animals

MDPI

Article

Evaluation of Inter-Observer Reliability of Animal Welfare Indicators: Which Is the Best Index to Use?

Mauro Giammarino [1], Silvana Mattiello [2], Monica Battini [2], Piero Quatto [3], Luca Maria Battaglini [4], Ana C. L. Vieira [5], George Stilwell [6] and Manuela Renna [7,*]

1 Department of Prevention, Asl TO3, Veterinary Service, Area Animal Sanity, 10045 Piossasco, Italy; mgiammarino@aslto3.piemonte.it
2 Department of Agricultural and Environmental Sciences—Production, Landscape, Agroenergy, University of Milan, 20133 Milan, Italy; silvana.mattiello@unimi.it (S.M.); monica.battini@unimi.it (M.B.)
3 Department of Economics, Management and Statistics, University of Milan-Bicocca, 20126 Milan, Italy; piero.quatto@unimib.it
4 Department of Agricultural, Forest and Food Sciences, University of Turin, 10095 Grugliasco, Italy; luca.battaglini@unito.it
5 Centre for Management Studies of Instituto Superior Técnico (CEG-IST), University of Lisbon, 1049-001 Lisbon, Portugal; ana.lopes.vieira@tecnico.ulisboa.pt
6 Department of Veterinary Medicine, University of Lisbon, 1300-477 Lisbon, Portugal; stilwell@fmv.ulisboa.pt
7 Department of Veterinary Sciences, University of Turin, 10095 Grugliasco, Italy
* Correspondence: manuela.renna@unito.it; Tel.: +39-011-670-8023

Simple Summary: In order to be effective, on-farm welfare assessment protocols should always rely on reliable, as well as valid and feasible, indicators. Inter-observer reliability refers to the extent to which two or more observers are observing and recording data in the same way. The present study focuses on the problem of assessing inter-observer reliability in the case of dichotomous (e.g., yes/no) welfare indicators and the presence of two observers, in order to decide about the inclusion of indicators in welfare assessment protocols. We compared the performance of the most popular currently available agreement indexes. Some widely used indexes showed their inappropriateness to evaluate the inter-observer reliability when the agreement between observers was high. Other less used indexes, such as Bangdiwala's B or Gwet's $\gamma(AC_1)$, were found to perform better and are therefore suggested to assess the inter-observer reliability of dichotomous indicators.

Abstract: This study focuses on the problem of assessing inter-observer reliability (IOR) in the case of dichotomous categorical animal-based welfare indicators and the presence of two observers. Based on observations obtained from Animal Welfare Indicators (AWIN) project surveys conducted on nine dairy goat farms, and using udder asymmetry as an indicator, we compared the performance of the most popular agreement indexes available in the literature: Scott's π, Cohen's k, k_{PABAK}, Holsti's H, Krippendorff's α, Hubert's Γ, Janson and Vegelius' J, Bangdiwala's B, Andrés and Marzo's Δ, and Gwet's $\gamma(AC_1)$. Confidence intervals were calculated using closed formulas of variance estimates for π, k, k_{PABAK}, H, α, Γ, J, Δ, and $\gamma(AC_1)$, while the bootstrap and exact bootstrap methods were used for all the indexes. All the indexes and closed formulas of variance estimates were calculated using Microsoft Excel. The bootstrap method was performed with R software, while the exact bootstrap method was performed with SAS software. k, π, and α exhibited a paradoxical behavior, showing unacceptably low values even in the presence of very high concordance rates. B and $\gamma(AC_1)$ showed values very close to the concordance rate, independently of its value. Both bootstrap and exact bootstrap methods turned out to be simpler compared to the implementation of closed variance formulas and provided effective confidence intervals for all the considered indexes. The best approach for measuring IOR in these cases is the use of B or $\gamma(AC_1)$, with bootstrap or exact bootstrap methods for confidence interval calculation.

