$ (see appendix). If the observer-specific kappa is significantly smaller than the global kappa, there is evidence that the corresponding observer disagrees with others. The test statistic of the exclusion test for observer A, $A = 1, \ldots, m$, is given by

$$\frac{\hat{\kappa}^{(global)} - \hat{\kappa}^{(A)}}{\sqrt{\hat{\mathrm{var}}(\hat{\kappa}^{(global)}) + \hat{\mathrm{var}}(\hat{\kappa}^{(A)}) - 2 \cdot \hat{\mathrm{cov}}(\hat{\kappa}^{(global)}, \hat{\kappa}^{(A)})}}.$$

The null-hypothesis (observer *A* agrees with the other observers) can be rejected with significance level α, if the value of test statistic exceeds the $(1-\alpha)$ quantile of standard normal distribution. Estimators for variance (vâr) and covariance (côv) of kappa statistics are given in the appendix. The significance level for statistical tests conducted in this study is fixed at 5%.

Detailed assessment of disagreeing observers. If an observer is identified as disagreeing significantly, the size and direction of disagreement is of interest. For this purpose, a contingency table was set up comparing the ratings of the deviating observer with the cumulative ratings of the $(m-1)$ other observers. The proportion of agreeing and disagreeing ratings was analysed graphically. To check the asymmetry of the table the cumulative frequencies were divided by $(m-1)$ and, depending on the size of the contingency table, the McNemar test or the Bowker test was conducted using PROC FREQ in SAS 9.3 (SAS Institute Inc., 2012). Test results and graphics help to assess if the observer rated non-systematically differently or took systematically higher or lower score levels than others. In case of a non-systematically disagreeing observer, the definition of the variable itself may be unclear and should be clarified in further training. If the observer disagreed systematically, a training session with regard to defining the specific rating levels should be performed.

Special case: homogeneous study population. Assessing agreement of ratings is reasonable only if the animals differ in their conditions concerning the variable of interest. To check this, we examined for each variable the most frequently given score per hen. If at least two hens were differently scored, we accepted the study population as being heterogeneous.

Special case: an observer rated all objects identically. If one observer, say observer *Z*, rated all hens identically with the same score level concerning one variable, the calculated observed agreement and the expected agreement between this observer and any other observer is equal. Consequently, the estimates for his or her observer-specific kappa and the according variance are both zero and it is not reasonable to calculate the exclusion test statistic. If there was evidence for a heterogeneous study population for this variable (for decision rule see above), a 41st artificial hen was constructed, to which identical ratings by all observers were assigned, with a score level differing from that of observer *Z*. The analysis of this variable was then conducted with the modified dataset.

Results

In total the inter-observer agreement was assessed for 17 out of 24 variables. Fifteen of these 17 variables were analysed directly using the exclusion test. For the two variables lesion at cloaca and occurrence of broken claws the analysis was conducted after adding a 41st artificial hen to the dataset. The ratings concerning the variables lesions at head, neck, back, wing, abdomen and feet as well as occurrence of mites led to the assumption of homogeneous study population (Table 1). Therefore, the agreement of observers was not assessed.

The estimates of global kappa coefficients, presented in Figure 1, show that the overall agreement was beyond chance for all variables with estimates varying between 0.37 (lesions at comb) and 0.94 (plumage condition of back).

Using the exclusion test we identified one significantly disagreeing observer for head plumage condition, two for tail plumage condition, one for foot pad hyperceratosis and one for toe pad hyperceratosis. The detailed assessment of disagreeing observers is displayed in Table 2 and Figure 2 for plumage condition of tail with two disagreeing observers, here named as observer *X* and observer *Y*. In 15.5% of cases observer *X* assessed the hens' plumage condition with a lower score, in 22.0% of cases with a higher score compared with the others and therefore differed non-systematically. The Bowker test showed no statistically significant asymmetry with $P = 0.954$. In 50% of the cases observer *Y* assessed the plumage condition with a lower score and in 4% of cases with a higher score in comparison to the other observers. With $P = 0.008$ this observer differed systematically from others. In the cases of plumage condition at head and hyperceratosis at foot pad results of the McNemar test showed that identified observers did not differ systematically ($P = 0.550$, 0.210). In the case of hyperceratosis at toe pads the McNemar test result showed that the identified observer differed systematically ($P = 0.006$).

