Italian Journal of Animal Science

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tjas20

Evaluation of inter-observer reliability in the case of trichotomous and four-level animal-based welfare indicators with two observers

Benedetta Torsiello, Mauro Giammarino, Piero Quatto, Monica Battini, Silvana Mattiello, Luca Battaglini & Manuela Renna

To cite this article: Benedetta Torsiello, Mauro Giammarino, Piero Quatto, Monica Battini, Silvana Mattiello, Luca Battaglini & Manuela Renna (2024) Evaluation of inter-observer reliability in the case of trichotomous and four-level animal-based welfare indicators with two observers, Italian Journal of Animal Science, 23:1, 938-960, DOI: 10.1080/1828051X.2024.2367681

To link to this article: https://doi.org/10.1080/1828051X.2024.2367681

ITALIAN JOURNAL OF ANIMAL SCIENCE
2024, VOL. 23, NO. 1, 938–960
https://doi.org/10.1080/1828051X.2024.2367681

RESEARCH ARTICLE

OPEN ACCESS

Check for updates

Evaluation of inter-observer reliability in the case of trichotomous and four-level animal-based welfare indicators with two observers

Benedetta Torsiello[a], Mauro Giammarino[b], Piero Quatto[c], Monica Battini[d], Silvana Mattiello[d], Luca Battaglini[a] and Manuela Renna[e]

[a]Dipartimento di Scienze Agrarie, Forestali e Alimentari, Università degli Studi di Torino, Grugliasco, Italy; [b]Dipartimento di Prevenzione, Asl TO3, Servizio Veterinario, Area Sanità Animale, Piossasco, Italy; [c]Dipartimento di Economia, Metodi Quantitativi e Strategie d'Impresa, Università degli Studi Milano-Bicocca, Milano, Italy; [d]Dipartimento di Scienze Agrarie e Ambientali - Produzione, Territorio, Agroenergia, Università degli Studi di Milano, Milano, Italy; [e]Dipartimento di Scienze Veterinarie, Università degli Studi di Torino, Grugliasco, Italy

ABSTRACT

This study focuses on assessing inter-observer reliability (IOR) between two observers in the case of trichotomous and four-level animal-based welfare indicators assessed at individual level. The Body Condition Score (BCS) and Knee calluses (KNC) were chosen as trichotomous indicators; data were collected in fourteen intensively managed dairy goat farms in Italy (ITF1 to ITF7) and Portugal (PTF1 to PTF7) and in extensively managed dairy goat farms exploiting three alpine pastures (AP1, AP2 and AP3) in Italy. The Ear posture (EP) and Eye white (EW) were chosen as four-level indicators; data were collected in three intensively managed dairy cattle farms (F1, F2 and F3) in Italy. The performance of the most documented agreement indices was compared. In the case of trichotomous indicators, Scott's π, Cohen's K, Cohen's K_C, Cohen's *weighted* K and Krippendorff's α were affected by the paradox effect: when the concordance rate (P_0) was high, they sometimes gave very low or even negative values (e.g. $P_{0(\text{BCS-ITF3})} = 74\%$; Scott's $\pi = 0.05$; Cohen's $K = 0.09$; Krippendorff's $\alpha = 0.06$; $P_{0(\text{BCS-AP3})} = 74\%$; Scott's $\pi = -0.12$; Cohen's K = Krippendorff's $\alpha = -0.11$). Bangdiwala's B, Gwet's $\gamma(AC_1)$ and Quatto's *weighted* S were not affected by this phenomenon and provided values very close to P_0 (e.g. $P_{0(\text{KNC-PTF1})} = 88\%$; Bangdiwala's B = Gwet's $\gamma(AC_1) = 0.85$; $P_{0(\text{BCS-AP1})} = 82\%$; Bangdiwala's B = Gwet's $\gamma(AC_1) = 0.79$). In the case of four-level indicators, Cohen's K and Krippendorff's α were not affected by the paradox behaviour. However, Cohen's K_C in some cases exceeded the observed P_0 (e.g. $P_{0(\text{EP-F3})} = 78\%$; Cohen's $K_C = 1$). Gwet's $\gamma(AC_1)$ showed the best results for four-level indicators (e.g. $P_{0(\text{EP-F1})} = 88\%$; Gwet's $\gamma(AC_1) = 0.86$), followed by Quatto's S and Holley and Guilford's G (e.g. $P_{0(\text{EP-F1})} = 88\%$; Quatto's S = Holley and Guilford's $G = 0.84$). To evaluate IOR between two observers, Bangdiwala's B, Gwet's $\gamma(AC_1)$ and Quatto's *weighted* S are suggested for trichotomous indicators, while Gwet's $\gamma(AC_1)$, Quatto's S and Holley and Guilford's G are suggested for four-level indicators.

ARTICLE HISTORY
Received 20 November 2023
Revised 13 May 2024
Accepted 22 May 2024

KEYWORDS
Agreement index; animal-based measure; inter-observer reliability; bootstrap method; three- and four-level indicators

HIGHLIGHTS

- Scott's π, Cohen's K, Cohen's K_C, Cohen's *weighted* K and Krippendorff's α can be affected by a paradox behaviour.
- Bangdiwala's B, Gwet's $\gamma(AC_1)$ and Quatto's *weighted* S are suggested to evaluate IOR between two observers for trichotomous indicators.
- Gwet's $\gamma(AC_1)$, Quatto's S and Holley and Guilford's G are suggested to evaluate IOR between two observers for four-level indicators.

Introduction

Animal-based welfare indicators are considered the most suitable for a comprehensive welfare assessment, as they are based on evaluations made on the animal itself (EFSA 2012; De Rosa et al. 2015). Animal-based indicators currently included in welfare assessment protocols are mainly dichotomous variables [e.g. udder asymmetry in the Animal Welfare Indicators (AWIN) welfare assessment protocol for goats; scores: 0 = absence of asymmetry; 1 = presence of asymmetry

CONTACT Mauro Giammarino mauro.giammarino@aslto3.piemonte.it

Supplemental data for this article can be accessed online at https://doi.org/10.1080/1828051X.2024.2367681.

(AWIN 2015a); coughing in the Welfare Quality® assessment for pigs; scores: 0 = no evidence of coughing; 2 = evidence of coughing (Welfare Quality® 2009a)]. However, trichotomous and four-level indicators are also found. Examples of trichotomous animal-based welfare indicators are the foot pad dermatitis in the Welfare Quality® Assessment protocol for poultry [scores: 0 = feet intact, no or minimal proliferation of epithelium; 1 = necrosis or proliferation of epithelium or chronic bumble foot with no or moderate swelling; 2 = swollen (dorsally visible); Welfare Quality® 2009b] and the bursitis in the Welfare Quality® assessment for pigs [scores: 0 = no evidence of bursae; 1 = one or several small bursae on the same leg or one large bursa; 2 = several large bursae on the same leg, or one extremely large bursae, or any bursa that is eroded (Welfare Quality® 2009a). Among the four-level indicators included in welfare assessment protocols, it is possible to find the body and head lesions in the AWIN welfare assessment protocol for sheep [scores: 0 = no lesions; 1 = minor lesions; 2 = major lesions; 3 = myiasis (AWIN 2015b)], and the lesions at mouth corners in the AWIN welfare assessment protocol for horses [scores: 0 = no lesion; 1 = hardened spots; 2 = redness; 3 = open wounds; (AWIN 2015c)]. Other examples of trichotomous and four-level animal-based welfare indicators can be found in published literature (e.g. Buczinski et al. 2016; Munoz et al. 2017; Navarro et al. 2020; Nannarone et al. 2024).

The inclusion of animal-based welfare indicators into welfare assessment protocols implies that such indicators must be valid, feasible and reliable (Vieira et al. 2018). Reliability needs to be assessed both when an observer performs the welfare assessment on the same subjects several times (intra-observer reliability) and when different observers perform the welfare assessment on the same subjects contemporarily and independently one from the other (inter-observer reliability; IOR) (Martin and Bateson 2007). To assess the IOR, the level of agreement among the observers is calculated processing the scores assigned by the observers to each variable using different statistical indices, defined as agreement indices. If the percentage of agreement (i.e. concordance rate, P_0) among observers is low, the reliability of the indicator will be equally low; therefore, the indicator will not be suitable to assess animal welfare properly and it will need to be redefined (De Rosa et al. 2009).

In published literature, the agreement indices belonging to the Kappa statistics are the most implemented ones for the evaluation of IOR of trichotomous and four-level categorical animal-based welfare indicators assessed at individual level. Even though it is not our purpose here to give an exhaustive literature review, we intend to provide some examples. Cohen's *K* (Cohen 1960) was implemented both by Pedersen et al. (2011) when assessing the reliability of a three-level faecal consistency in growing pigs, and by Buczinski et al. (2016) when evaluating the IOR for four-level indicators (namely rectal temperature, cough, ocular discharge, nasal discharge, and ear position) in pre-weaned dairy cattle. Cohen's *weighted K* (Cohen 1968) was instead implemented both by Vieira et al. (2018) who evaluated the IOR of BCS and Knee calluses (KNC) in dairy goats, and by Munoz et al. (2017) who evaluated the IOR of the trichotomous indicators fleece conditions and hoof overgrowth, of the four-level indicator foot-wall integrity, and of a 5-level BCS, in dairy ewes. Thomsen and Baadsgaard (2006) evaluated the IOR of the trichotomous indicators lameness and cutaneous lesions in dairy cattle using prevalence-adjusted, bias-adjusted kappa (PABAK) (Byrt et al. 1993). Czycholl et al. (2019) assessed the reliability of the Horse Grimace Scale (a combination of different animal-based welfare indicators evaluated using a 3-level assessment scale), of 4-level integument alterations assessed in various parts of the body of the horse, and of a 5-level BCS, contemporarily using Cohen's *K*, Cohen's *weighted K* and PABAK.

However, the Kappa statistics are sometimes affected by a paradoxical behaviour (Feinstein and Cicchetti 1990) and other agreement indices have therefore been proposed in literature (Giammarino et al. 2021). A critical issue is that, when assessing the reliability, a part of the agreement among the observers might be due to chance, being defined as '*chance agreement*'. During the evaluation of the agreement among observers, the rate of agreement due to chance (P_e) must be removed from the rate of the observed agreement (P_0) (Gwet 2001). To assess the agreement among observers properly, it is essential to determine the most appropriate way to calculate the rate of agreement due to chance (Gwet 2001). For this purpose, many chance-corrected agreement indices, used in the case of the presence of two observers, are proposed in the literature. For example, Scott (1955) assumed that the chance agreement is related to the classification probabilities of the subjects within the same category by the two observers. Cohen (1960) criticised this assumption, since the classification of all the subjects within the same category means that the chance agreement is equal to 1 and that the IOR is 0. Therefore, Scott's π (Scott 1955) is suitable only when the level of agreement between the observers in assigning the subjects to the same category is poor, so that the rate of agreement due to chance results lower. Chance agreement calculation of Cohen's *K* (Cohen 1960) differs from that of Scott's π; indeed, for the implementation of the rate of agreement due to chance, Cohen considered the number of times that the observers assign the subjects to each of the considered categories. Despite this, Cohen's *K* is characterised by the same problems that affect Scott's π: when the observers assign all the subjects to the same category, the chance agreement will be equal to 1. Consequently, when the agreement due to chance is high, Cohen's *K* assumes a low value, despite a high observed P_0. As stated by Feinstein and Cicchetti (1990), this is due to the unbalanced marginal distributions within the concordance matrix. According to Bennet et al. (1954), the chance agreement can also be considered as the inverse of the number of categories. Subsequently, this principle was proposed by Holley and Guilford (1964) by means of the Holley and Guilford's *G*, and later by Falotico and Quatto (2010) by means of Quatto's *S* (2004), these indices being closely related to each other. As Holley and Guilford's *G* and Quatto's *S*, Gwet's $\gamma(AC_1)$ (Gwet 2008) considers the number of the categories that characterises the variable, but the implementation of the chance agreement is different and more complex. According to Gwet (2008), not only the number of categories characterising the variable, but also the frequency with which the scores are attributed to each subject by each involved observer, must be considered.

The choice of the agreement indices is not only linked to the number of categories which characterises the variable under analysis, but also to the number of observers involved during the evaluation process (Gisev et al. 2013). For this reason, it is crucial to calculate agreement indices which can estimate the concordance between two or more observers properly, conferring reliable agreement results (Gwet 2001) and guaranteeing the possibility of including new animal-based welfare indicators in welfare assessment protocols (Vieira et al. 2018).

In a previous study, Giammarino et al. (2021) identified Bangdiwala's *B* (Bangdiwala 1985) and Gwet's $\gamma(AC_1)$ (Gwet 2008) as the best agreement indices to evaluate the IOR between two observers in the case of dichotomous categorical animal-based welfare indicators. With this study, we aimed at identifying the best indices for measuring the agreement between two observers, and calculating the related confidence intervals, when evaluating trichotomous and four-level animal-based welfare indicators. To do so, we selected two trichotomous animal-based indicators, namely the BCS and KNC from a prototype (Battini et al. 2016; Can et al. 2016) and a modified (Battini et al. 2021) Animal Welfare Indicators (AWIN) welfare assessment protocol for goats (AWIN 2015a), and two four-level animal-based indicators from published literature (Battini et al. 2019), namely the EP and EW in dairy cows, and we used them as examples to test the performance of the most documented agreement indices proposed in the literature.

Materials and methods

Dataset

Trichotomous animal-based welfare indicators

A prototype of the AWIN welfare assessment protocol was applied by two observers in seven intensively managed dairy goat farms in Italy (ITF1, $n = 49$; ITF2, $n = 37$; ITF3, $n = 43$; ITF4, $n = 30$; ITF5, $n = 30$; ITF6, $n = 34$; ITF7, $n = 39$) and in seven intensively managed dairy goat farms in Portugal (PTF1, $n = 48$; PTF2, $n = 38$; PTF3, $n = 25$; PTF4, $n = 39$; PTF5, $n = 32$; PTF6, $n = 38$; PTF7, $n = 35$) between January and March 2014 (Battini et al. 2016; Can et al. 2016). The two Italian observers had different background and experience with dairy goats, as one was an animal scientist with more than three years of experience with dairy goats, while the other was a veterinarian without any experience with dairy goats. On the other hand, the two Portuguese observers had both a veterinary background but different level of experience, as one had more than three years of experience with dairy goats, while the other was just graduated from a veterinary school (Vieira et al. 2018). From the application of this prototype, we used the data collected for two trichotomous welfare indicators assessed at individual level, namely the BCS and KNC.

In addition, further BCS data to be used in the current study were obtained from the application of a modified AWIN protocol for goat welfare assessment (Battini et al. 2021) by two observers in extensively managed dairy goat farms exploiting three alpine pastures (AP1, $n = 44$; AP2, $n = 70$; AP3, $n = 46$) in Italy between June and August 2021. In this case, the observers were students enrolled in the second year of the MSc in Animal Science and of the MSc in Science and Technologies of Forest Systems and Territories at the University of Turin (Italy). Both observers had no previous experience with dairy goats. Before data collection, the observers received a common training on goat welfare assessment, including both theoretical and practical sessions, given by one author of the original AWIN welfare assessment protocol for goats kept in intensive or semi-intensive production systems (AWIN 2015a). They also received, as training material, both the original AWIN welfare assessment protocol for goats (AWIN 2015a) and a publication on the application of the AWIN welfare assessment protocol for goats under semi-extensive conditions (Battini et al. 2021).