Keywords: agreement index; animal-based measure; dichotomous categorical indicator; inter-rater reliability

check for updates

Citation: Giammarino, M.; Mattiello, S.; Battini, M.; Quatto, P.; Battaglini, L.M.; Vieira, A.C.L.; Stilwell, G.; Renna, M. Evaluation of Inter-Observer Reliability of Animal Welfare Indicators: Which Is the Best Index to Use? *Animals* **2021**, *11*, 1445. https://doi.org/10.3390/ani11051445

Academic Editor: Angelo Peli

Received: 25 March 2021
Accepted: 10 May 2021
Published: 18 May 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Animal-based indicators for the assessment of animal welfare need to meet three essential requirements: validity, feasibility, and reliability [1]. The concept of reliability is closely related to the concept of assessment reproducibility, whether the same observer expresses a measure at different times (intra-observer reliability) or whether there are multiple observers to express the measure at the same moment (inter-observer reliability, IOR). The IOR is a fundamental attribute for reliable welfare assessments, especially when the evaluation is carried out using animal-based indicators, which may be associated with a certain level of subjectivity, and biased by the assessors' previous experience and level of empathy with the animals [2]. However, in animal welfare and behavioral studies, IOR is frequently neglected due to different reasons (e.g., debate on the particular type of statistic to be used, difficulties in involving multiple observers) [3].

While the term "agreement" means the measure of concordance between observers (concordance rate, P_o), "reliability" is what we would like to infer from the agreement [4]. Reliability measures the concordance between observers, net of chance agreement [5]. If reliability is low, the indicator is inappropriate and should be redefined, ensuring good data recording and/or better training of the observers [6]. The reliability of animal welfare indicators should be bias-free and, furthermore, the indexes should be robust. The reproducibility is the most important interpretation of reliability [4], and it is necessary that the agreement estimates can ensure the reproducibility of judgments. The need to ascertain the agreement between observers, beyond the agreement due to chance, implies the possibility of having reliable statistical methods for assessing the quality of measurements [7].

According to Krippendorff [4], an agreement coefficient can become an index of reliability only if: (i) it is applied to proper reliable data, (ii) it treats units of analysis as separately describable or categorizable without presuming any knowledge about the correctness of their descriptions or categories (absence of gold standard), and (iii) its values correlate with the conditions under which one is willing to rely on imperfect data.

To our knowledge, only a few studies have been specifically designed to test the IOR of animal welfare indicators [8–14]. For this purpose, the most frequently used agreement index has been Cohen's k [5]. Some recent reviews, aimed at identifying promising indicators for welfare assessments in ruminants, confirmed that reliability, and particularly IOR, has been scarcely investigated so far [1,15,16], and highlighted the need for further investigation of this issue. One explanation could be that, although the literature is rich in agreement indexes, the problem of finding the best one for different application contexts has not yet been solved [17]. According to Ato et al. [18], all agreement indexes for categorical data can be traced back to three distinct approaches. The first approach, the most widely used in the literature, dates back to Scott's intuition [19] of having to correct the agreement, described as a percentage of concordant cases out of the total number of observed cases, by eliminating the concordance due to chance. The π index [19], the σ index [20], the k index [5], and the $\gamma(AC_1)$ index [21] belong to this approach. Loglinear modeling is the second approach, which aims to analyze agreement and disagreement patterns by accounting for the discrepancies between the data and expected values, under the hypothesis of independence [22]. Loglinear models generalize to a mixture model by including an unobserved categorical latent variable [18]. In such models, the population is composed of two subpopulations (latent classes): the subpopulation of objects easy to classify by both observers (systematic agreement) and the subpopulation of objects difficult to classify (random agreement and disagreement). This approach allowed Aickin [23] to define a new measure of agreement called the α-coefficient. The third approach is inspired by the traditional multiple-choice test, which allowed Andrés and Marzo [24] to define the Δ index, under the assumption that each observer can choose only one of N possible answers for each object to be evaluated.