As the chosen methods do not account for missing values, hens with incomplete ratings were not included in the agreement analysis. Accordingly, the sample size is reduced in some cases (see Table 1 for frequencies).

Discussion

In this paper kappa-based methods are provided for analysing observer-specific agreement in settings where a fixed group of observers rates the same objects during a one-time calibration meeting. The presented exclusion test was developed by Krummenauer (2005 and 2006) for application in the field of improving diagnostic findings in human medicine with the idea, that excluding identified observers will make the remaining ratings more consistent. In the application of reliability studies all participating observers will participate in future studies as well. Therefore, it is not reasonable to exclude identified observers but to train them specifically to improve the overall agreement.

Kappa coefficients are frequently criticised (e.g. Byrt *et al*., 1993) as the magnitude of kappa strongly depends on the marginal distributions of ratings, the number of score levels and the applied weights for calculations. Unweighted kappa decrease when more score levels exist, while weighted kappa with quadratic weight functions increase with the

Table 1 *Observed proportions of ratings for the body condition variables included in the reliability study (eight observers, 40 hens)*

		Marginal proportions for categories in %					
Variable	Scale[1]	0	1	2	3	4	na
Overall condition	Binary	88.42	11.58				
Plumage condition of ...							
Head	Binary	88.13	9.69				2.19
Neck	Ordinal (4)		–	8.75	35.63	55.31	0.31
Back	Ordinal (4)		5.63	8.75	5.31	80.0	0.31
Wing	Ordinal (4)		–	2.19	11.25	86.56	
Tail	Ordinal (4)		6.88	10.63	24.83	58.13	
Breast	Ordinal (4)		21.56	35.00	30.00	13.44	
Abdomen	Ordinal (4)		–	3.75	29.69	66.56	
Lesions at ...							
Comb	Nominal (3)	29.38	68.13	2.50			
Head	Nominal (3)	99.06	0.94	–			
Neck	Nominal (3)	100.00	–	–			
Back	Nominal (3)	97.50	2.50	–			
Wing	Nominal (3)	100.00	–	–			
Breast	Nominal (3)	90.63	9.38	–			
Abdomen	Nominal (3)	99.69	0.31	–			
Cloaca	Nominal (3)	96.88	3.13	–			
Feet	Nominal (3)	99.06	0.94	–			
Further variables							
Mites	Binary	89.38	0.64				9.69
Keel bone status	Ordinal (3)			8.75	27.50	63.44	0.31
Hyperceratosis[2]	Binary	74.06	25.94				
Hyperceratosis[3]	Binary	79.37	20.63				
Epithelial lesion[2]	Ordinal (4)		4.06	29.38	6.88	59.38	0.31
Epithelial lesion[3]	Ordinal (4)		2.81	2.81	5.63	88.75	
Broken claws	Binary	95.31	3.44				1.25

na: missing values, – : no observations.
[1](Number of levels).
[2]Of foot pads.
[3]Of toe pads.

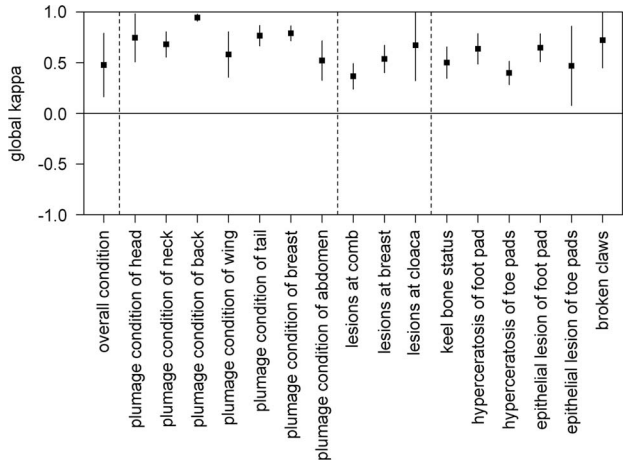

Figure 1 Global kappa coefficients with 95% asymptotic confidence intervals.

number of score levels (Sim and Wright, 2005). Therefore, a straightforward interpretation is difficult. The statistical test method used in this study is based on the difference between two kappa coefficients (global *v.* observer-specific) and is therefore partly based on the same ratings. Assuming that the prevalences of scores and the marginal distributions in both sets of data are similar, the two coefficients are comparable and the results of the exclusion test are reliable. However, it is not recommended to compare kappa values between different variables or different studies. Thus, we decided to avoid using published benchmarks for coefficient interpretation (Landis and Koch, 1977), which are, as commented by the authors, 'clearly arbitrary'.