Each goat was assigned to one of three mutually exclusive and exhaustive categories. For BCS: very thin goat $= -1$; normal goat $= 0$; very fat goat $= 1$; for KNC: no lesions, hair loss or skin thickening $= 0$; skin damage with/without hair loss and reddened skin, but no enlargement of any joint $= 1$; skin damage with hair loss, and enlargement of at least one joint, showing a thick callus $= 2$.

Four-level animal-based welfare indicators

In the current study, we used data from 219 photos taken from March to June 2018 in three intensively managed dairy cattle farms (F1, $n = 126$; F2, $n = 42$; F3, $n = 51$) located in Italy. Each photo was scored by two observers for EP and EW. Following the classification proposed by Battini et al. (2019), each cow was assigned to one of four mutually exclusive categories. Considering EW: eye white clearly visible $= 1$; eye white barely visible $= 2$; eye white not visible, with eye normally open $= 3$; half-closed eye $= 4$. Considering EP: ears held up $= 1$; ears held horizontally $= 2$; ears held back along the head $= 3$; ears held downwards $= 4$. The observers were students of the MSc in Animal Production Sciences and Technologies of the University of Milan (Italy), one graduating while the other just graduated. The observers had no previous experience with dairy cows, and they received specific training to score a set of sample photos.

Agreement measures

A rough measure of the reliability is the concordance rate (P_0), which is given by the ratio between the sum of the concordant cases and the total number of observations (Bajpai et al. 2015). P_0 is expressed as a percentage, and it is implemented creating an agreement matrix, where the rows and columns represent the total marginal distributions, obtained summing the frequencies of the scores assigned by each observer to the variable of interest during the IOR evaluation (McHugh 2012). However, this measure does not consider the chance agreement (P_e). For this reason, to obtain a proper IOR estimation, the use of agreement indices, which also consider the P_e, is mandatory.

A summary of the most documented agreement indices for trichotomous and four-level animal-based welfare indicators in the case of the evaluation performed by two observers is reported in Table 1. In particular, to evaluate the IOR between two observers for trichotomous indicators, the most documented agreement indices in the literature are: Scott's π (Scott 1955), Cohen's K (Cohen 1960), Cohen's K_C (Cohen 1960), Holley and Guilford's G (Holley and Guilford 1964), Cohen's *weighted K* (K^*) (Cohen 1968), Krippendorff's α (Krippendorff 1970), Hubert's Γ (Hubert 1977a), Janson and Vegelius' J (Janson and Vegelius 1978), Bangdiwala's B (Bangdiwala 1985), Andrés and Marzo's Δ (Andrés and Marzo 2004), Quatto's S (Quatto 2004), Gwet's $\gamma(AC_1)$ (Gwet 2008) and Quatto's *weighted S* (S^*) (Marasini et al. 2016). Holsti's H (Holsti 1969), even if suitable to assess the IOR of trichotomous variables in the presence of two observers, was not considered in the current study, as this index does not consider the P_e and, therefore, we considered it unable to confer reliable agreement results (Giammarino et al. 2021). To evaluate the IOR between two observers for four-level indicators, the most documented agreement indices in the literature are: Cohen's K, Cohen's K_C, Holley and Guilford's G, Krippendorff's α, Quatto's S and Gwet's $\gamma(AC_1)$. An exhaustive explanation of each of the abovementioned indices, as well as their closed formulas of variance estimates, are reported by Giammarino et al. (2021). However, in the current paper, some modifications and implementations were adopted, as detailed in Appendix A and briefly summarised here below. In particular, the formula for Janson and Vegelius' J is calculated differently from what reported in Giammarino et al. (2021) as, for variables characterised by more than two categories, the development of the formula for this index changes (Janson and Vegelius 1982). Moreover, the closed formulas for Cohen's *weighted K* [not considered in Giammarino et al. (2021), as this index can be implemented in the presence of ordinal variables, but not in the presence of categorical variables] and for Andrés and Marzo's Δ are not reported in Appendix A, as they were too complex to be implemented manually for trichotomous variables (their implementation was possible in R software, only). Finally, the closed formula for Quatto's *weighted S* is included in Appendix A, as this index can be adopted to evaluate the IOR for ordinal variables only, and therefore it was not considered by Giammarino et al. (2021).

Confidence intervals for agreement indices

For a proper estimation of the agreement between observers, the calculation of confidence intervals

Table 1. Agreement indices implemented for each animal-based welfare indicator.

Variable	Agreement index	References for the agreement index	Confidence intervals	References for the confidence intervals	R packages[2]	R functions[2]
BODY CONDITION SCORE and KNEE CALLUSES (trichotomous animal-based welfare indicators)	*Scott's π*	Scott (1955)	Closed formula of variance Bootstrap	Scott (1955)	library(boot)	boot_result var(boot) boot.ci

	Cohen's K	Cohen (1960)	Closed formula of variance
			Bootstrap
				Altman (2000)	library(irr)
					library(agrmt)
					library(raters)
					library(vcd)
					library(boot)
						boot_result var(boot)
						boot.ci
						confint(res.k)

	Cohen's Kc	Cohen (1960)	Closed formula of variance Bootstrap	Altman (2000)	library(boot)	boot_result var(boot) boot.ci
	Holley and Guilford's G	Holley and Guilford (1964)	Closed formula of variance Bootstrap	Gwet (2001)	library(boot)	boot_result var(boot) boot.ci

	Cohen's weighted K	Cohen (1968)	Bootstrap[a]	Cohen (1968)	library(vcd)
					library(boot)
						boot_result
						var(boot)
						boot.ci
						confint(res.k)

	Krippendorff's α	Krippendorff (1970)	Closed formula of variance
			Bootstrap
				Altman (2000)	library(irr)
					library(boot)
						boot_result
						var(boot)
						boot.ci
						kripp.alpha

	Hubert's Γ	Hubert (1977a)	Closed formula of variance Bootstrap	Janson and Vegelius (1982)	library(boot)	boot_result var(boot) boot.ci
	Janson and Vegelius' J	Janson and Vegelius (1978)	Closed formula of variance Bootstrap	Janson and Vegelius (1982)	library(boot)	boot_result var(boot) boot.ci

	Bangdiwala's B	Bangdiwala (1985)	Bootstrap[a]	Bangdiwala (1985)	library(boot)
					library(vcd)
						boot_result
						var(boot)
						boot.ci
						agreementplot(xtab)

	Andrés and Marzo's Δ	Andrés and Marzo (2004)	Bootstrap[a]	Andrés and Marzo (2004)	library(boot) library (DeltaMAN)	boot_result var(boot) boot.ci Delta

	Quatto's S	Quatto (2004)	Closed formula of variance
			Bootstrap
				Quatto (2004)	library(irr)
					library(agrmt)
					library(raters)
					library(boot)
						boot_result
						var(boot)
						boot.ci
						concordance

	Gwet's γ(AC$_1$)	Gwet (2008)	Closed formula of variance
			Bootstrap
				Gwet (2008)	library(boot)
					library(irrCAC)
						boot_result
						var(boot)
						boot.ci
						gwet.ac1

	Quatto's weighted S	Marasini et al. (2016)	Bootstrap[a]	Marasini et al. (2016)	library(irr)
					library(agrmt)
					library(raters)
					library(boot)
						boot_result
						var(boot)
						boot.ci
						wlin.conc

EAR POSTURE AND EYE WHITE (four-level animal-based welfare indicators)	*Cohen's K*	Cohen (1960)	Closed formula of variance
			Bootstrap
				Altman (2000)	library(irr)
					library(agrmt)
					library(raters)
					library(vcd)
					library(boot)
						boot_result
						var(boot)
						boot.ci
						confint(res.k)

	Cohen's Kc	Cohen (1960)	Closed formula of variance Bootstrap	Altman (2000)	library(boot)	boot_result var(boot) boot.ci
	Holley and Guilford's G	Holley and Guilford (1964)	Closed formula of variance Bootstrap	Gwet (2001)	library(boot)	boot_result var(boot) boot.ci

	Krippendorff's α	Krippendorff (1970)	Closed formula of variance
			Bootstrap
				Altman (2000)	library(irr)
					library(boot)
						boot_result
						var(boot)
						boot.ci
						kripp.alpha

	Quatto's S	Quatto (2004)	Closed formula of variance
			Bootstrap
				Quatto (2004)	library(irr)
					library(agrmt)
					library(raters)
					library(boot)
						boot_result
						var(boot)
						boot.ci
						concordance

	Gwet's γ(AC$_1$)	Gwet (2008)	Closed formula of variance
			Bootstrap
				Gwet (2008)	library(boot)
					library(irrCAC)
						boot_result
						var(boot)
						boot.ci
						gwet.ac1

[a] Bootstrap: for Cohen's *weighted K*, Bangdiwala's *B*, Andrès and Marzo's Δ, and Quatto's *weighted S* the formula of variance is too complex to be implemented manually. Consequently, for the above-mentioned indices, the confidence intervals were calculated through the Bootstrap Method, only.

(inference on the estimated parameter) for each index is recommended. To create the confidence intervals, it is necessary to calculate the variance for each index, which gives information about the variability of the values assumed by the index itself.

For all the agreement indices, the variance estimates and the confidence intervals were implemented by the Bootstrap Method, which is a resampling technique that guarantees reliable confidence intervals (DiCiccio and Efron 1996). At this regard, one of the most useful and easiest method is the Bootstrap *t*-Method proposed by Efron (1979), which is a generalisation of the Student's *t*-Method.

When it was possible, adopting a 95% confidence limit and 1.96 as a constant (that is, for Scott's π, Cohen's *K*, Cohen's K_C, Holley and Guilford's *G*, Krippendorff's α, Hubert's Γ, Janson and Vegelius' *J*, Quatto's *S* and Gwet's $\gamma(AC_1)$), confidence intervals were also implemented using closed formulas of variance estimates. An exhaustive explanation of the closed formulas used in the current paper for the calculation of the variance for each of the implemented agreement indices is reported by Giammarino et al. (2021). However, in the current paper, some modifications and implementations were adopted, as detailed in Appendix B and briefly summarised here below. In particular, the closed formulas of variance estimates for Cohen's *weighted K*, Bangdiwala's *B,* Andrés and Marzo's Δ and Quatto's *weighted S* were not included in Appendix B, as they were too complex to be implemented manually. The same difficulty was already reported by Giammarino et al. (2021) when considering the manual calculation of the variance estimates for Bangdiwala's *B* in the case of dichotomous animal-based welfare indicators; on the contrary, for Andrés and Marzo's Δ the variance estimates were easier to be calculated manually in the case of dichotomous rather than trichotomous indicators (Andrés and Marzo 2004). When the closed formulas of variance estimates were too complex to be implemented manually (i.e. for Cohen's *weighted K*, Bangdiwala's *B*, Andrés and Marzo's Δ and Quatto's *weighted S*), confidence intervals were calculated using the Bootstrap Method, only.

For some agreement indices (i.e. for Cohen's *K*, Cohen's *weighted K*, Quatto's *S*, Gwet's $\gamma(AC_1)$, and Quatto's *weighted S*) specific functions are available in R Commander that allow the confidence intervals to be easily calculated (Table 1). Therefore, for the above-mentioned agreement indices, the confidence intervals were also calculated using R functions.

Statistical analyses

Both Microsoft Excel (2019) and R Commander (*version R × 64 4.2.2*) were used to calculate the values of the agreement indices and their respective confidence intervals. Due to the complexity in calculating the agreement values using closed formulas in Microsoft Excel, Cohen's *weighted K* and Andrés and Marzo's Δ were implemented in R Commander, only. For the same reason, the confidence intervals for Cohen's *weighted K*, Bangdiwala's *B*, Andrés and Marzo's Δ and Quatto's *weighted S* were implemented in R Commander, only. Moreover, in R Commander the Bootstrap Method was developed to calculate the values of all the agreement indices and their respective confidence intervals, implementing scripts specifically created for each index. Specific packages and R functions were also used to calculate the values only (i.e. Krippendorff's α, Bangdiwala's *B* and Andrés and Marzo's Δ), or both the values and their respective confidence intervals (i.e. Cohen's *K*, Cohen's *weighted K*, Quatto's *S*, Gwet's $\gamma(AC_1)$ and Quatto's *weighted S*) of some agreement indices. A summary of all the R packages and functions implemented for each agreement index is reported in Table 1.

In R software, for Bangdiwala's *B* the agreement chart was also created using specific packages and functions, which are reported in Table 1. Indeed, the *B*-statistic proposed by Bangdiwala (1985) derives from a graphical representation, which easily identifies the level of agreement between two observers (Munoz and Bangdiwala 1997; Bangdiwala et al. 2008). In particular, the agreement chart allows the reader for an immediate visual evaluation of the agreement between observers, which could result easier when compared to the implementation and subsequent interpretation of the *B* index.

Results

Trichotomous animal-based welfare indicators

Agreement measures for Body condition score and Knee calluses

The values of the agreement indices obtained for BCS and KNC are reported in Tables 2 and 3, respectively. The concordance rate (P_0) was the same for all the considered indices, except for Cohen's *weighted K* and Quatto's *weighted S*; in the latter cases, the concordance rate (P_0*) showed higher values when compared to P_0.

In some cases [i.e. for BCS: ITF3, ITF5, ITF7, PTF4, AP1, and AP2 (Table 2); for KNC: ITF2 and PTF2 (Table 3)], Scott's π, Cohen's *K*, Cohen's K_C, Cohen's *weighted*

K and Krippendorff's α showed very low agreement values when compared to the obtained P_0 and P_0*. The same indices even resulted in null or negative values in some cases [i.e. for BCS: ITF4, PTF3, and AP3 (Table 2); for KNC: ITF4, ITF5, ITF6, ITF7, PTF5, and PTF7 (Table 3)]. When P_0 was equal to 100% [i.e. for KNC: ITF3 and PTF4 (Table 3)], the above-mentioned agreement indices were not computable. Moreover, in some cases [i.e. for BCS: ITF1, PTF1, PTF6, and PTF7 (Table 2); for KNC: PTF1, PTF3, and PTF6 (Table 3)], Cohen's K_C exceeded the P_0 values.

Except for BCS in ITF7, in all the cases in which P_0 was $\leq$ 95% Andrés and Marzo's Δ showed higher agreement values than Hubert's Γ. Andrés and Marzo's Δ was not computable when the P_0 was equal to 100% [i.e. for KNC: ITF3 and PTF4 (Table 3)]. Analysing the cases in which Scott's π, Cohen's *K*, Cohen's K_C, Cohen's *weighted K* and Krippendorff's α conferred very low agreement results if compared to their respective P_0 and P_0*, it can be seen that Andrés and Marzo's Δ, Hubert's Γ and Janson and Vegelius' *J* were able to give higher agreement results (Tables 2 and 3). However, in all the cases, Hubert's Γ, Andrès and Marzo's Δ and Janson and Vegelius' *J* conferred agreement results that were lower and further from P_0 when compared to those obtained implementing Bangdiwala's *B*, Gwet's $\gamma(AC_1)$, Quatto's *weighted S*, Holley and Guilford's *G* and Quatto's *S*.