The aims of this study are to compare the most popular agreement indexes, as to ascertain the best practice for measuring the agreement between two observers, and to calculate the related confidence intervals when evaluating dichotomous categorical animal-

based welfare indicators. To do so, we selected one dichotomous animal-based indicator from the Animal Welfare Indicators (AWIN) welfare assessment protocol for goats [25,26], namely the udder asymmetry, and we used it as an example to test the performance of the different considered agreement indexes.

2. Materials and Methods

2.1. Dataset

The AWIN protocol for welfare assessment [25,26] was applied by two observers in 10 Italian intensive (AWIN prototype protocol applied from Feb to Jun 2014 [12]), 10 Portuguese intensive (AWIN prototype protocol applied from Jan to Apr 2014 [12]), and 13 Italian extensive (AWIN protocol adapted to extensive conditions applied from Apr to Jul 2019; unpublished data) dairy goat farms. Both assessors were students of the second year of the MSc in Animal Science at the University of Turin (Grugliasco, Italy). Assessor A also had an MSc in Veterinary Science and in Biostatistics, had worked as a veterinarian in the Public Health Service, and had more than 10 years of experience with dairy goats. Assessor B had no specific experience with dairy goats. Before the beginning of the study, both assessors received a common 1-day training including both theoretical and practical sessions, and received the AWIN protocol [25] as training material. The training was provided by two authors of the AWIN welfare assessment protocol for goats kept in intensive or semi-intensive production systems [26].

The collected data were preliminarily analyzed to identify a dichotomous variable that presented a wide variation of concordance rate between two observers. Among the six dichotomous categorical individual animal-based welfare indicators included in the AWIN protocol for goats (i.e., fecal soiling, nasal discharge, ocular discharge, severe lameness, overgrown claws, and udder asymmetry), udder asymmetry was chosen to test different methods for assessing IOR, because it was the variable where we observed the widest variability of agreement between observers in the visited farms. According to the AWIN protocol, during the assessment, each goat was assigned to one of two mutually exclusive and exhaustive categories: presence of asymmetry = 1; absence of asymmetry = 0. The presence of udder asymmetry was confirmed when one half of the udder was at least 25% longer than the other, excluding the teats [26].

To perform our analysis, we used data collected from nine farms (out of the initially considered 33 farms), which represented the whole range of variability (75 to 100%) in terms of agreement between observers: seven Italian intensive farms (from I-IT1 to I-IT7), one Portuguese intensive farm (I-PT1), and one Italian extensive farm (E-IT1).

2.2. Agreement Measures

A crude measure of reliability (P_o) is given by the proportion of concordant cases (agreement) out of the total observed cases. However, this measure is distorted in favor of situations with fewer categories [19] and does not account for the chance agreement.

The IOR of udder asymmetry was evaluated measuring the most popular agreement indexes currently available in the literature: Scott's π [19], Cohen's k [5], k_{PABAK} and the related indexes (σ [20], G [27], and S [28]), Holsti's H [29], Krippendorff's α [30], Hubert's Γ [31], Janson and Vegelius' J [32], Bangdiwala's B [33], Andrés and Marzo's Δ [24], and Gwet's $\gamma(AC_1)$ [21]. A detailed description of each considered agreement index is presented in Appendix A.

2.3. Confidence Intervals for Agreement Indexes

Closed formulas of variance estimates are available for almost all the considered agreement indexes. The application of such formulas is handy for some indexes, including π [19], k and k_C [5], k_{PABAK} and the related indexes (σ [20], G [27], and S [28]), Holsti's H [29], α [30], Γ [31], J [32], Δ [24], and $\gamma(AC_1)$ [21]. A detailed description of the applied formulas of variance estimates for the above-mentioned indexes is presented in Appendix B.