Independently from the statistical method, the objects to be rated should show certain variability in their conditions to evaluate observer agreement properly concerning the variables of interest. If this is not given, the true reliability concerning the variable cannot be assessed. In our example we accepted the study population to be heterogeneous, if at least two hens had a different most frequent score. It should be noted that this is an arbitrarily chosen cut-off and might be worth a discussion. As true conditions are unknown, reasons for the absence of heterogeneity in ratings might be on the one hand that the scoring system was not discriminating enough. Another reason could be that the study

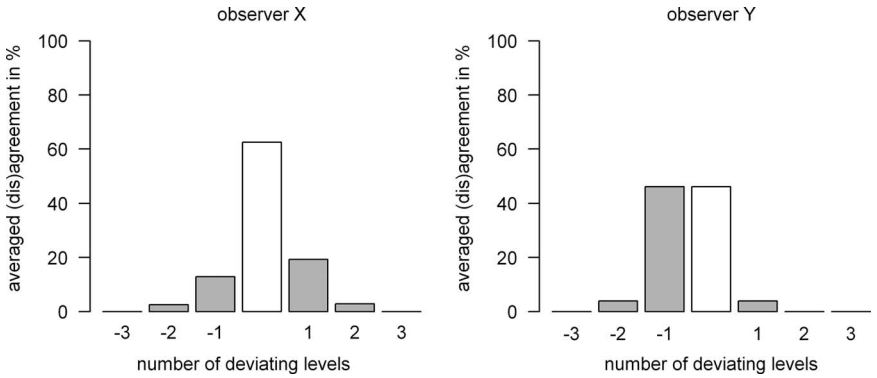

Figure 2 Proportions of (dis)agreement for the two statistically significant disagreeing observers in rating plumage condition of tail (four levels) for 40 laying hens. The white bar indicates the proportion of ratings agreeing with the remaining seven observers, the left grey bars of each figure indicate disagreement in rating lower scores than others, the right grey bars indicate disagreement in rating higher scores than others, respectively ($100\% = 40 \times 7$).

Table 2 *(a) Comparison of ratings from the statistically significant disagreeing observer* X *with the distribution of ratings from the remaining seven observers for plumage condition of tail for 40 laying hens ($40 \times 7 = 280$, levels: 4: <6 feathers damaged, 3: 6 to 8 feathers damaged, 2: 9 to 12 feathers damaged, 1: 13 or more feathers damaged); (b) analogue table for the disagreeing observer* Y

		Distribution of ratings from the other seven observers						Distribution of ratings from the other seven observers			
(a)	Score	1	2	3	4	(b)	Score	1	2	3	4
Observer *X*	1	0	6	1	0	Observer *Y*	1	14	5	2	0
	2	16	8	5	6		2	5	15	27	9
	3	5	12	35	25		3	0	6	30	97
	4	0	3	26	132		4	0	0	0	70

population was too homogeneous. Homogeneity was apparent in our study and unfortunately could not be avoided due to the investigated hens originating from the same housing system. In order to offer comparable results for agreement assessment in the situation where one (or more) observer(s) rated all hens identically and others did not, we analysed the data including an additional hen, to investigate if these differences between observers are given by random or if they are significant. The approach of creating one additional hen to enable the assessment of agreement is based on the idea of adding constants in cases of empty cells for contingency table analysis, which is performed by many researchers (Agresti, 2002). In conclusion, we decided that this approach is helpful to deal with this situation seeing that the alternative would be to exclude the concerned observer or even the whole variable from assessment.