Bangdiwala's *B*, Gwet's $\gamma(AC_1)$ and Quatto's *weighted S* resulted in very similar values each other; such values were very close to the obtained P_0 and P_0*. In all cases, values for Holley and Guilford's *G* and Quatto's *S* were identical.

Confidence intervals for Body condition score and Knee calluses

The values of the confidence intervals obtained for the trichotomous indicators and implemented in Microsoft Excel using the closed formulas of the variance estimates and in R Commander using the Bootstrap Method and specific R functions are reported in Table 4 (BCS) and Table 5 (KNC). In most of the cases, we observed a substantial agreement between the confidence intervals obtained using the closed formulas of variance and those obtained using the Bootstrap Method. However, the closed formulas are built on an approximate calculation of the variance (DiCiccio and Efron 1996) and are sometimes difficult to be implemented manually [i.e. in the case of Cohen's *weighted K*, Bangdiwala's *B*, Andrés and Marzo's Δ and Quatto's *weighted S* for both BCS (Table 4) and KNC (Table 5)]. Moreover, in some cases [i.e. for BCS: the formula for Cohen's K_C in ITF4 and PTF3 (Table 4); for KNC: the formula for Cohen's K_C in ITF3, ITF4, ITF5, ITF6, ITF7, PTF4, PTF5, PTF7 (Table 5); for KNC: the formulas for Scott's π, Cohen's *K*, and Krippendorff's α in ITF3 and PTF4 (Table 5)] the closed formulas were not able to give any number. On the contrary, the Bootstrap Method allowed to calculate the confidence intervals for all the considered

Table 2. Values of the concordance rate and of the agreement indices obtained for Body condition score (BCS) for the three alpine pastures and for the fourteen intensively managed Italian and Portuguese dairy goat farms.

				Agreement Index											
	n	P_0	P_0*	π	K	K_C	G	K*	α	Γ	J	B	Δ	S	$\gamma(AC_1)$	S*
ITF1	49	84%	91%	0.53	0.55	0.91	0.76	0.51	0.54	0.44	0.64	0.81	0.71	0.76	0.80	0.79
ITF2	37	78%	88%	0.43	0.43	0.46	0.68	0.40	0.44	0.30	0.53	0.75	0.60	0.68	0.73	0.73
ITF3	43	74%	87%	0.05	0.09	0.20	0.62	0.12	0.06	0.22	0.45	0.72	0.34	0.62	0.70	0.71
ITF4	30	90%	95%	−0.04	0	N.C.	0.85	0	−0.02	0.63	0.78	0.90	0.74	0.85	0.90	0.89
ITF5	30	83%	92%	0.19	0.21	0.40	0.75	0.21	0.20	0.43	0.63	0.81	0.63	0.75	0.81	0.81
ITF6	34	82%	90%	0.68	0.68	0.72	0.74	0.68	0.69	0.40	0.58	0.73	0.71	0.74	0.76	0.77
ITF7	39	74%	87%	0.26	0.27	0.43	0.62	0.32	0.27	0.22	0.44	0.69	0.17	0.62	0.69	0.71
PTF1	48	98%	99%	0.95	0.95	1	0.97	0.96	0.96	0.92	0.93	0.96	0.88	0.97	0.97	0.98
PTF2	38	92%	96%	0.72	0.72	0.80	0.88	0.72	0.73	0.70	0.79	0.89	0.78	0.88	0.91	0.91
PTF3	25	96%	98%	−0.02	0	N.C.	0.94	0	0	0.84	0.91	0.96	0.83	0.94	0.96	0.96
PTF4	39	82%	91%	0.35	0.35	0.39	0.73	0.35	0.36	0.40	0.58	0.77	0.60	0.73	0.79	0.80
PTF5	32	88%	94%	0.59	0.59	0.74	0.81	0.59	0.60	0.55	0.68	0.83	0.70	0.81	0.85	0.86
PTF6	38	97%	99%	0.91	0.91	1	0.96	0.92	0.91	0.89	0.93	0.97	0.86	0.96	0.97	0.97
PTF7	35	97%	99%	0.89	0.89	1	0.96	0.89	0.89	0.89	0.92	0.96	0.87	0.96	0.97	0.97
AP1	44	82%	91%	0.26	0.26	0.36	0.73	0.29	0.27	0.39	0.59	0.79	0.50	0.73	0.79	0.80
AP2	70	73%	86%	0.04	0.05	0.08	0.59	0.07	0.05	0.20	0.42	0.69	0.22	0.59	0.68	0.69
AP3	46	74%	87%	−0.12	−0.11	−0.15	0.61	−0.08	−0.11	0.21	0.46	0.71	0.27	0.61	0.70	0.71

n = sample size; P_0 = concordance rate for all the indices, except for K* (Cohen's *weighted K*) and S* (Quatto's *weighted S*); P_0* = concordance rate for Cohen's *weighted K* and Quatto's *weighted S*; π = Scott's π; K = Cohen's K; K_C = Cohen's K_C; G = Holley and Guilford's G; K* = Cohen's *weighted K*; α = Krippendorff's α; Γ = Hubert's Γ; J = Janson and Vegelius' J; B = Bangdiwala's B; Δ = Andrés and Marzo's Δ; S = Quatto's S; $\gamma(AC_1)$ = Gwet's $\gamma(AC_1)$; S* = Quatto's *weighted S*; ITF1 = Italian farm 1; ITF2 = Italian farm 2; ITF3 = Italian farm 3; ITF4 = Italian farm 4; ITF5 = Italian farm 5; ITF6 = Italian farm 6; ITF7 = Italian farm 7; PTF1 = Portuguese farm 1; PTF2 = Portuguese farm 2; PTF3 = Portuguese farm 3; PTF4 = Portuguese farm 4; PTF5 = Portuguese farm 5; PTF6 = Portuguese farm 6; PTF7 = Portuguese farm 7; AP1 = alpine pasture 1; AP2 = alpine pasture 2; AP3 = alpine pasture 3; N.C. = not computable.

Table 3. Values of the concordance rate and of the agreement indices obtained for knee calluses (KNC) for the fourteen intensively managed Italian and Portuguese dairy goat farms.

				Agreement Index												
	n	P_0	P_0*	π	*K*	K_C	*G*	*K**	α	Γ	*J*	*B*	Δ	*S*	$\gamma(AC_1)$	*S**
ITF1	49	94%	97%	0.63	0.63	0.72	0.91	0.63	0.64	0.77	0.85	0.93	0.83	0.91	0.93	0.93
ITF2	37	78%	89%	0.09	0.12	0.26	0.68	0.13	0.10	0.30	0.54	0.76	0.49	0.68	0.75	0.76
ITF3	43	100%	100%	N.C.	N.C.	N.C.	1	N.C.	N.C.	1	1	1	N.C.	1	1	1
ITF4	30	97%	98%	−0.02	0	N.C.	0.95	0	0	0.87	0.93	0.97	0.85	0.95	0.97	0.96
ITF5	30	77%	88%	−0.10	0	N.C.	0.65	0	−0.08	0.26	0.51	0.77	0.57	0.65	0.74	0.74
ITF6	34	85%	93%	−0.08	0	N.C.	0.78	0	−0.06	0.48	0.69	0.85	0.70	0.78	0.84	0.83
ITF7	39	92%	96%	−0.04	0	N.C.	0.88	0	−0.03	0.71	0.83	0.92	0.81	0.88	0.92	0.91
PTF1	48	88%	94%	0.64	0.65	1	0.81	0.67	0.65	0.55	0.70	0.85	0.75	0.81	0.85	0.86
PTF2	38	82%	91%	0.29	0.29	0.42	0.72	0.32	0.30	0.38	0.59	0.79	0.55	0.72	0.79	0.79
PTF3	25	92%	96%	0.63	0.63	1	0.88	0.64	0.64	0.69	0.81	0.91	0.75	0.88	0.91	0.91
PTF4	39	100%	100%	N.C.	N.C.	N.C.	1	N.C.	N.C.	1	1	1	N.C.	1	1	1
PTF5	32	97%	98%	−0.02	0	N.C.	0.95	0	0	0.88	0.93	0.97	0.86	0.95	0.97	0.96
PTF6	38	95%	97%	0.88	0.88	1	0.92	0.89	0.88	0.80	0.85	0.92	0.82	0.92	0.93	0.94
PTF7	35	94%	97%	−0.02	0	N.C.	0.91	0	−0.01	0.78	0.87	0.94	0.81	0.91	0.94	0.94

n = sample size; P_0 = concordance rate for all the indices, except for *K** (Cohen's *weighted K*) and *S** (Quatto's *weighted S*); P_0* = concordance rate for Cohen's *weighted K* and Quatto's *weighted S*; π = Scott's π; *K* = Cohen's *K*; K_C = Cohen's K_C; *G* = Holley and Guilford's *G*; *K** = Cohen's *weighted K*; α = Krippendorff's α; Γ = Hubert's Γ; *J* = Janson and Vegelius' *J*; *B* = Bangdiwala's *B*; Δ = Andrés and Marzo's Δ; *S* = Quatto's *S*; $\gamma(AC_1)$ = Gwet's $\gamma(AC_1)$; *S** = Quatto's *weighted S*; ITF1 = Italian farm 1; ITF2 = Italian farm 2; ITF3 = Italian farm 3; ITF4 = Italian farm 4; ITF5 = Italian farm 5; ITF6 = Italian farm 6; ITF7 = Italian farm 7; PTF1 = Portuguese farm 1; PTF2 = Portuguese farm 2; PTF3 = Portuguese farm 3; PTF4 = Portuguese farm 4; PTF5 = Portuguese farm 5; PTF6 = Portuguese farm 6; PTF7 = Portuguese farm 7; N.C. = not computable.

agreement indices (with very few exceptions), conferring more accurate results (DiCiccio and Efron 1996).

For some agreement indices (i.e. Cohen's *K*, Cohen's *weighted K*, Quatto's *S*, Gwet's $\gamma(AC_1)$ and Quattos' *weighted S*), R functions are available for confidence intervals calculation. In all the cases, the confidence intervals obtained using R functions were close (and in some cases identical) to those obtained using the Bootstrap Method. Indeed, the R functions 'concordance' and 'wlin.conc' were developed to calculate the confidence intervals for Quatto's *S* and Quatto's *weighted S* starting from the Bootstrap Method.

Considering the above-mentioned issues, we decided to rely on the Bootstrap Method to describe the differences in the results obtained for confidence intervals among the considered agreement indices. For all the farms and alpine pastures, the confidence intervals obtained for Holley and Guilford's *G* and Quatto's *S* were identical. Furthermore, in all the considered cases, Bangdiwala's *B*, Gwet's $\gamma(AC_1)$ and Quatto's *weighted S* showed the narrowest confidence intervals, followed by Quatto's *S* (Tables 4 and 5).

Considering BCS, the confidence intervals obtained for Scott's π, Cohen's *K*, Cohen's K_C, Cohen's *weighted K* and Krippendorff's α were wide, with few exceptions recorded (i.e. in ITF4, PTF1, PTF3, PTF6, and AP3, in which negative values were also found). Wide confidence intervals were also often found for Andrés and Marzo's Δ. Janson and Vegelius' *J* and Hubert's Γ were characterised by confidence intervals with similar width, except in AP1.

The confidence intervals results obtained for KNC showed the same trend as that observed for BCS.

Finally, even using the Bootstrap Method, in few cases the confidence intervals calculated for Scott's π, Cohen's *K*, Cohen's *weighted K*, Krippendorff's α [i.e. for KNC: ITF3, PTF4 (Table 5)], Cohen K_C [i.e. for BCS: ITF4, PTF3 (Table 4); for KNC: ITF3, ITF4, ITF5, ITF6, ITF7, PTF4, PTF5, PTF7 (Table 5)], and Andrés and Marzo's Δ [i.e. for KNC: ITF3, PTF4 (Table 5)] did not return any number.

Bangdiwala's agreement chart for Body condition score

To provide examples of Bangdiwala's agreement charts, three cases were considered. The charts were developed for the BCS recorded in the three alpine pastures, and are shown in Appendix C. Within the chart, the agreement is defined as the proportion between the black areas inside the chart and the remaining part of the matrix, which is represented by the total marginal distributions of the rows and columns.

Four-level animal-based welfare indicators

Agreement measures for Ear posture and Eye white

Differently from what was observed in the case of the considered trichotomous indicators (BCS; KNC), in all the cattle farms, Cohen's *K* and Krippendorff's α showed agreement values not far from P_0 for both EP and EW (Table 6). On the other hand, in some circumstances, Cohen's K_C coincided with P_0 (EW-F3) or, as already observed for BCS and KNC, even exceeded P_0 (EP-F2, EP-F3, EW-F2), therefore showing anomalous values. As for BCS and KNC, also for the four-level indicators (i) Quatto's *S* and Holley and Guilford's *G*

Table 4. Values of the confidence intervals for the agreement indices obtained for Body condition score (BCS) implemented using closed formulas, Bootstrap-*t* Method, and R functions in the three alpine pastures and in the fourteen intensively managed Italian and Portuguese dairy goat farms.