Closed formulas of variance estimates are instead cumbersome for the *B* index. Therefore, for such index, we used confidence intervals based on the bootstrap method [34] and the exact bootstrap method for small samples [35]. Bootstrapping is a general method for estimating the distribution of a given statistic by resampling with the replacement of the data set at hand [34]. The bootstrap procedure uses such an empirical distribution as a substitute for the true distribution in order to provide variance estimates and confidence intervals. A criticism of the standard bootstrap procedure is that different observers may reach, by chance, different conclusions [35]. The exact bootstrap method prevents the possibility of different conclusions. This method was proposed for Cohen's *k* when the proportion of agreement was high, and the sample size *n* was small ($\leq$200), but it was never applied to other agreement indexes so far. The exact bootstrap method attributes the probability $1/n$ to each element of a small population with size *n*, so that we can extract with replacement $n\hat{}n$ samples from the population, which allows providing $n\hat{}n$ values of the considered agreement index, whose empirical distribution is known as the exact bootstrap distribution [35]. In particular, 95% bootstrap and exact bootstrap confidence intervals can be constructed by the percentile method, which employs the 2.5th and 97.5th percentiles of the bootstrap and exact bootstrap distribution, respectively [35,36].

2.4. Statistical Analyses

Microsoft Excel (2010) was used to calculate the index values (using the formulas reported in Appendix A) and their confidence intervals (using the closed formulas of variance estimates reported in Appendix B). For bootstrapping, the following packages of the R software (v. 3.5.2; R Core Team, Wien, Austria, 2018) were used: "raters" [37], "vcd" [38] and "bootstrap" [39]. The SAS software (v. 9.0; SAS Institute Inc., Cary, NC, USA) was used for the exact bootstrap method, using the script reported by Klar et al. [35] for Cohen's *k*; the scripts were modified by adapting them to the all the other considered agreement indexes.

3. Results

3.1. Agreement Measures

Three hundred and eighty-eight dairy goats were examined in the nine selected farms. The frequency of cases for the indicator "udder asymmetry" in each of the nine selected farms is reported in the agreement tables (Table S1).

For each farm, Table 1 shows the values expressed by the considered agreement indexes for the AWIN indicator "udder asymmetry".

As expected, the *H* index coincided with the concordance rate (P_o). The *k* index and α index on the one hand, and the Γ index and *J* index on the other hand, showed the same values. The π index, *k* index, and α index expressed unacceptably low values, even in the presence of high concordance rates (e.g., farms I-IT2, I-IT5, and I-IT7). When the concordance between observers was perfect, and cell *n11* of the agreement table (Table S2) showed a value equal to zero, π index, *k* index, and α index did not express any value. When the concordance was not perfect for a single or few objects, and cell *n22* showed a value equal to zero (farms I-IT5 and I-IT7), Cohen's *k* and Scott's π showed value zero or a negative value since one of the marginals relating to the probability table was zero.

The distance from the concordance rate of the values expressed by the k_{PABAK} index (the values of which coincided with those of the related σ, G, and S indexes) gradually decreased as the concordance rate increased (Table 1) until it expressed the value 1 to perfect concordance.

Table 1. Values of the agreement indexes for the AWIN animal-based welfare indicator "udder asymmetry" for the nine selected dairy goat farms, sorted by increasing concordance rate (P_0).