A further requirement for a proper evaluation of observer agreement is that each score level of the variable of interest should have been observed. If score levels are provided but not assigned at all, this implicates that the agreement and reliability of observers concerning these score levels cannot be assessed. If score levels are given with low frequency, the methods described can be applied, but validity of results can be limited for the concerned score levels. Either way, the observed frequencies (as shown in Table 1) should be considered in any case for interpreting the calculated agreements.

The statistical methods are presented here in the context of reliability studies, but can be applied in any situation with binary, nominal or ordinal outcomes where the same observers rated the same animals (or other objects of interest). Multi-personnel study design is common in field studies when diagnostic findings depend on observer ratings in the context of clinical symptoms (e.g. Elbers *et al*., 2004), when health scoring systems are applied (e.g. Blokhuis *et al*., 2007) or when behavioural traits shall be assessed (e.g. Ott *et al*., 2011). In any multi-observer study measures have to be taken before the initiation of a study to minimise the observer effect. Therefore, it is advisable to conduct calibration meetings where observers are trained and the current level of agreement is assessed. Generally, the aspired level of agreement has to be defined under consideration of the investigated items. In doing so, it is useful to get an overall impression of the observer agreement per investigated item by employing a statistical test. The overall agreement between all observers should be calculated and significantly deviating observers should be identified. The proposed approach can be applied on studies where no gold-standard observer can be determined. This implies the possibility of applying the method on diagnostic test evaluation studies, where no gold-standard test is available, as well. However, kappa-based methods depend on the apparent prevalence (Gardner *et al*., 2000). Therefore, latent class models are preferable for evaluation studies. Nevertheless, as regards

diagnostic tests, our kappa-based method can be applied on laboratory validation studies (ring-trials).

In conclusion, the presented kappa statistics are appropriate for assessing agreement among multiple observers in ordinal, nominal and binary data and a statistical test is used to identify disagreeing observers. Graphics are provided to describe the direction and degree of deviance. Results obtained applying these methods can be used in reliability studies to train observers more individually and correct for the identified bias in order to improve agreement among several observers. To improve the quality of a reliability study in general, it must be ensured that the study population is heterogeneous concerning all variables among all score levels of interest. In the presented example the method was applied to laying hens and body conditions. However, in general the method can be applied to any binary, nominally or ordinally scaled measures to assess the welfare (EFSA Panel on Animal Health and Welfare, 2012), health and behaviour of any animal species and even to analyse laboratory validation studies (ring-trials). Conclusively, it is suggested to further improve studies with numerous observers by conducting calibration meetings and accounting for observer bias by applying an approach like the one presented here.

Acknowledgements

This work was partially financed by the German Federal Ministry of Food, Agriculture and Consumer Protection (BMELV) through the Federal Agency of Agriculture and Nutrition (BLE), grant number PGI-06.01-28-1-36.004-07. The authors thank all participating partners and the farm manager at the Friedrich-Loeffler-Institut in Celle for providing the hens and settings for this study.

References

Agresti A 2002. Categorical data analysis. Wiley-Interscience, Hoboken, NJ, USA.

Blokhuis HJ, Van Niekerk TF, Bessei W, Elson A, Guemene D, Kjaer JB, Levrino GAM, Nicol CJ, Tauson R, Weeks CA and De Weerd HAV 2007. The LayWel project: welfare implications of changes in production systems for laying hens. Worlds Poultry Science Journal 63, 101–114.

Brenninkmeyer C, Dippel S, March S, Brinkmann J, Winckler C and Knierim U 2007. Reliability of a subjective lameness scoring system for dairy cows. Animal Welfare 16, 127–129.

Byrt T, Bishop J and Carlin JB 1993. Bias, prevalence and kappa. Journal of Clinical Epidemiology 46, 423–429.

EFSA Panel on Animal Health and Welfare 2012. Statement on the use of animal-based measures to assess the welfare of animals. EFSA Journal 10, 1–29.

Elbers ARW, Vos JH, Bouma A and Stegeman JA 2004. Ability of veterinary pathologists to diagnose classical swine fever from clinical signs and gross pathological findings. Preventive Veterinary Medicine 66, 239–246.

Fleiss JL and Cohen J 1973. Equivalence of weighted kappa and intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement 33, 613–619.