		Confidence intervals		
Trichotomous variables (BCS)	Agreement indices	By closed formulas	By Bootstrap-*t* Method	By R functions
ITF1	π	0.24; 0.83	0.26; 0.84	N.A.
($n = 49$)	K	0.26; 0.83	0.28; 0.83	confint(res.k): 0.29; 0.80
	K_C	0.63; 1.18	0.71; 1.11	N.A.
	G	0.55; 0.96	0.60; 0.91	N.A.
	K^*	N.C.	0.24; 0.79	confint(res.k): 0.24; 0.79
	α	0.25; 0.83	0.28; 0.84	N.A.
	Γ	0.39; 0.50	0.13; 0.72	N.A.
	J	0.58; 0.70	0.41; 0.85	N.A.
	B	N.C.	0.69; 0.94	N.A.
	Δ	N.C.	0.49; 1.02	N.A.
	S	0.56; 0.95	0.60; 0.91	concordance: 0.60; 0.88
	$\gamma(AC_1)$	0.67; 0.93	0.67; 0.94	gwet.ac1: 0.66; 0.94
	S^*	N.C.	0.66; 0.93	wlin.conc: 0.66; 0.91
ITF2	π	0.07; 0.78	0.10; 0.79	N.A.
($n = 37$)	K	0.08; 0.78	0.12; 0.76	confint(res.k): 0.12; 0.74
	K_C	0.32; 0.60	−0.01; 0.78	N.A.
	G	0.41; 0.94	0.48; 0.88	N.A.
	K^*	N.C.	0.08; 0.74	confint(res.k): 0.07; 0.73
	α	0.22; 0.66	0.14; 0.76	N.A.
	Γ	0.23; 0.38	0.00; 0.59	N.A.
	J	0.46; 0.60	0.25; 0.80	N.A.
	B	N.C.	0.58; 0.91	N.A.
	Δ	N.C.	0.29; 0.97	N.A.
	S	0.45; 0.90	0.48; 0.88	concordance: 0.47; 0.88
	$\gamma(AC_1)$	0.55; 0.91	0.56; 0.91	gwet.ac1: 0.55; 0.92
	S^*	N.C.	0.55; 0.90	wlin.conc: 0.51; 0.88
ITF3	π	−0.45; 0.54	−0.23; 0.33	N.A.
($n = 43$)	K	−0.38; 0.55	−0.16; 0.35	confint(res.k): −0.17; 0.34
	K_C	−0.19; 0.60	−0.54; 0.92	N.A.
	G	0.36; 0.88	0.43; 0.80	N.A.
	K^*	N.C.	−0.14; 0.37	confint(res.k): −0.14; 0.37
	α	−0.43; 0.55	−0.21; 0.34	N.A.
	Γ	0.16; 0.28	−0.07; 0.46	N.A.
	J	0.39; 0.51	0.20; 0.68	N.A.
	B	N.C.	0.57; 0.87	N.A.
	Δ	N.C.	−0.16; 0.80	N.A.
	S	0.40; 0.83	0.43; 0.80	concordance: 0.41; 0.83
	$\gamma(AC_1)$	0.53; 0.87	0.54; 0.88	gwet.ac1: 0.53; 0.88
	S^*	N.C.	0.57; 0.85	wlin.conc: 0.56; 0.84
ITF4	π	−1.18; 1.10	−0.08; 0.01	N.A.
($n = 30$)	K	−1.07; 1.07	0.00; 0.00	confint(res.k): 0.00; 0.00
	K_C	N.A.N.	N.A.N.	N.A.
	G	0.64; 1.06	0.69; 1.00	N.A.
	K^*	N.C.	0.00; 0.00	confint(res.k): 0.00; 0.00
	α	−1.09; 1.05	−0.48; 0.35	N.A.
	Γ	0.54; 0.72	0.27; 0.96	N.A.
	J	0.69; 0.87	0.54; 1.01	N.A.
	B	N.C.	0.79; 1.00	N.A.
	Δ	N.C.	0.59; 0.88	N.A.
	S	0.60; 1.10	0.69; 1.00	concordance: 0.65; 1.00
	$\gamma(AC_1)$	0.78; 1.02	0.78; 1.02	gwet.ac1: 0.77; 1.00
	S^*	N.C.	0.77; 1.01	wlin.conc: 0.78; 1.00
ITF5	π	−0.47; 0.85	−0.27; 0.69	N.A.
($n = 30$)	K	−0.42; 0.84	−0.23; 0.68	confint(res.k): −0.24; 0.66
	K_C	−0.11; 0.91	−0.51; 1.30	N.A.
	G	0.48; 1.02	0.55; 0.95	N.A.
	K^*	N.C.	−0.23; 0.69	confint(res.k): −0.24; 0.66
	α	−0.36; 0.76	−0.23; 0.70	N.A.
	Γ	0.33; 0.52	0.05; 0.76	N.A.
	J	0.54; 0.72	0.37; 0.87	N.A.
	B	N.C.	0.66; 0.97	N.A.
	Δ	N.C.	0.39; 0.84	N.A.
	S	0.50; 1.00	0.55; 0.95	concordance: 0.55; 0.95
	$\gamma(AC_1)$	0.65; 0.97	0.66; 0.99	gwet.ac1: 0.65; 0.98
	S^*	N.C.	0.67; 0.96	wlin.conc: 0.66; 0.96
ITF6	π	0.45; 0.92	0.46; 0.93	N.A.
($n = 34$)	K	0.45; 0.91	0.45; 0.92	confint(res.k): 0.46; 0.91

(*continued*)

Table 4. Continued.

		Confidence intervals		
Trichotomous variables (BCS)	Agreement indices	By closed formulas	By Bootstrap-*t* Method	By R functions
	K_C	0.62; 0.82	0.42; 0.90	N.A.
	G	0.48; 0.99	0.55; 0.93	N.A.
	K^*	N.C.	0.44; 0.93	confint(res.k): 0.43; 0.92
	α	0.46; 0.92	0.48; 0.92	N.A.
	Γ	0.32; 0.48	0.06; 0.71	N.A.
	J	0.50; 0.66	0.29; 0.83	N.A.
	B	N.C.	0.53; 0.92	N.A.
	Δ	N.C.	0.51; 1.04	N.A.
	S	0.50; 0.97	0.55; 0.93	concordance: 0.56; 0.91
	$\gamma(AC_1)$	0.58; 0.94	0.58; 0.93	gwet.ac1: 0.56; 0.95
	S^*	N.C.	0.59; 0.95	wlin.conc: 0.57; 0.93
ITF7	π	−0.14; 0.66	−0.05; 0.61	N.A.
($n = 39$)	K	−0.11; 0.66	−0.03; 0.60	confint(res.k): −0.05; 0.59
	K_C	0.13; 0.73	−0.08; 0.97	N.A.
	G	0.34; 0.89	0.40; 0.83	N.A.
	K^*	N.C.	0.04; 0.63	confint(res.k): 0.03; 0.62
	α	−0.12; 0.66	−0.05; 0.64	N.A.
	Γ	0.15; 0.29	−0.08; 0.47	N.A.
	J	0.37; 0.51	0.20; 0.66	N.A.
	B	N.C.	0.52; 0.86	N.A.
	Δ	N.C.	−0.50; 0.80	N.A.
	S	0.39; 0.84	0.40; 0.83	concordance: 0.38; 0.81
	$\gamma(AC_1)$	0.51; 0.87	0.51; 0.87	gwet.ac1: 0.50; 0.88
	S^*	N.C.	0.56; 0.87	wlin.conc: 0.54; 0.86
PTF1	π	0.87; 1.04	0.86; 1.04	N.A.
($n = 48$)	K	0.87; 1.04	0.86; 1.04	confint(res.k): 0.87; 1.00
	K_C	0.91; 1.09	1.00; 1.00	N.A.
	G	0.89; 1.05	0.91; 1.03	N.A.
	K^*	N.C.	0.87; 1.04	confint(res.k): 0.87; 1.00
	α	0.86; 1.06	0.87; 1.04	N.A.
	Γ	0.86; 0.97	0.76; 1.07	N.A.
	J	0.87; 0.99	0.81; 1.06	N.A.
	B	N.C.	0.90; 1.03	N.A.
	Δ	N.C.	0.82; 0.94	N.A.
	S	0.77; 1.17	0.91; 1.03	concordance: 0.91; 1.00
	$\gamma(AC_1)$	0.92; 1.02	0.92; 1.02	gwet.ac1: 0.92; 1.00
	S^*	N.C.	0.93; 1.02	wlin.conc: 0.93; 1.00
PTF2	π	0.42; 1.03	0.41; 1.10	N.A.
($n = 38$)	K	0.42; 1.02	0.41; 1.07	confint(res.k): 0.43; 1.00
	K_C	0.62; 0.97	0.41; 1.09	N.A.
	G	0.71; 1.05	0.76; 1.01	N.A.
	K^*	N.C.	0.41; 1.07	confint(res.k): 0.43; 1.00
	α	0.43; 1.03	0.41; 1.07	N.A.
	Γ	0.63; 0.77	0.41; 0.99	N.A.
	J	0.72; 0.86	0.57; 1.01	N.A.
	B	N.C.	0.78; 1.02	N.A.
	Δ	N.C.	0.62; 0.94	N.A.
	S	0.66; 1.11	0.76; 1.01	concordance: 0.76; 1.00
	$\gamma(AC_1)$	0.81; 1.01	0.81; 1.01	gwet.ac1: 0.80; 1.00
	S^*	N.C.	0.81; 1.01	wlin.conc: 0.79; 1.00
PTF3	π	−2.02; 1.98	−0.04; 0.03	N.A.
($n = 25$)	K	−1.92; 1.92	0.00; 0.00	confint(res.k): 0.00; 0.00
	K_C	N.A.N.	N.A.N.	N.A.
	G	0.79; 1.09	0.82; 1.07	N.A.
	K^*	N.C.	0.00; 0.00	confint(res.k): 0.00; 0.00
	α	−1.54; 1.54	−1.34; 0.59	N.A.
	Γ	0.73; 0.95	0.55; 1.10	N.A.
	J	0.80; 1.02	0.75; 1.07	N.A.
	B	N.C.	0.88; 1.04	N.A.
	Δ	N.C.	0.77; 0.92	N.A.
	S	0.66; 1.22	0.82; 1.07	concordance: 0.82; 1.00
	$\gamma(AC_1)$	0.88; 1.04	0.88; 1.04	gwet.ac1: 0.87; 1.00
	S^*	N.C.	0.87; 1.04	wlin.conc: 0.87; 1.00
PTF4	π	−0.09; 0.79	−0.03; 0.79	N.A.
($n = 39$)	K	−0.08; 0.79	−0.02; 0.76	confint(res.k): −0.03; 0.73
	K_C	0.21; 0.57	−0.14; 0.86	N.A.
	G	0.49; 0.97	0.55; 0.91	N.A.
	K^*	N.C.	−0.01; 0.75	confint(res.k): −0.03; 0.73
	α	−0.07; 0.79	−0.03; 0.79	N.A.
	Γ	0.32; 0.47	0.07; 0.67	N.A.

(*continued*)

Table 4. Continued.

		Confidence intervals		
Trichotomous variables (BCS)	Agreement indices	By closed formulas	By Bootstrap-*t* Method	By R functions
	J	0.51; 0.65	0.33; 0.81	N.A.
	B	N.C.	0.61; 0.94	N.A.
	Δ	N.C.	0.36; 0.81	N.A.
	S	0.51; 0.95	0.55; 0.91	concordance: 0.54; 0.88
	$\gamma(AC_1)$	0.64; 0.94	0.65; 0.94	gwet.ac1: 0.64; 0.95
	S^*	N.C.	0.66; 0.94	wlin.conc: 0.65; 0.91
PTF5	π	0.21; 0.97	0.25; 0.99	N.A.
($n = 32$)	K	0.22; 0.97	0.23; 1.00	confint(res.k): 0.24; 0.95
	K_C	0.47; 1.02	0.26; 1.21	N.A.
	G	0.58; 1.04	0.65; 0.98	N.A.
	K^*	N.C.	0.23; 0.99	confint(res.k): 0.24; 0.95
	α	0.26; 0.94	0.24; 0.99	N.A.
	Γ	0.46; 0.64	0.18; 0.87	N.A.
	J	0.59; 0.77	0.42; 0.94	N.A.
	B	N.C.	0.67; 1.00	N.A.
	Δ	N.C.	0.49; 0.87	N.A.
	S	0.57; 1.06	0.65; 0.98	concordance: 0.63; 0.95
	$\gamma(AC_1)$	0.71; 0.99	0.72; 0.99	gwet.ac1: 0.70; 1.00
	S^*	N.C.	0.73; 0.98	wlin.conc: 0.72; 0.96
PTF6	π	0.74; 1.08	0.71; 1.13	N.A.
($n = 38$)	K	0.74; 1.08	0.71; 1.13	confint(res.k): 0.74; 1.00
	K_C	0.83; 1.17	1.00; 1.00	N.A.
	G	0.86; 1.06	0.88; 1.04	N.A.
	K^*	N.C.	0.80; 1.11	confint(res.k): 0.75; 1.00
	α	0.73; 1.09	0.71; 1.13	N.A.
	Γ	0.82; 0.97	0.70; 1.09	N.A.
	J	0.86; 1.00	0.81; 1.06	N.A.
	B	N.C.	0.90; 1.03	N.A.
	Δ	N.C.	0.78; 0.93	N.A.
	S	0.74; 1.19	0.88; 1.04	concordance: 0.88; 1.00
	$\gamma(AC_1)$	0.91; 1.03	0.91; 1.03	gwet.ac1: 0.91; 1.00
	S^*	N.C.	0.91; 1.03	wlin.conc: 0.91; 1.00
PTF7	π	0.68; 1.10	0.64; 1.17	N.A.
($n = 35$)	K	0.68; 1.10	0.63; 1.18	confint(res.k): 0.69; 1.00
	K_C	0.79; 1.21	1.00; 1.00	N.A.
	G	0.85; 1.07	0.87; 1.04	N.A.
	K^*	N.C.	0.66; 1.15	confint(res.k): 0.69; 1.00
	α	0.68; 1.10	0.67; 1.12	N.A.
	Γ	0.81; 0.96	0.67; 1.11	N.A.
	J	0.84; 1.00	0.77; 1.06	N.A.
	B	N.C.	0.88; 1.03	N.A.
	Δ	N.C.	0.79; 0.95	N.A.
	S	0.72; 1.19	0.87; 1.04	concordance: 0.87; 1.00
	$\gamma(AC_1)$	0.91; 1.03	0.90; 1.03	gwet.ac1: 0.90; 1.00
	S^*	N.C.	0.91; 1.03	wlin.conc: 0.90; 1.00
AP1	π	−0.21; 0.73	−0.11; 0.65	N.A.
($n = 44$)	K	−0.20; 0.72	−0.10; 0.64	confint(res.k): −0.10; 0.63
	K_C	0.06; 0.66	−0.23; 0.93	N.A.
	G	0.50; 0.96	0.55; 0.90	N.A.
	K^*	N.C.	−0.05; 0.66	confint(res.k): −0.06; 0.64
	α	−0.71; 1.25	−0.08; 0.67	N.A.
	Γ	0.33; 0.45	0.10; 0.67	N.A.
	J	0.53; 0.66	0.36; 0.82	N.A.
	B	N.C.	0.66; 0.93	N.A.
	Δ	N.C.	0.11; 0.87	N.A.
	S	0.52; 0.94	0.55; 0.90	concordance: 0.56; 0.90
	$\gamma(AC_1)$	0.65; 0.93	0.65; 0.93	gwet.ac1: 0.65; 0.94
	S^*	N.C.	0.67; 0.92	wlin.conc: 0.67; 0.92
AP2	π	−0.33; 0.41	−0.19; 0.27	N.A.
($n = 70$)	K	−0.31; 0.42	−0.17; 0.27	confint(res.k): −0.17; 0.27
	K_C	−0.17; 0.32	−0.32; 0.47	N.A.
	G	0.38; 0.80	0.43; 0.75	N.A.
	K^*	N.C.	−0.13; 0.29	confint(res.k): −0.14; 0.29
	α	−0.32; 0.42	−0.18; 0.28	N.A.
	Γ	0.16; 0.24	−0.02; 0.38	N.A.
	J	0.38; 0.46	0.23; 0.60	N.A.
	B	N.C.	0.56; 0.82	N.A.
	Δ	N.C.	−0.26; 0.69	N.A.
	S	0.43; 0.76	0.43; 0.75	concordance: 0.42; 0.74
	$\gamma(AC_1)$	0.55; 0.82	0.55; 0.82	gwet.ac1: 0.55; 0.82

(*continued*)

Table 4. Continued.