	Agreement Index [1]											
Farm	P_o [2]	π	k	k_c	k_{PABAK} [3]	H	α	Γ	J	B	Δ	$\gamma(AC_1)$
E-IT1	75	0.15	0.16	0.23	0.51	75	0.15	0.25	0.25	0.70	0.52	0.65
I-IT1	77	0.24	0.24	0.24	0.54	77	0.24	0.28	0.30	0.71	0.54	0.68
I-IT2	88	0.27	0.27	0.43	0.77	88	0.28	0.58	0.58	0.87	0.79	0.86
I-IT3	92	0.55	0.55	0.55	0.84	92	0.56	0.69	0.70	0.90	0.84	0.90
I-IT4	95	0.64	0.64	1.00	0.89	95	0.64	0.79	0.79	0.94	0.95	0.94
I-IT5	95	−0.02	0.00	0.00	0.90	95	−0.01	0.80	0.81	0.95	0.95	0.95
I-IT6	97	0.78	0.78	1.00	0.93	97	0.78	0.87	0.87	0.96	0.97	0.96
I-IT7	97	−0.02	0.00	0.00	0.93	97	0.00	0.87	0.87	0.97	0.97	0.96
I-PT1	100	1.00	1.00	1.00	1.00	100	1.00	1.00	1.00	1.00	1.00	1.00

Abbreviations: E, extensive; I, intensive; IT, Italian; PT, Portuguese. [1] π [19]; k [5]; k_c [5]; H [29]; α [30]; Γ [31]; J [32]; B [33]; Δ [24]; $\gamma(AC_1)$ [21]. [2] Concordance rate (P_o, %), calculated as: $(n_{11} + n_{22})/N$. [3] The related indexes (σ index [20], G index [27], and S index [28]) gave the same results.

The Δ index showed an intermediate behavior between the k_{PABAK}, with which it shared the values when the concordance rate ranged from 75 to 92%, and the B index, the values of which were similar to those expressed by the Δ index at higher concordance rates (95 to 100%). The distances between the values expressed by the Δ index and the concordance rate were wider at medium-high values of concordance rate (75 to 92%), but soon they decreased, and Δ index coincided with the concordance rate in the case of higher concordance rates (95 to 100%; Table 1).

The Γ index expressed low values especially when the concordance rate was equal to 75 and 77% (farms E-IT1 and I-IT1). The distances from the values of P_o were high, up to values of 97% of the concordance rate (Table 1).

The B index showed values very close to those of the concordance rate in all the cases examined in this study. When the concordance rate showed its minimum (75%; farm E-IT1), the B index showed the highest value among the values presented by the analyzed indexes (Table 1). The B index values were always very close to those of the observed concordance rate until they early coincided with them (when P_o = 88%, B index = 0.87; farm I-IT2). The Bangdiwala's observer agreement chart (Table S1) graphically represents the B index, providing an immediate and very useful visual representation of the obtainable results.

The $\gamma(AC_1)$ index expressed almost the same values of the B index, with the exception of cases with medium-high values of concordance rate (75 and 77%; farms E-IT1 and I-IT1), when the $\gamma(AC_1)$ index showed lower values than the B index.

3.2. Confidence Intervals for Agreement Indexes

Figure 1 shows the boxplot of the values obtained for each considered agreement index with the bootstrap method and the exact bootstrap method for the nine selected farms.

The best performing indexes are expressed by values closer to the concordance rate (that coincided with Holsti's H) and by narrower confidence intervals. For all the considered indexes and for all the farms, we observed a substantial overlapping of confidence intervals results when implemented with the bootstrap and exact bootstrap methods. The inadequacy of the values expressed by Cohen's k and Scott's π is evident in the case of low concordance rates (farms E-IT1, I-ITI, I-IT2, I-IT5, and I-IT7). In all the cases, confidence intervals ranges were wide for π and k indexes, even when no paradox effect was observed. In almost all cases, the Γ e k_{PABAK} indexes also showed wider ranges of confidence intervals when compared to the other considered agreement indexes. The exact bootstrap method expressed confidence intervals for π and k indexes even when cell *n22* of the agreement table showed a value equal to zero (Figure 1, boxplots for farms I-IT5 and I-IT7). The boxplots also graphically highlight the paradox effect (farms E-IT1, I-IT1 and I-IT2).