Gardner IA, Stryhn H, Lind P and Collins MT 2000. Conditional dependence between tests affects the diagnosis and surveillance of animal diseases. Preventive Veterinary Medicine 45, 107–122.

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

Krummenauer F 2005. Methoden zur Evaluation bildgebender Verfahren von begrenzter Reproduzierbarkeit. Shaker Verlag, Aachen, Germany.

Krummenauer F 2006. The comparison of clinical imaging devices with respect to parallel readings in both devices. European Journal of Medical Research 11, 119–122.

Landis JR and Koch GG 1977. Measurement of observer agreement for categorical data. Biometrics 33, 159–174.

March S, Brinkmann J and Winkler C 2007. Effect of training on the inter-observer reliability of lameness scoring in dairy cattle. Animal Welfare 16, 131–133.

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

Ott S, Schalke E, Campe A and Hackbarth H 2011. Urteilsübereinstimmung bei zwei Beobachterpaaren in einem Verhaltenstest für Hunde. In Proceedings of 16. Internationale DVG-Fachtagung zum Thema Tierschutz, Nürtingen, Deutschland, pp. 307–319.

Pedersen KS, Holyoake P, Stege H and Nielsen JP 2011. Observations of variable inter-observer agreement for clinical evaluation of faecal consistency in grow-finishing pigs. Preventive Veterinary Medicine 98, 284–287.

Petersen HH, Enoe C and Nielsen EO 2004. Observer agreement on pen level prevalence of clinical signs in finishing pigs. Preventive Veterinary Medicine 64, 147–156.

SAS Institute Inc. 2012. SAS/SAT user's guide. SAS Institute Inc., Cary, NC, USA.

Sim J and Wright CC 2005. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Physical Therapy 85, 257–268.

Svartberg K 2005. A comparison of behaviour in test and in everyday life: evidence of three consistent boldness-related personality traits in dogs. Applied Animal Behaviour Science 91, 103–128.

Thomsen PT and Baadsgaard NP 2006. Intra- and inter-observer agreement of a protocol for clinical examination of dairy cows. Preventive Veterinary Medicine 75, 133–139.

Thomsen PT, Munksgaard L and Togersen FA 2008. Evaluation of a lameness scoring system for dairy cows. Journal of Dairy Science 91, 119–126.

Winckler C and Willen S 2001. The reliability and repeatability of a lameness scoring system for use as an indicator of welfare in dairy cattle. Acta Agriculturae Scandinavica Section a-Animal Science 51, 103–107.

Appendix

Global kappa and observer-specific kappa

Let n be the number of objects, which are rated by the same m observers for a variable with c categories. Concerning the ratings of two observers A and B with $A, B = 1, \ldots, m$ and $A \neq B$, p_{ij} denotes the relative frequency for observer A giving category i in combination with observer B giving category j with $i, j = 1, \ldots, c$. Furthermore, $p_i^{(A)}$ denotes the relative frequency for observer A giving category i over all n objects and $p_j^{(B)}$ denotes the relative frequency for observer B giving category j over all n objects. The observed and expected agreement between observer A and B is estimated by

$$\hat{o}^{(A,B)} = \sum_{i=1}^{c} \sum_{j=1}^{c} w_{ij}^{(A,B)} p_{ij} \quad \text{and}$$

$$\hat{e}^{(A,B)} = \sum_{i=1}^{c} \sum_{j=1}^{c} w_{ij}^{(A,B)} p_i^{(A)} p_j^{(B)},$$

where $w_{ij}^{(A,B)}$ denotes the weight associated with ratings i and j, $i, j = 1, \ldots, c$. Within this study, for ordinal variables the quadratic weight function

$$w_{ij}^{(A,B)} = 1 - \frac{(i-j)^2}{(c-1)^2}$$

was chosen (Fleiss and Cohen, 1973). For nominal or binary variables

$$w_{ij}^{(A,B)} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

was used, which correspond to unweighted agreement calculations.