		Confidence intervals		
Trichotomous variables (BCS)	Agreement indices	By closed formulas	By Bootstrap-*t* Method	By R functions
	S^*	N.C.	0.58; 0.81	wlin.conc: 0.58; 0.81
AP3 ($n = 46$)	π	−0.67; 0.43	−0.18; −0.05	N.A.
	K	−0.65; 0.43	−0.19; −0.05	confint(res.k): −0.18; −0.04
	K_C	−0.46; 0.15	−0.24; −0.03	N.A.
	G	0.35; 0.86	0.43; 0.80	N.A.
	K^*	N.C.	−0.15; −0.02	confint(res.k): −0.15; −0.01
	α	−0.66; 0.44	−0.17; −0.04	N.A.
	Γ	0.15; 0.27	−0.05; 0.44	N.A.
	J	0.40; 0.52	0.24; 0.67	N.A.
	B	N.C.	0.57; 0.86	N.A.
	Δ	N.C.	−0.24; 0.78	N.A.
	S	0.40; 0.81	0.43; 0.80	concordance: 0.41; 0.80
	$\gamma(AC_1)$	0.54; 0.87	0.54; 0.86	gwet.ac1: 0.54; 0.87
	S^*	N.C.	0.56; 0.86	wlin.conc: 0.56; 0.83

n = sample size; π = Scott's π; K = Cohen's K; K_C = Cohen's K_C; G = Holley and Guilford's G; K^* = Cohen's *weighted* K; α = Krippendorff's α; Γ = Hubert's Γ; J = Janson and Vegelius' J; B = Bangdiwala's B; Δ = Andrés and Marzo's Δ; S = Quatto's S; $\gamma(AC_1)$ = Gwet's $\gamma(AC_1)$; S^* = Quatto's *weighted* S; ITF1 = Italian farm 1; ITF2 = Italian farm 2; ITF3 = Italian farm 3; ITF4 = Italian farm 4; ITF5 = Italian farm 5; ITF6 = Italian farm 6; ITF7 = Italian farm 7; PTF1 = Portuguese farm 1; PTF2 = Portuguese farm 2; PTF3 = Portuguese farm 3; PTF4 = Portuguese farm 4; PTF5 = Portuguese farm 5; PTF6 = Portuguese farm 6; PTF7 = Portuguese farm 7; AP1 = alpine pasture 1; AP2 = alpine pasture 2; AP3 = alpine pasture 3; N.C. = not computable (i.e. the closed formula is too complex to be implemented manually); N.A.N. = not a number (i.e. the closed formula or the Bootstrap Method does not return any number); N.A. = not available (i.e. no R function available to compute confidence intervals).

showed identical values and (ii) Gwet's $\gamma(AC_1)$ conferred the agreement results closest to P_0.

Confidence intervals for Ear posture and Eye white

As already observed for the trichotomous indicators, also for the four-level indicators there was a substantial agreement between the confidence intervals implemented with the closed formulas of variance estimates and the confidence intervals obtained using the Bootstrap Method (Table 7). The confidence intervals obtained using R functions (when available) were also very close to those obtained using both closed formulas and the Bootstrap Method.

In all the considered cases, the confidence intervals obtained for Holley and Guilford's G and Quatto's S were identical. In addition, also Cohen's K and Krippendorff's α showed very similar or identical confidence intervals (Table 7).

All the agreement indices implemented for the four-level indicators showed confidence intervals characterised by similar widths (Table 7).

Discussion

Evaluation of IOR for trichotomous animal-based welfare indicators

The BCS and KNC, which were chosen as examples of trichotomous animal-based welfare indicators in the current study, behave both like categorical variables (variables that express values divided into pre-established categories, which cannot be ordered) and ordinal variables (variables which express countable and orderable values) (Stevens 1946). For this reason, all the indices used to evaluate the agreement between two observers in the case of dichotomous categorical indicators (e.g. udder asymmetry; Giammarino et al. 2021) are also suitable to evaluate the agreement between two observers in the case of trichotomous categorical indicators. Exceptions are (i) Cohen's *weighted K* and Quatto *weighted S*, which can be used for ordinal variables, only; and (ii) *K* PABAK as, according to Byrt et al. (1993), can be implemented for dichotomous variables only, even if there are examples of its use for trichotomous indicators (Thomsen and Baadsgaard 2006).

As reported by Giammarino et al. (2021) in the case of dichotomous animal-based indicators and the presence of two observers, also for trichotomous indicators Scott's π, Cohen's K and Krippendorff's α gave very low agreement results in some of the Italian and Portuguese farms and in all the alpine pastures, if compared to the obtained P_0 (Tables 2 and 3). This phenomenon was identified by Feinstein and Cicchetti (1990) for the Kappa statistics (Cohen 1960; Fleiss 1971; Hubert 1977b), being defined as 'paradox behaviour', which occurs when, despite a high P_0, some indices confer low agreement values. The main explanation of this effect was already highlighted by Kraemer (1979), who showed the problem of the prevalence, defined as the frequency attribution of a subject to the same category by the observers. If the prevalence is high, the lack of variability in assigning the variables to the categories makes the marginal distributions unbalanced within the concordance matrix (Feinstein and Cicchetti 1990). This leads to an increase of P_e which, in some cases, results in negative values of Cohen's K, as observed in the current study

Table 5. Values of the confidence intervals for the agreement indices obtained for knee calluses (KNC) implemented using closed formulas, Bootstrap-*t* Method, and R functions in the fourteen Italian and Portuguese dairy goat farms.

		Confidence intervals		
Trichotomous variables (Knee calluses)	Agreement indices	By closed formulas	By Bootstrap-*t* Method	By R functions
ITF1	π	0.23; 1.04	0.20; 1.11	N.A.
($n = 49$)	K	0.23; 1.04	0.20; 1.13	confint(res.k): 0.25; 1.00
	K_C	0.48; 0.96	0.15; 1.20	N.A.
	G	0.77; 1.04	0.81; 1.01	N.A.
	K^*	N.C.	0.21; 1.10	confint(res.k): 0.25; 1.00
	α	0.20; 1.08	0.22; 1.10	N.A.
	Γ	0.71; 0.82	0.53; 0.99	N.A.
	J	0.79; 0.91	0.69; 1.01	N.A.
	B	N.C.	0.85; 1.01	N.A.
	Δ	N.C.	0.71; 0.94	N.A.
	S	0.71; 1.11	0.81; 1.01	concordance: 0.82; 1.00
	$\gamma(AC_1)$	0.85; 1.01	0.86; 1.01	gwet.ac1: 0.86; 1.00
	S^*	N.C.	0.85; 1.01	wlin.conc: 0.84; 1.00
ITF2	π	−0.48; 0.66	−0.25; 0.47	N.A.
($n = 37$)	K	−0.43; 0.66	−0.18; 0.45	confint(res.k): −0.21; 0.45
	K_C	−0.19; 0.71	−0.50; 1.05	N.A.
	G	0.41; 0.94	0.48; 0.88	N.A.
	K^*	N.C.	−0.17; 0.46	confint(res.k): −0.18; 0.45
	α	−0.44; 0.64	−0.23; 0.47	N.A.
	Γ	0.23; 0.38	−0.03; 0.61	N.A.
	J	0.47; 0.61	0.29; 0.76	N.A.
	B	N.C.	0.61; 0.92	N.A.
	Δ	N.C.	0.12; 0.86	N.A.
	S	0.45; 0.90	0.48; 0.88	concordance: 0.43; 0.88
	$\gamma(AC_1)$	0.59; 0.91	0.59; 0.93	gwet.ac1: 0.58; 0.93
	S^*	N.C.	0.61; 0.91	wlin.conc: 0.60; 0.91
ITF3	π	N.A.N.	N.A.N.	N.A.
($n = 43$)	K	N.A.N.	N.A.N.	confint(res.k): N.A.N.
	K_C	N.A.N.	N.A.N.	N.A.
	G	1.00; 1.00	1.00; 1.00	N.A.
	K^*	N.C.	N.A.N.	confint(res.k): N.A.N.
	α	N.A.N.	N.A.N.	N.A.
	Γ	0.94; 1.06	1.00; 1.00	N.A.
	J	0.94; 1.06	1.00; 1.00	N.A.
	B	N.C.	1.00; 1.00	N.A.
	Δ	N.C.	N.A.N.	N.A.
	S	0.79; 1.21	1.00; 1.00	concordance:1.00; 1.00
	$\gamma(AC_1)$	1.00; 1.00	1.00; 1.00	gwet.ac1: 1.00; 1.00
	S^*	N.C.	1.00; 1.00	wlin.conc: 1.00; 1.00
ITF4	π	−2.01; 1.98	−0.03; 0.02	N.A.
($n = 30$)	K	−1.93; 1.93	0.00; 0.00	confint(res.k): 0.00; 0.00
	K_C	N.A.N.	N.A.N.	N.A.
	G	0.82; 1.08	0.86; 1.04	N.A.
	K^*	N.C.	0.00; 0.00	confint(res.k): 0.00; 0.00
	α	−1.78; 1.78	−1.33; 0.59	N.A.
	Γ	0.77; 0.96	0.62; 1.11	N.A.
	J	0.84; 1.02	0.79; 1.07	N.A.
	B	N.C.	0.90; 1.03	N.A.
	Δ	N.C.	0.81; 0.94	N.A.
	S	0.70; 1.20	0.86; 1.04	concordance: 0.85; 1.00
	$\gamma(AC_1)$	0.90; 1.04	0.90; 1.03	gwet.ac1: 0.90; 1.00
	S^*	N.C.	0.89; 1.03	wlin.conc: 0.89; 1.00
ITF5	π	−0.83; 0.62	−0.17; −0.02	N.A.
($n = 30$)	K	−0.65; 0.65	0.00; 0.00	confint(res.k): 0.00; 0.00
	K_C	N.A.N.	N.A.N.	N.A.
	G	0.35; 0.95	0.43; 0.87	N.A.
	K^*	N.C.	0.00; 0.00	confint(res.k): 0.00; 0.00
	α	−0.71; 0.55	−0.18; 0.02	N.A.
	Γ	0.17; 0.35	−0.11; 0.57	N.A.
	J	0.42; 0.60	0.24; 0.76	N.A.
	B	N.C.	0.61; 0.92	N.A.
	Δ	N.C.	0.37; 0.77	N.A.
	S	0.40; 0.90	0.43; 0.87	concordance: 0.40; 0.85
	$\gamma(AC_1)$	0.55; 0.93	0.55; 0.94	gwet.ac1: 0.54; 0.94
	S^*	N.C.	0.57; 0.90	wlin.conc: 0.55; 0.89
ITF6	π	−0.97; 0.81	−0.15; −0.01	N.A.
($n = 34$)	K	−0.81; 0.81	0.00; 0.00	confint(res.k): 0.00; 0.00
	K_C	N.A.N.	N.A.N.	N.A.

(*continued*)

Table 5. Continued.

		Confidence intervals		
Trichotomous variables (Knee calluses)	Agreement indices	By closed formulas	By Bootstrap-*t* Method	By R functions
	G	0.54; 1.02	0.60; 0.97	N.A.
	K^*	N.C.	0.00; 0.00	confint(res.k): 0.00; 0.00
	α	−0.86; 0.74	−0.24; 0.11	N.A.
	Γ	0.40; 0.56	0.14; 0.79	N.A.
	J	0.61; 0.77	0.46; 0.91	N.A.
	B	N.C.	0.73; 0.97	N.A.
	Δ	N.C.	0.55; 0.85	N.A.
	S	0.54; 1.02	0.60; 0.97	concordance: 0.56; 0.96
	$\gamma(AC_1)$	0.70; 0.98	0.70; 0.99	gwet.ac1: 0.70; 0.99
	S^*	N.C.	0.70; 0.97	wlin.conc: 0.70; 0.97
ITF7 ($n = 39$)	π	−1.19; 1.11	−0.08; 0.01	N.A.
	K	−1.09; 1.09	0.00; 0.00	confint(res.k): 0.00; 0.00
	K_C	N.A.N.	N.A.N.	N.A.
	G	0.72; 1.05	0.76; 1.00	N.A.
	K^*	N.C.	0.00; 0.00	confint(res.k): 0.00; 0.00
	α	−1.25; 1.19	−0.50; 0.35	N.A.
	Γ	0.64; 0.78	0.41; 0.99	N.A.
	J	0.76; 0.90	0.66; 0.99	N.A.
	B	N.C.	0.84; 1.01	N.A.
	Δ	N.C.	0.71; 0.92	N.A.
	S	0.66; 1.11	0.76; 1.00	concordance: 0.73; 1.00
	$\gamma(AC_1)$	0.83; 1.01	0.83; 1.01	gwet.ac1: 0.83; 1.00
	S^*	N.C.	0.82; 1.01	wlin.conc: 0.80; 1.00
PTF1 ($n = 48$)	π	0.38; 0.91	0.39; 0.94	N.A.
	K	0.39; 0.91	0.41; 0.93	confint(res.k): 0.40; 0.90
	K_C	0.74; 1.26	1.00; 1.00	N.A.
	G	0.63; 1.00	0.67; 0.96	N.A.
	K^*	N.C.	0.43; 0.93	confint(res.k): 0.43; 0.91
	α	0.36; 0.94	0.37; 0.96	N.A.
	Γ	0.50; 0.61	0.27; 0.83	N.A.
	J	0.64; 0.76	0.49; 0.90	N.A.
	B	N.C.	0.74; 0.96	N.A.
	Δ	N.C.	0.58; 0.87	N.A.
	S	0.61; 1.01	0.67; 0.96	concordance: 0.66; 0.94
	$\gamma(AC_1)$	0.73; 0.97	0.73; 0.96	gwet.ac1: 0.73; 0.97
	S^*	N.C.	0.75; 0.96	wlin.conc: 0.74; 0.95
PTF2 ($n = 38$)	π	−0.20; 0.77	−0.08; 0.72	N.A.
	K	−0.18; 0.77	−0.10; 0.70	confint(res.k): −0.09; 0.67
	K_C	0.09; 0.75	−0.26; 1.04	N.A.
	G	0.48; 0.97	0.54; 0.91	N.A.
	K^*	N.C.	−0.05; 0.73	confint(res.k): −0.05; 0.69
	α	−0.16; 0.76	−0.10; 0.72	N.A.
	Γ	0.31; 0.46	0.05; 0.69	N.A.
	J	0.52; 0.66	0.33; 0.82	N.A.
	B	N.C.	0.64; 0.93	N.A.
	Δ	N.C.	0.25; 0.84	N.A.
	S	0.50; 0.95	0.54; 0.91	concordance: 0.53; 0.92
	$\gamma(AC_1)$	0.64; 0.94	0.64; 0.94	gwet.ac1: 0.63; 0.95
	S^*	N.C.	0.65; 0.93	wlin.conc: 0.64; 0.91
PTF3 ($n = 25$)	π	0.12; 1.13	0.10; 1.30	N.A.
	K	0.14; 1.12	0.11; 1.24	confint(res.k): 0.18; 1.00
	K_C	0.51; 1.49	1.00; 1.00	N.A.
	G	0.67; 1.09	0.72; 1.04	N.A.
	K^*	N.C.	0.13; 1.24	confint(res.k): 0.19; 1.00
	α	0.25; 1.03	0.10; 1.24	N.A.
	Γ	0.58; 0.80	0.31; 1.03	N.A.
	J	0.70; 0.92	0.56; 1.05	N.A.
	B	N.C.	0.79; 1.03	N.A.
	Δ	N.C.	0.58; 0.89	N.A.
	S	0.60; 1.16	0.72; 1.04	concordance: 0.70; 1.00
	$\gamma(AC_1)$	0.79; 1.03	0.78; 1.05	gwet.ac1: 0.78; 1.00
	S^*	N.C	0.79; 1.03	wlin.conc: 0.78; 1.00
PTF4 ($n = 39$)	π	N.A.N.	N.A.N.	N.A.
	K	N.A.N.	N.A.N.	confint(res.k): N.A.N.
	K_C	N.A.N.	N.A.N.	N.A.
	G	1.00; 1.00	1.00; 1.00	N.A.
	K^*	N.C.	N.A.N.	confint(res.k): N.A.N.
	α	N.A.N.	N.A.N.	N.A.
	Γ	0.93; 1.07	1.00; 1.00	N.A.
	J	0.93; 1.07	1.00; 1.00	N.A.