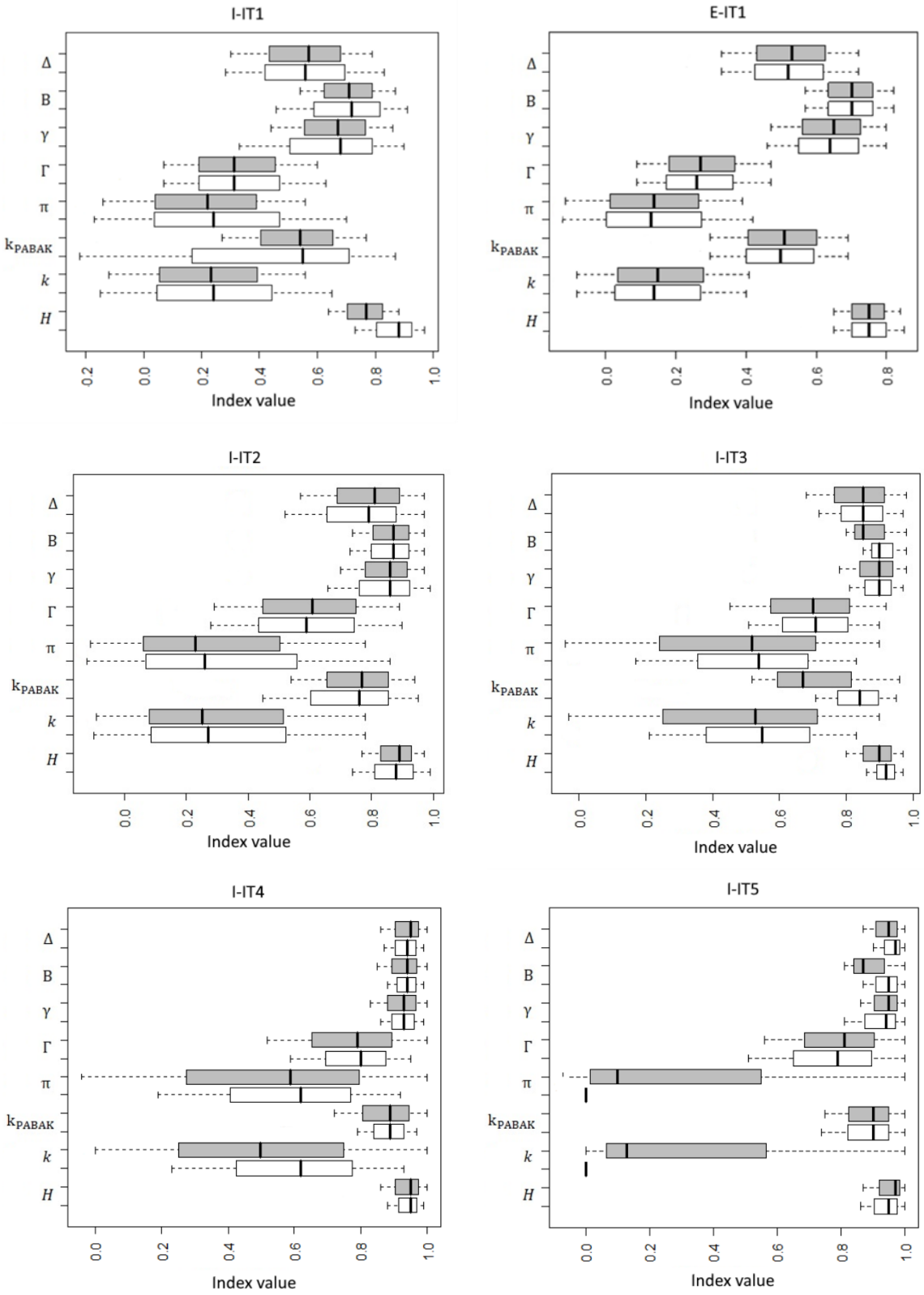

Figure 1. *Cont*.