The *global weighted* kappa is then estimated by

$$\hat{\kappa}^{(global)} = \frac{\hat{o}^{(global)} - \hat{e}^{(global)}}{1 - \hat{e}^{(global)}}$$

with an estimated observed global agreement of

$$\hat{o}^{(global)} = \frac{2}{m(m-1)} \sum_{A=1}^{m-1} \sum_{B>A}^{m} \hat{o}^{(A,B)}$$

and an estimated expected global agreement of

$$\hat{e}^{(global)} = \frac{2}{m(m-1)} \sum_{A=1}^{m-1} \sum_{B>A}^{m} \hat{e}^{(A,B)}.$$

The corresponding variance estimator can be calculated by

$$\hat{\text{var}}\left(\hat{\kappa}^{(global)}\right) = \frac{\frac{1}{n}\sum_{k=1}^{n} \left(d^{(global)[k]}\right)^2 - \left(s^{(global)}\right)^2}{n\left(1 - \hat{e}^{(global)}\right)^4},$$

with

$$d^{(global)[k]} = \left(1 - \hat{e}^{(global)}\right) \sum_{A=1}^{m} \sum_{\substack{B=1 \\ B \neq A}}^{m} \frac{w_{i_A i_B}^{(A,B)[k]}}{m(m-1)} - 2\left(1 - \hat{o}^{(global)}\right) \sum_{A=1}^{m} \sum_{\substack{B=1 \\ B \neq A}}^{m} \frac{\sum_{i=1}^{c} p_i^{(A)} \cdot w_{ij_B}^{(A,B)[k]}}{m(m-1)},$$

$$s^{(global)} = \hat{e}^{(global)} \cdot \hat{o}^{(global)} - 2 \cdot \hat{e}^{(global)} + \hat{o}^{(global)},$$

where i_A and j_B denote the specific category given by observer A and B for object k, $k = 1, \ldots, n$ (Krummenauer, 2005).

The *observer-specific weighted* kappa for observer A, $A = 1, \ldots, m$, is estimated by

$$\hat{\kappa}^{(A)} = \frac{\hat{o}^{(A)} - \hat{e}^{(A)}}{1 - \hat{e}^{(A)}},$$

where the observed and expected agreement between observer A and the $(m-1)$ others is estimated by

$$\hat{o}^{(A)} = \frac{1}{m-1} \sum_{\substack{B=1 \\ B \neq A}}^{m} \hat{o}^{(A,B)} \quad \text{and} \quad \hat{e}^{(A)} = \frac{1}{m-1} \sum_{\substack{B=1 \\ B \neq A}}^{m} \hat{e}^{(A,B)}.$$

The corresponding variance estimator can be calculated by

$$\hat{\text{var}}\left(\hat{\kappa}^{(A)}\right) = \frac{\frac{1}{n}\sum_{k=1}^{n} \left(d^{(A)[k]}\right)^2 - \left(s^{(A)}\right)^2}{n\left(1 - \hat{e}^{(A)}\right)^4}$$

with

$$d^{(A)[k]} = \left(1 - \hat{e}^{(A)}\right) \sum_{\substack{B=1 \\ B \neq A}}^{m} \frac{w_{i_A i_B}^{(A,B)[k]}}{m-1} - 2\left(1 - \hat{o}^{(A)}\right) \sum_{\substack{B=1 \\ B \neq A}}^{m} \frac{\sum_{i=1}^{c} p_i^{(A)} w_{ij_B}^{(A,B)[k]} + \sum_{j=1}^{c} p_j^{(B)} w_{i_A j}^{(A,B)[k]}}{m-1},$$

$$s^{(A)} = \hat{e}^{(A)} \cdot \hat{o}^{(A)} - 2 \cdot \hat{e}^{(A)} + \hat{o}^{(A)},$$

where i_A and j_B denote the specific category given by observer A and B for object k, $k = 1, \ldots, n$.

The covariance between the global kappa and the observer-specific kappa is estimated by

$$\hat{\text{cov}}\left(\hat{\kappa}^{(global)}, \hat{\kappa}^{(A)}\right) = \frac{\frac{1}{n}\sum_{k=1}^{n} d^{(global)[k]} \cdot d^{(A)[k]} - s^{(global)} \cdot s^{(A)}}{n\left(1 - \hat{e}^{(global)}\right)^2 \left(1 - \hat{e}^{(A)}\right)^2}$$

with $d^{(global)[k]}$, $d^{(A)[k]}$, $s^{(global)}$ and $s^{(A)}$ as described above (Krummenauer, 2005).