(*continued*)

Table 5. Continued.

		Confidence intervals		
Trichotomous variables (Knee calluses)	Agreement indices	By closed formulas	By Bootstrap-*t* Method	By R functions
	B	N.C.	1.00; 1.00	N.A.
	Δ	N.C.	N.A.N.	N.A.
	S	0.78; 1.22	1.00; 1.00	concordance: 1.00; 1.00
	$\gamma(AC_1)$	1.00; 1.00	1.00; 1.00	gwet.ac1: 1.00; 1.00
	*S**	N.C.	1.00; 1.00	wlin.conc: 1.00; 1.00
PTF5	π	−2.00; 1.98	−0.03; 0.02	N.A.
(*n* = 32)	*K*	−1.93; 1.93	0.00; 0.00	confint(res.k): 0.00; 0.00
	K_C	N.A.N.	N.A.N.	N.A.
	G	0.83; 1.07	0.86; 1.04	N.A.
	*K**	N.C.	0.00; 0.00	confint(res.k): 0.00; 0.00
	α	−1.78; 1.78	−1.32; 0.59	N.A.
	Γ	0.79; 0.96	0.64; 1.10	N.A.
	J	0.84; 1.02	0.80; 1.06	N.A.
	B	N.C.	0.91; 1.03	N.A.
	Δ	N.C.	0.82; 0.94	N.A.
	S	0.71; 1.20	0.86; 1.04	concordance: 0.86; 1.00
	$\gamma(AC_1)$	0.91; 1.03	0.91; 1.03	gwet.ac1: 0.90; 1.00
	*S**	N.C.	0.90; 1.03	wlin.conc: 0.89; 1.00
PTF6	π	0.72; 1.04	0.71; 1.06	N.A.
(*n* = 38)	*K*	0.72; 1.04	0.72; 1.06	confint(res.k): 0.72; 1.00
	K_C	0.84; 1.16	1.00; 1.00	N.A.
	G	0.78; 1.06	0.82; 1.03	N.A.
	*K**	N.C.	0.73; 1.05	confint(res.k): 0.74; 1.00
	α	0.72; 1.04	0.71; 1.07	N.A.
	Γ	0.72; 0.87	0.54; 1.05	N.A.
	J	0.78; 0.92	0.66; 1.03	N.A.
	B	N.C.	0.82; 1.03	N.A.
	Δ	N.C.	0.73; 0.91	N.A.
	S	0.70; 1.15	0.82; 1.03	concordance: 0.80; 1.00
	$\gamma(AC_1)$	0.84; 1.02	0.84; 1.03	gwet.ac1: 0.84; 1.00
	*S**	N.C.	0.86; 1.02	wlin.conc: 0.85; 1.00
PTF7	π	−1.42; 1.37	−0.04; 0.01	N.A.
(*n* = 35)	*K*	−1.35; 1.35	0.00; 0.00	confint(res.k): 0.00; 0.00
	K_C	N.A.N.	N.A.N.	N.A.
	G	0.76; 1.07	0.80; 1.03	N.A.
	*K**	N.C.	0.00; 0.00	confint(res.k): 0.00; 0.00
	α	−1.25; 1.23	−0.84; 0.55	N.A.
	Γ	0.70; 0.86	0.50; 1.04	N.A.
	J	0.79; 0.95	0.70; 1.03	N.A.
	B	N.C.	0.87; 1.02	N.A.
	Δ	N.C.	0.71; 0.91	N.A.
	S	0.68; 1.15	0.80; 1.03	concordance: 0.79; 1.00
	$\gamma(AC_1)$	0.86; 1.02	0.86; 1.03	gwet.ac1: 0.86; 1.00
	*S**	N.C.	0.85; 1.02	wlin.conc: 0.84; 1.00

n = sample size; π = Scott's π; *K* = Cohen's *K*; K_C = Cohen's K_C; *G* = Holley and Guilford's *G*; *K** = Cohen's *weighted K*; α = Krippendorff's α; Γ = Hubert's Γ; *J* = Janson and Vegelius' *J*; *B* = Bangdiwala's *B*; Δ = Andrés and Marzo's Δ; *S* = Quatto's *S*; $\gamma(AC_1)$ = Gwet's $\gamma(AC_1)$; *S** = Quatto's *weighted S*; ITF1 = Italian farm 1; ITF2 = Italian farm 2; ITF3 = Italian farm 3; ITF4 = Italian farm 4; ITF5 = Italian farm 5; ITF6 = Italian farm 6; ITF7 = Italian farm 7; PTF1 = Portuguese farm 1; PTF2 = Portuguese farm 2; PTF3 = Portuguese farm 3; PTF4 = Portuguese farm 4; PTF5 = Portuguese farm 5; PTF6 = Portuguese farm 6; PTF7 = Portuguese farm 7; N.C. = not computable (i.e. the closed formula is too complex to be implemented manually); N.A.N. = not a number (i.e. the closed formula, the Bootstrap Method or the R function does not return any number); N.A. = not available (i.e. no R function available to compute confidence intervals).

for AP3 (Table 2). Although the paradox effect was preliminarily studied for the Kappa statistics, Scott's π and Krippendorff's α are affected by the same problem, sometimes giving negative values (Tables 2 and 3), as also observed by Giammarino et al. (2021) in the case of dichotomous indicators. For both Scott's π and Cohen's *K*, when the observers assign all the subjects to the same category, the chance agreement is equal to 1, producing low agreement results, despite a high P_0 (Gwet 2001). Krippendorff's α can be implemented to evaluate the IOR for both ordinal and categorical variables characterised by two or more categories, and in the presence of two or several observers (Krippendorff 2011). Despite considering both the level of agreement and disagreement between the observers (Krippendorff 2011), Krippendorff's α follows the same statistical approach of Scott's π and Cohen's *K*, suffering from the paradox behaviour too (Giammarino et al. 2021). Moreover, when the P_0 was equal to 100% (Table 3), Scott's π, Cohen's *K* and Krippendorff's α were not computable, being both the P_0 and the P_e equal to 1, giving a ratio of 0/0 (see Appendix A in Giammarino et al. 2021).

In the case of trichotomous indicators, Cohen's *weighted K*, which is an extension of Cohen's *K* for ordinal variables, was affected by the paradox behaviour too. Specifically, while implementing the P_e for Cohen's *weighted K*, the linear weights proposed by Cicchetti and Allison (1971) are used, as they are less sensitive than the quadratic weights to the number of categories of the variable (Brenner and Kliebsch 1996). Furthermore, during the implementation of Cohen's *weighted K*, both the level of agreement and disagreement between observers are considered, improving in some cases the performance of Cohen's *K*. Indeed, the agreement values conferred by Cohen's *weighted K* were higher than those given by Cohen's *K* (Tables 2 and 3) but, in any case, very low if compared to P_0*, confirming the presence of the paradox behaviour also for this index. As observed for Scott's π, Cohen's *K* and Krippendorff's α, also Cohen's *weighted K* was not computable when the observed P_0* was equal to 100%, for the same reasons explained for the former indices. Moreover, in all the considered cases, the concordance rate (P_0*) obtained for Cohen's *weighted K* was different and higher than the concordance rate (P_0) obtained for Cohen's *K* (Tables 2 and 3). This occurs because Cohen's *weighted K* is implemented using a different matrix, in which both the level of agreement and disagreement between observers is considered; on the contrary, in the classic matrix

Table 6. Values of the concordance rate and of the agreement indices obtained for ear posture (EP) and eye white (EW) for the three intensively managed italian dairy cattle farms.

			Agreement Index					
	n	P_0	K	K_C	G	α	S	$\gamma(AC_1)$
EAR POSTURE								
EP-F1	126	88%	0.76	0.86	0.84	0.76	0.84	0.86
EP-F2	42	81%	0.70	0.82	0.75	0.70	0.75	0.76
EP-F3	51	78%	0.69	1	0.71	0.69	0.71	0.72
EYE WHITE								
EW-F1	126	63%	0.49	0.62	0.51	0.49	0.51	0.52
EW-F2	42	62%	0.50	0.69	0.49	0.50	0.49	0.49
EW-F3	51	80%	0.67	0.80	0.74	0.67	0.74	0.76

n = sample size; P_0 = concordance rate; K = Cohen's K; K_C = Cohen's K_C; G = Holley and Guilford's G; α = Krippendorff's α; S = Quatto's S; $\gamma(AC_1)$ = Gwet's $\gamma(AC_1)$; EP = Ear posture; EW = Eye white; F1 = Farm 1; F2 = Farm 2; F3 = Farm 3.

Table 7. Values of the confidence intervals for the agreement indices obtained for ear posture (EP) and eye white (EW) implemented using closed formulas, Bootstrap-*t* Method, and R functions in the three intensively managed italian dairy cattle farms.

		Confidence intervals		
Four-level variables (EP-EW)	Agreement indices	By closed formulas	By Bootstrap-*t* Method	By R functions
EP-F1 (n = 126)	K	0.65; 0.87	0.66; 0.87	confint(res.k): 0.66; 0.87
	K_C	0.78; 0.94	0.76; 0.94	N.A.
	G	0.73; 0.95	0.77; 0.92	N.A.
	α	0.65; 0.87	0.66; 0.87	N.A.
	S	0.74; 0.94	0.77; 0.92	concordance: 0.77; 0.92
	$\gamma(AC_1)$	0.78; 0.93	0.79; 0.92	gwet.ac1: 0.79; 0.93
EP-F2 (n = 42)	K	0.51; 0.89	0.52; 0.88	confint(res.k): 0.52; 0.88
	K_C	0.68; 0.96	0.61; 0.99	N.A.
	G	0.51; 0.98	0.59; 0.90	N.A.
	α	0.51; 0.89	0.53; 0.88	N.A.
	S	0.57; 0.92	0.59; 0.90	concordance: 0.59; 0.90
	$\gamma(AC_1)$	0.60; 0.92	0.61; 0.91	gwet.ac1: 0.60; 0.92
EP-F3 (n = 51)	K	0.53; 0.85	0.53; 0.85	confint(res.k): 0.53; 0.85
	K_C	0.84; 1.16	1.00; 1.00	N.A.
	G	0.49; 0.94	0.55; 0.87	N.A.
	α	0.52; 0.86	0.51; 0.86	N.A.
	S	0.55; 0.87	0.55; 0.87	concordance: 0.56; 0.84
	$\gamma(AC_1)$	0.57; 0.88	0.57; 0.87	gwet.ac1: 0.57; 0.87
EW-F1 (n = 126)	K	0.37; 0.61	0.38; 0.60	confint(res.k): 0.38; 0.60
	K_C	0.53; 0.71	0.49; 0.73	N.A.
	G	0.35; 0.68	0.40; 0.62	N.A.
	α	0.37; 0.61	0.37; 0.60	N.A.
	S	0.41; 0.61	0.40; 0.62	concordance: 0.40; 0.62
	$\gamma(AC_1)$	0.40; 0.64	0.41; 0.63	gwet.ac1: 0.41; 0.64
EW-F2 (n = 42)	K	0.31; 0.69	0.31; 0.69	confint(res.k): 0.31; 0.68
	K_C	0.53; 0.86	0.45; 0.92	N.A.
	G	0.20; 0.79	0.30; 0.68	N.A.
	α	0.30; 0.70	0.30; 0.69	N.A.
	S	0.32; 0.67	0.30; 0.68	concordance: 0.30; 0.68
	$\gamma(AC_1)$	0.28; 0.70	0.29; 0.69	gwet.ac1: 0.29; 0.70
EWF3 (n = 51)	K	0.48; 0.85	0.49; 0.84	confint(res.k): 0.50; 0.84
	K_C	0.66; 0.94	0.64; 0.92	N.A.
	G	0.52; 0.96	0.60; 0.88	N.A.
	α	0.48; 0.86	0.51; 0.84	N.A.
	S	0.58; 0.90	0.60; 0.88	concordance: 0.58; 0.87
	$\gamma(AC_1)$	0.61; 0.90	0.62; 0.90	gwet.ac1: 0.61; 0.90

n = sample size; K = Cohen's K; K_C = Cohen's K_C; G = Holley and Guilford's G; α = Krippendorff's α; S = Quatto's S; $\gamma(AC_1)$ = Gwet's $\gamma(AC_1)$; EP = Ear posture; EW = Eye white; F1 = Farm 1; F2 = Farm 2; F3 = Farm 3; N.A. = not available (i.e. no R function available to compute confidence intervals).

only the level of agreement is considered, as the agreement is based only on a categorical scale.

The paradox behaviour was also highlighted in the current study through the implementation of the confidence intervals (Tables 4 and 5). Generally, the best agreement indices are those characterised by confidence intervals with a narrow width (Giammarino et al. 2021), as the values assumed by the indices are not dispersed in the sample. Wide confidence intervals were obtained for Scott's π, Cohen's K, Cohen's *weighted* K and Krippendorff's α in many cases for both BCS (Table 4) and KNC (Table 5). However, in some cases, these indices showed very narrow confidence intervals (e.g. BCS in AP3; Table 4). This is due to the paradox effect, as the negative values assumed by the above-mentioned indices in AP3 and the lack of variability in assigning the subjects to the categories, paradoxically produce confidence intervals characterised by negative extremes. As already highlighted by Giammarino et al. (2021) in the presence of dichotomous indicators, in most of the intensively dairy goat farms in the current study, for both BCS and KNC the confidence intervals were wide for Scott's π, Cohen's K, Cohen's *weighted* K and Krippendorff's α, even when the paradox behaviour did not occur. Moreover, when the values assumed by Scott's π, Cohen's K, Cohen's *weighted* K and Krippendorff's α were not computable, also the confidence intervals for these indices were not computable, as all the possible values assumed by these indices during the Bootstrap resampling resulted in 'not a number'.

While reading the published literature on this topic, we highlighted the paradox behaviour in some studies for dichotomous (Vieira et al. 2018; Munoz et al. 2017) and trichotomous (Vieira et al. 2018; Pedersen et al. 2011) animal-based welfare indicators, in the case of an evaluation performed by two observers. When considering trichotomous indicators, for example in Vieira et al. (2018), this problem occurred during the evaluation of IOR for BCS and KNC, the same variables used in our study. These Authors assessed the IOR by computing Cohen's *weighted* K. Signs of paradox were prominent for KNC ($P_0{*} = 91\%$; Cohen's *weighted* $K = 0.27$) and evident also for BCS ($P_0{*} = 79\%$; Cohen's *weighted* $K = 0.46$) in the Italian farms evaluated by Vieira et al. (2018). Pedersen et al. (2011) evaluated the IOR of faecal consistency, a trichotomous categorical indicator used to assess welfare in grow-finishing pigs. The concordance was evaluated among three pairs of observers (AB; AC; BC) by computing Cohen's K which, in one case, was affected by the paradox behaviour ($P_{0AB} = 61\%$; Cohen's $K = 0.24$).

To solve the paradox problem, Cohen proposed the Cohen's K_C as an alternative to the original Cohen's K but, as highlighted in the current study, also Cohen's K_C is affected by the paradox behaviour, conferring low agreement results in all the alpine pastures, and in some of the Italian and Portuguese farms both for BCS and KNC, if compared to P_0 (Tables 2 and 3). In particular, the computation of Cohen's K_C is based on the maximum Kappa (K_M), defined as the proportion of the standardisation of the difference between the maximum value of P_0 (P_{0max}) and the value of P_e, and the difference between the maximum value of Kappa ($K = 1$) and the value of P_e. The maximum value of Kappa is reached when the values outside the diagonal of the matrix are equal to 0, and the total marginal distributions are equal each other (Giammarino et al. 2021). However, if the marginal distributions are unbalanced, the maximum value of Kappa will not be equal to 1 (Cohen 1960). Cohen's K_C is given by the ratio between Cohen's K and K_M, so that this index is strongly influenced by the values assumed by both of them. In the case of trichotomous indicators, the values assumed by Cohen's K were low while the values assumed by K_M were high, resulting in Cohen's K_C which significantly improved the performance of Cohen's K in most of the considered cases, especially when considering BCS results (Table 2). The negative value of Cohen's K_C in AP3 is due to the negative value obtained by Cohen's K, which is involved in Cohen's K_C calculation, as previously explained. In several cases, for both BCS and KNC, Cohen's K_C was not computable, as the values assumed by Cohen's K and K_M were equal to 0. As already observed for Scott's π, Cohen's K, Cohen's *weighted* K and Krippendorff's α, also the confidence intervals for Cohen's K_C were wide in most cases, even when no paradox behaviour was detected (Tables 4 and 5); moreover, when the value of Cohen's K_C was not computable (Tables 2 and 3), it was not possible to obtain the confidence intervals for this index, as all the possible values assumed by Cohen's K_C inside the sample during the application of the Bootstrap resampling resulted in 'not a number'.

Andrés and Marzo (2004) also tried to overcome the limitations of the Kappa statistics by means of Andrés and Marzo's Δ. This index was initially created in the case of 2 × 2 tables (dichotomous variables and the presence of two observers). Andrés and Marzo's Δ fits quite well for the IOR evaluation in the case of dichotomous indicators and the presence of two observers (Giammarino et al. 2021), but its performance gets worse when dealing with trichotomous indicators, especially in the presence of concordance rates lower than 75% (Tables 2 and 3). In particular, in most of the considered cases, Andrés and Marzo's Δ improved the performance of Cohen's *K*, especially when the latter index was affected by the paradox behaviour. The confidence intervals based on the Bootstrap resampling for Andrés and Marzo's Δ were wide in several cases.

The values obtained for Andrés and Marzo's Δ were similar to those obtained implementing Hubert's Γ. The latter index conferred low agreement values if compared to P_0, especially when the P_0 was lower than 85% (Table 2). This phenomenon was also identified by Giammarino et al. (2021) for dichotomous indicators, with Hubert's Γ resulting in better agreement results when the P_0 was higher than 80%.

In the case of trichotomous indicators, Janson and Vegelius' *J* overcame the problems which affect Hubert's Γ. For this reason, differently from what was observed in the case of dichotomous indicators (Giammarino et al. 2021), Janson and Vegelius' *J* conferred better agreement results than those given by Hubert's Γ for trichotomous indicators in the current study. The confidence intervals obtained for Hubert's Γ and Janson and Vegelius' *J* were in most of the considered cases characterised by similar widths (Tables 4 and 5).

As already reported by Giammarino et al. (2021) for dichotomous indicators and the presence of two observers, Bangdiwala's *B* and Gwet's $\gamma(AC_1)$ were not affected by the paradox behaviour, and these indices, together with Quatto's *weighted S*, conferred the best agreement results for trichotomous indicators in all the considered cases, followed by Quatto's *S* and Holley and Guilford's *G* (Tables 2 and 3). Moreover, Bangdiwala's *B*, Gwet's $\gamma(AC_1)$ and Quatto's *weighted S* showed the tightest confidence intervals in all the cases, confirming their goodness in evaluating IOR for trichotomous indicators, again followed by Quatto's *S* and Holley and Guilford's *G* (Tables 4 and 5). For Bangdiwala's *B* the agreement between the two observers for BCS is easily seen from Appendix C, where the agreement charts obtained for the three alpine pastures are reported as examples.

Quatto (2004), Marasini et al. (2016) and Gwet (2008), proposed Quatto's *S*, Quatto's *weighted S* and Gwet's $\gamma(AC_1)$, respectively, as alternative agreement indices to solve the problems of the paradox behaviour. These indices are based on a new implementation of P_e, which considers the number of categories that characterises the variables under analysis. This different calculation method leads to a reduction of the chance agreement, solving the problem of paradox. Quatto's *S* defined P_e as the sum of the probabilities in assigning randomly a couple of values to the same variable, so that P_e is given by the ratio between 1 and the number of the response categories (Falotico and Quatto 2010). Quatto's *S* follows the same statistical approach of Holley and Guilford's *G* in the calculation of P_e; this is the reason why Quatto's *S* and Holley and Guilford's *G* conferred identical results in all the cases considered in the current study (Tables 2 and 3). Holley and Guilford's *G* was initially created in the case of 2×2 tables (Gwet 2001), but this index can be also extended to evaluate the IOR of variables characterised by a number of categories > 2.

Quatto's *weighted S* is an extension of Quatto's *S* to evaluate IOR for ordinal variables in the presence of two or more observers, and a number of categories $\geq$ 2. Indeed, Quatto's *S* is suitable to evaluate IOR for categorical variables characterised by any number of categories and the presence of two or more observers (Quatto 2004). For all the cases considered in the current study, and according to an ordered scale, the concordance rate (P_0*) obtained for Quatto's *weighted S* was different from that obtained for all the other agreement indices, but equal to that obtained for Cohen's *weighted K*; indeed, the percentage of observed agreement for these two indices is calculated using the same matrix, where both concordant and discordant pairs are considered (Marasini et al. 2016). Moreover, the implementation of P_e for Quatto's *weighted S* also considers the number of the categories which characterises the variables but, differently from Quatto's *S*, it is developed using the linear weights, for the same reasons explained when computing the P_e for Cohen's *weighted K*. However, in the presence of ordinal variables, it is necessary to highlight that the efficiency of the agreement indices in calculating the concordance among the observers, is related to the subjectivity during the selection of the weights.

In the implementation of P_e, Gwet's $\gamma(AC_1)$ differs from Quatto's *S* and Quatto's *weighted S*, specifying that the expected agreement occurs when at least one observer classifies randomly a variable into a pre-established category (Gwet 2008).

Evaluation of IOR for four-level animal-based welfare indicators

The number of indices implemented to evaluate the agreement between the observers is limited for EP and EW, if compared with those involved in the evaluation of IOR for BCS and KNC because, in our study, these two four-level indicators behave only as categorical variables and are characterised by a number of categories > 3.

The paradox effect was not detected for the four-level indicators in the current study (Table 6). Indeed, having a higher number of categories which characterises the variables, the possibility of choice for the observers to assign each variable to a category is higher and the prevalence decreases (Byrt et al. 1993). This leads to a lower unbalance of the marginal distributions within the concordance matrix and consequently to the reduction of P_e, which implies a lower probability of the presence of the paradox behaviour (Feinstein and Cicchetti 1990).

Despite the paradox behaviour was not detected in the current study for EP and EW, we highlighted signs of paradox in some published studies when four-level indicators were evaluated by two observers. For example, in Buczinski et al. (2016) the reliability of the categorical four-level indicators rectal temperature, cough, nasal discharge, eye discharge and ear position were evaluated by two observers on pre-weaned dairy cows. The concordance was calculated using Cohen's *K*, which demonstrated to be affected by the paradox behaviour for most of the analysed indicators. Indeed, this index showed very low or even negative agreement results if compared to P_0 for cough (P_0 = 78%; Cohen's $K = 0.10$), nasal discharge (P_0 = 62%; Cohen's $K = 0.24$), eye discharge (P_0 = 63%; Cohen's $K = 0.11$) and ear position (P_0 = 85%; Cohen's $K = -0.04$). In particular, for eye discharge and ear position the observers classified the variables into three and two categories only, respectively, even though four different categories were available. This led to a minor heterogeneity in classifying each variable into the categories, producing a higher prevalence and a major unbalance of the marginal distribution within the concordance matrix, resulting in an increase of P_e and in the presence of the paradox behaviour. The paradox behaviour was also detected in Munoz et al. (2017) for the foot-wall integrity, an ordinal four-level indicator used to assess welfare in dairy ewes. The concordance was calculated between three pairs of observers (AB; AC; BC) using Cohen's *weighted K*, showing the paradox behaviour in all the cases (P_{0AB} = 90%; Cohen's *weighted* $K = 0.47$; P_{0AC} = 97%; Cohen's *weighted* $K = 0.21$; P_{0BC} = 95%; Cohen's *weighted* $K = 0.55$).

To better understand the role of P_0 and marginal distributions on the paradox behaviour of Cohen's *K* and Krippendorff's α for four-level indicators, starting from the real matrices we had on EP in F1 and EW in F2, we created three fictitious matrices in each case, and then we calculated the agreement indices and the related confidence intervals (Appendix D). We observed that, when having unbalanced marginal distributions, Cohen's *K* and Krippendorff's α were affected by the paradox behaviour, conferring low agreement results despite high P_0 values [EP-F1 Forced matrices 1, 2 and 3; EW-F2 Forced matrices 2 and 3 (Appendix D)]. In such cases, the confidence intervals were wide for the above-mentioned indices (Appendix D). Only in one case [EW-F2 – Forced matrix 1 (Appendix D)], where the heterogeneity in assigning the scores to the variables was higher (as it was the case of the real data presented in the current study), which resulted in more balanced marginal distributions inside the concordance matrix, the paradox behaviour was not found. On the other hand, even forcing the matrices, Gwet's $\gamma(AC_1)$, followed by Quatto's *S* and Holley and Guilford's *G*, conferred the best agreement results (Appendix D), confirming the results obtained with the real data.

In the case of four-level indicators, Cohen's K_C improved the agreement results obtained with Cohen's *K* both for EP and EW (Table 6). However, in some cases, it exceeded the agreement between the observers, conferring identical or even higher agreement results if compared to P_0 (Table 6). This was already reported by Giammarino et al. (2021) for dichotomous animal-based indicators and the presence of two observers, and it was also observed for the trichotomous indicators analysed in the current study (Tables 2 and 3). The same problem was also observed when forcing the matrices of four-level indicators [EP-F1 Forced matrix 3; EW-F2 Forced matrix 1 (Appendix D)].

As already demonstrated for dichotomous (Giammarino et al. 2021) and trichotomous (current study) indicators, our results show that Gwet's $\gamma(AC_1)$ conferred the best agreement results also for four-level indicators (Table 6), confirming the ability of this index to fit well in the presence of variables characterised by different number of categories when the evaluation is performed by two observers. Only in EW-F2 Gwet's $\gamma(AC_1)$, as well as Quatto's *S* and Holley and Guilford's *G*, gave a slightly lower agreement values than those conferred by Cohen's *K* and Krippendorff's α (Table 6). Indeed, the higher possibility of choice for the observers to assign the scores produced very balanced, and sometimes equal, marginal distributions, resulting in a higher agreement for the latter indices. Following Gwet's $\gamma(AC_1)$, Quatto's *S* and Holley and

Guilford's *G* also gave the highest agreement results during the evaluation of IOR for four-level indicators.

The confidence intervals results obtained with the Bootstrap Method showed that, although Gwet's $\gamma(AC_1)$ was characterised by the tightest confidence intervals, followed by Quatto's *S* and Holley and Guilford's *G*, the differences between the confidence intervals for all the implemented indices were negligible (Table 7). This phenomenon, differently from what was observed in the case of trichotomous indicators, is due to the lack of the paradox behaviour (Feinstein and Cicchetti 1990) for Cohen's *K* and Krippendorff's α, producing a reduction of the dispersion of the possible values assumed by the indices within the sample, and confidence intervals characterised by widths similar to those conferred by Gwet's $\gamma(AC_1)$, Quatto's *S* and Holley and Guilford's *G*. However, when analysing the results obtained with the forced matrices, it is observed that when Cohen's *K* and Krippendorff's α were affected by the paradox behaviour, their confidence intervals sometimes resulted in negative values [EP-F1 Forced matrix 3; EW-F2 Forced matrix 2 (Appendix D)]; such results confirm those obtained with both dichotomous (Giammarino et al. 2021) and trichotomous (current study) indicators.

Considering Cohen's K_C, in one case, when the value of the index was equal to 1 (i.e. EP-F3; Table 6), the confidence intervals obtained with the Bootstrap Method were also equal to 1 (Table 7), due to a reduction of variability of all the possible values assumed by Cohen's K_C in the sample.

EP and EW could be promising animal-based indicators to be included into the animal welfare protocols. Unfortunately, the low P_0 observed in some cases among the observers for EW (i.e. 63% and 62% for EW-F1 and EW-F2, respectively; Table 6) suggests that a reduction of the number of categories (e.g. dichotomous variable; 0 = Eye white not visible; 1 = Eye white visible) would improve the reliability for this indicator. Indeed, the high number of categories which characterises the variable could lead in some cases to a reduction of the concordance rate among the observers, as the possibility of choice for the observers in assigning the scores to the variables increases. Reducing the number of categories, the P_0 increases, as the possibility of choice for the observers in classifying the variable into a specific category is lower.

Conclusions

From the obtained results, it is evident that not all the agreement indices available in the literature are suitable to evaluate the IOR between two observers for trichotomous or four-level animal-based welfare indicators assessed at individual level.

Bangdiwala's *B*, Gwet's $\gamma(AC_1)$ and Quatto's *weighted S* are promising for a proper evaluation of IOR in the case of trichotomous indicators and the presence of two observers, proving to be a valid alternative to Scott's π, Cohen's *K*, Cohen's K_C, Cohen's *weighted K* and Krippendorff's α, which are sometimes affected by the paradox behaviour. In the presence of two observers, Bangdiwala's *B* and Gwet's $\gamma(AC_1)$ can be used for trichotomous indicators which behave only as categorical variables, while Quatto's *weighted S* (using linear weights) is suggested to evaluate IOR for trichotomous indicators which behave only as ordinal variables. All these three agreement indices are suitable to evaluate IOR for trichotomous indicators which behave both as categorical and ordinal variables, and in the presence of two observers. However, it is important to specify that, in the presence of indicators that behave both ways, the observers can choose to consider them as categorical or as ordinal variables, which will imply the use of different agreement indices.

Gwet's $\gamma(AC_1)$, Quatto's *S* and Holley and Guilford's *G* confer the best agreement results also during the evaluation of IOR between two observers in the case of four-level indicators. Five-level animal-based welfare indicators are also present in welfare assessment protocols (AWIN 2015b, 2015c, 2015d; Welfare Quality® 2009b) as well as in published literature (Thomsen et al. 2008; Croyle et al. 2018). The results obtained in this study for four-level indicators can also be extended to categorical variables characterised by a higher number of categories, in the presence of two observers.

With the real data used in this study, the paradox behaviour was not detected for four-level indicators. However, as highlighted in some studies reported in published literature, and as also seen forcing the matrices in the current study, the paradox behaviour can also affect four-level indicators, despite the presence of a high number of categories.

Furthermore, considering any number of categories which characterises the variable under analysis, Quatto's *weighted S* is a reliable index to evaluate IOR for ordinal indicators.

For some agreement indices, closed formulas of variance were too complex to be implemented manually. Our results show that the Bootstrap Method is valid and represents an easiest and more accurate alternative to the closed formulas of variance for the estimation of the confidence intervals of all the agreement indices.

Further studies will be required to identify which agreement indices should be used for a proper evaluation of IOR in the presence of a number of observers greater than two.

Acknowledgments

We thank the farmers who allowed us to visit their farms. We also acknowledge Dr. Mauro Masino, Dr. Andrea Bonzanino, Prof. George Stilwell and his research team for Body Condition Score and Knee calluses data collection on goat farms, and Dr. Anna Agostini and Dr. Federica Manila Soli for Ear posture and Eye white data collection on cattle farms.

Ethical approval

The BCS data used in this study were obtained performing a trial that was approved by the Bioethics Committee of the University of Turin (Italy) (protocol n° 0587791). Ethical approval for collecting photos to evaluate EP and EW was not needed according to EU regulations because the experimental procedures were not likely to cause pain, suffering, distress or lasting harm equivalent to, or higher than, that caused by the introduction of a needle in accordance with good veterinary practice.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Ear posture and Eye white data were collected within the project 'LATTE.DOC – Development of an innovative supply chain management model to improve internal and external information flow, optimise processes and obtain sustainable and high-quality dairy products that meet consumer needs' [PSR 2014-2020. Op. 16.2], funded by Lombardy region.

ORCID

Benedetta Torsiello http://orcid.org/0000-0002-3859-2609
Mauro Giammarino http://orcid.org/0000-0002-9801-929X
Piero Quatto http://orcid.org/0000-0002-6679-7169
Monica Battini http://orcid.org/0000-0002-6134-7759
Silvana Mattiello http://orcid.org/0000-0003-1885-949X
Luca Battaglini http://orcid.org/0000-0002-2136-3826
Manuela Renna http://orcid.org/0000-0003-4296-7589

Data availability statement

The data that support the findings of this study are available from the corresponding author [M.G.] upon reasonable request.

References

Altman DG. 2000. Statistics in medical journals: some recent trends. Stat Med. 19:3275–3289. doi: 10.1002/10970258(20001215)19:23%3C3275::AIDSIM626%3E3.0.CO;2-M.

Andrés AM, Marzo PF. 2004. Delta: a new measure of agreement between two raters. Br J Math Stat Psychol. 57(Pt 1):1–19. doi: 10.1348/000711004849268.

AWIN. 2015a. AWIN welfare assessment protocol for goats. doi: 10.13130/AWIN_goats_2015.

AWIN. 2015b. AWIN welfare assessment protocol for sheep. doi: 10.13130/AWIN_sheep_2015.

AWIN. 2015c. AWIN welfare assessment protocol for horses. doi: 10.13130/AWIN_horses_2015.

AWIN. 2015d. AWIN welfare assessment protocol for donkeys. doi: 10.13130/AWIN_donkeys_2015.

Bajpai S, Bajpai RC, Chaturvedi HK. 2015. Evaluation of inter-rater agreement and inter-rater reliability for observational data: an overview of concepts and methods. J Indian Acad Appl Psychol. 41:20–27.

Bangdiwala SI. 1985. A graphical test for observer agreement. Proceedings of the 45th International Statistical Institute Meeting; August 12-22, Amsterdam (NL); Springer Ed. p. 307–308.

Bangdiwala SI, Haedo AS, Natal ML, Villaveces A. 2008. The agreement chart as an alternative to the receiver-operating characteristic curve for diagnostic tests. J Clin Epidemiol. 61(9):866–874. doi: 10.1016/j.jclinepi.2008.04.002.

Battini M, Barbieri S, Vieira A, Stilwell G, Mattiello S. 2016. Results of testing the prototype of the AWIN welfare assessment protocol for dairy goats in 30 intensive farms in Northern Italy. Ital J Anim Sci. 15(2):283–293. doi: 10.1080/1828051X.2016.1150795.

Battini M, Agostini A, Mattiello S. 2019. Understanding cows' emotions on farm: are eye white and ear posture reliable indicators? Animals. 9(8):1–12. doi: 10.3390/ani9080477.

Battini M, Renna M, Giammarino M, Battaglini L, Mattiello S. 2021. Feasibility and reliability of the AWIN welfare assessment protocol for dairy goats in semi-extensive farming conditions. Front Vet Sci. 8:731927. doi: 10.3389/fvets.2021.731927.

Bennet EM, Alpert R, Goldstein AC. 1954. Communications through limited response questioning. Public Opin Q. 18:303–308. doi: 10.1086/266520.

Brenner H, Kliebsch U. 1996. Dependence of weighted kappa coefficients on the number of categories. Epidemiology. 7(2):199–202. doi: 10.1097/00001648-199603000-00016.

Buczinski S, Faure C, Jolivet S, Abdallah A. 2016. Evaluation of inter-observer agreement when using a clinical respiratory scoring system in pre-weaned dairy calves. N Z Vet J. 64(4):243–247. doi: 10.1080/00480169.2016.1153439.

Byrt T, Bishop J, Carlin JB. 1993. Bias, prevalence and Kappa. J Clin Epidemiol. 46(5):423–429. doi: 10.1016/0895-4356(93)90018-V.

Can E, Vieira A, Battini M, Mattiello S, Stilwell G. 2016. On-farm welfare assessment of dairy goat farms using animal-based indicators: the example of 30 commercial farms in Portugal. Acta Agriculturae Scandinavica A Anim Sci. 66(1):43–55. doi: 10.1080/09064702.2016.1208267.

Cicchetti A, Allison T. 1971. A new procedure for assessing reliability of scoring EEG sleep recordings. Am J EEG Technol. 11:101–109. doi: 10.1080/00029238.1971.11080840.

Cohen J. 1960. A coefficient of agreement for nominal scales. Educ Psychol Meas. 20:37–46. doi: 10.1177/001316446002000104.

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

Croyle SL, Nash CGR, Bauman C, LeBlanc SJ, Haley DB, Khosa DK, Kelton DF. 2018. Training method for animal-based measures in dairy cattle welfare assessments. J Dairy Sci. 101(10):9463–9471. doi: 10.3168/jds.2018-14469.

Czycholl I, Klingbeil P, Krieter J. 2019. Interobserver reliability of the animal welfare indicators welfare assessment protocol for horses. J Equine Vet Sci. 75:112–121. doi: 10.1016/j.jevs.2019.02.005.

De Rosa G, Grasso F, Pacelli C, Napolitano F, Winckler C. 2009. The welfare of dairy buffalo. Ital J Anim Sci. 8:103–116. doi: 10.4081/ijas.2009.s1.103.

De Rosa G, Grasso F, Winckler C, Bilancione A, Pacelli C, Masucci F, Napolitano F. 2015. Application of the Welfare Quality protocol to dairy buffalo farms: prevalence and reliability of selected measures. J Dairy Sci. 98(10):6886–6896. doi: 10.3168/jds.2015-9350.

DiCiccio TJ, Efron B. 1996. Bootstrap confidence intervals. Stat Sci. 11:189–228. doi: 10.1214/ss/1032280214.

Efron B. 1979. Bootstrap methods: another look at the jackknife. Ann Stat. 7:1–26. doi: 10.1007/978-1-4612-4380-9_41.

EFSA Panel on Animal Health and Welfare (AHAW). 2012. Statement on the use of animal-based measures to assess the welfare of animals. EFSA J. 10(6):1–29. doi: 10.2903/j.efsa.2012.2767.

Falotico R, Quatto P. 2010. On avoiding paradoxes in assessing inter-rater agreement. Ital J Appl Stat. 22:151–160.

Feinstein AR, Cicchetti DV. 1990. High agreement but low Kappa: I. the problems of two paradoxes. J Clin Epidemiol. 43(6):543–549. doi: 10.1016/0895-4356(90)90158-L.

Fleiss JL. 1971. Measuring nominal scale agreement among many raters. Psychol Bull. 76:378–382. doi: 10.1037/h0031619.

Giammarino M, Mattiello S, Battini M, Quatto P, Battaglini LM, Vieira ACL, Stilwell G, Renna M. 2021. Evaluation of inter-observer reliability of animal welfare indicators: which is the best index to use? Animals. 11(5):1–16. doi: 10.3390/ani11051445.

Gisev N, Pharm B, Bell JS, Chen TF. 2013. Interrater agreement and interrater reliability: key concepts, approaches, and applications. Res Social Adm Pharm. 9(3):330–338. doi: 10.1016/j.sapharm.2012.04.004.

Gwet KL. 2001. Handbook of inter-rater reliability - how to estimate the level of agreement between two or multiple raters. Gaithersburg (MD): STATAXIS Publishing Company.

Gwet KL. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol. 61(Pt 1):29–48. doi: 10.1348/000711006X126600.

Holley GW, Guilford JP. 1964. A note on the G-index of agreement. Educ Psychol Meas. 24:749–753. doi: 10.1177/001316446402400402.

Holsti OR. 1969. Content analysis for the social sciences and humanities. Reading (MA): Addison-Wesley.

Hubert L. 1977a. Nominal scale response agreement as a generalized correlation. Br J Math Stat Psychol. 30:98–103. doi: 10.1111/j.2044-8317.1977.tb00728.x.

Hubert L. 1977b. Kappa revisited. Psychol Bull. 84(2):289–297. doi: 10.1037/0033-2909.84.2.289.

Janson S, Vegelius J. 1978. On the applicability of truncated component analysis based on correlation coefficients for nominal scales. Appl Psychol Meas. 2:135–145. doi: 10.1177/014662167800200113.

Janson S, Vegelius J. 1982. The J-index as a measure of nominal scale response agreement. Appl Psychol Meas. 6:111–121. doi: 10.1177/014662168200600111.

Kraemer HC. 1979. Ramifications of a population model for K as a coefficient of reliability. Psychometrika. 44:461–472. doi: 10.1007/BF02296208.

Krippendorff K. 1970. Estimating the reliability, systematic error and random error of interval data. Educ Psychol Meas. 30:61–70. doi: 10.1177/001316447003000105.

Krippendorff K. 2011. Computing Krippendorff's alpha-reliability. Philadelphia (PA): Annenberg School for Communication. [accessed: 2023 Jul 19]. https://repository.upenn.edu/asc_papers/43.

Marasini D, Quatto P, Ripamonti E. 2016. Assessing the interrater agreement for ordinal data through weighted indexes. Stat Methods Med Res. 25(6):2611–2633. doi: 10.1177/0962280214529560.

Martin P, Bateson P. 2007. Measuring behaviour: an introductory guide. 3rd ed. Cambridge: Cambridge University Press.

McHugh ML. 2012. Interrater reliability: the kappa statistic. Biochem Med. 22(3):276–282. doi: 10.11613/BM.2012.031.

Munoz S, Bangdiwala SI. 1997. Interpretation of kappa and B-statistics measures of agreement. J Appl Stat. 24:105–112. doi: 10.1080/02664769723918.

Munoz C, Campbell A, Hemsworth P, Doyle R. 2017. Animal-based measures to assess the welfare of extensively managed ewes. Animals. 8(1):1–16. doi: 10.3390/ani8010002.

Nannarone S, Ortolani F, Scilimati N, Gialletti R, Menchetti L. 2024. Refinement and revalidation of the equine ophthalmic pain scale: r-EOPS a new scale for ocular pain assessment in horses. Vet J. 304:106079. doi: 10.1016/j.tvjl.2024.106079.

Navarro E, Mainau E, Manteca X. 2020. Development of a facial expression scale using farrowing as a model of pain in sows. Animals. 10(11):2113. doi: 10.3390/ani10112113.

Pedersen KS, Holyoake P, Stege H, Nielsen JP. 2011. Observations of variable inter-observer agreement for clinical evaluation of faecal consistency in pigs. Prev Vet Med. 98(4):284–287. doi: 10.1016/j.prevetmed.2010.11.014.

Quatto P. 2004. Un test di concordanza tra più esaminatori [Testing agreement among multiple raters]. Statistica. 1:145–151. doi: 10.6092/issn.1973-2201/28.

Scott WA. 1955. Reliability of content analysis: the case of nominal scale coding. Public Opin Q. 19:321–325. doi: 10.1086/266577.

Stevens SS. 1946. On the theory of scales of measurement. Science. 103(2684):677–680. doi: 10.1126/science.103.2684.677.

Thomsen PT, Baadsgaard NP. 2006. Intra- and inter-observer agreement of a protocol for clinical examination of dairy cows. Prev Vet Med. 75(1–2):133–139. doi: 10.1016/j.prevetmed.2006.02.004.

Thomsen PT, Munksgaard L, Tøgersen FA. 2008. Evaluation of a lameness scoring system for dairy cows. J Dairy Sci. 91(1):119–126. doi: 10.3168/jds.2007-0496.

Vieira A, Battini M, Can E, Mattiello S, Stilwell G. 2018. Inter-observer reliability of animal-based welfare indicators included in the animal welfare indicators welfare assessment protocol for dairy goats. Animal. 12(9):1942–1949. doi: 10.1017/S1751731117003597.

Welfare Quality®. 2009a. Welfare Quality® assessment protocol for pigs (sows and piglets, growing and finishing pigs). Lelystad (Netherlands): Welfare Quality® Consortium.

Welfare Quality®. 2009b. Welfare Quality® assessment protocol for poultry. Lelystad (Netherlands): Welfare Quality® Consortium.