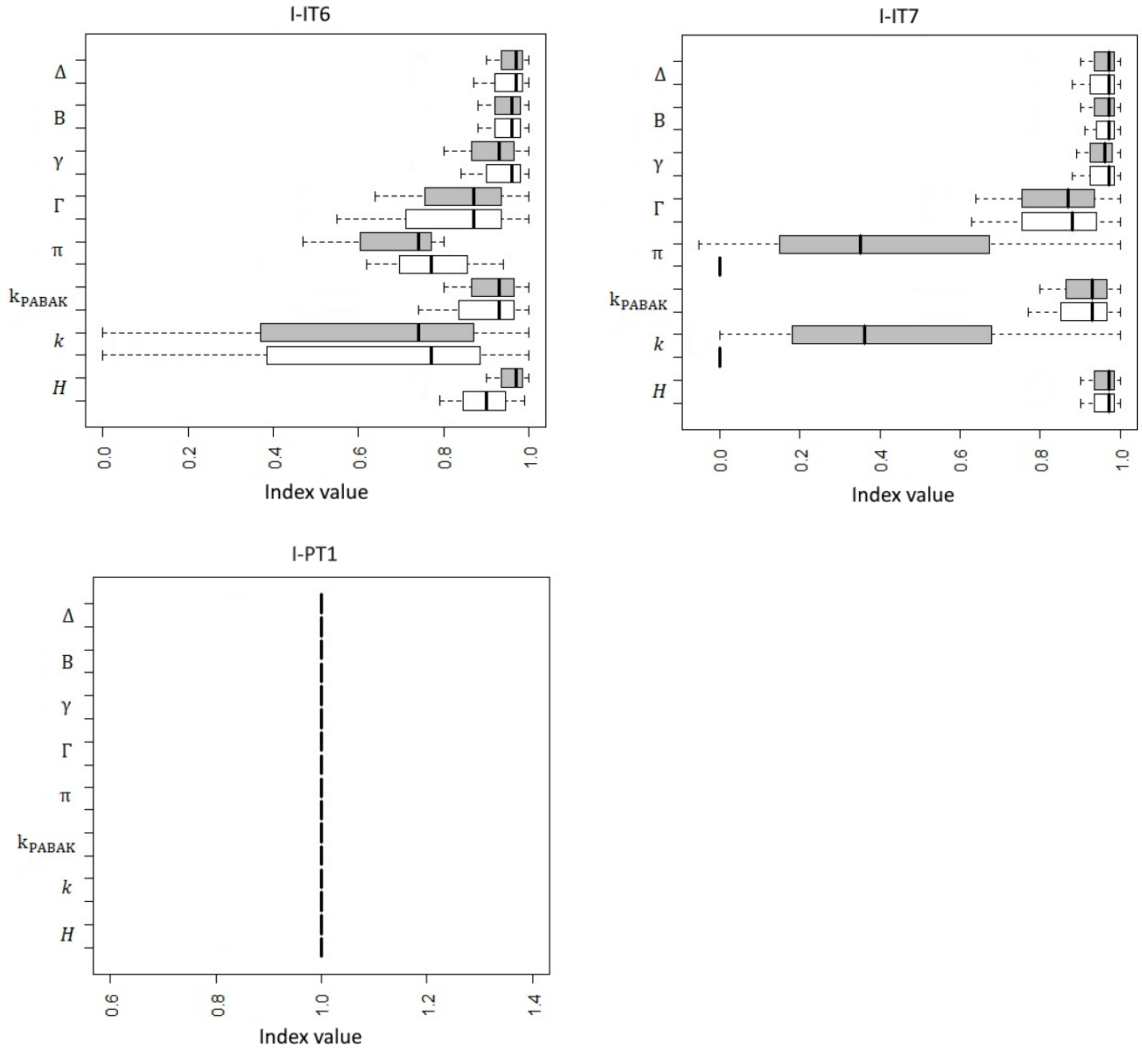

Figure 1. Boxplot of the agreement values obtained for each index with the bootstrap method and the exact bootstrap method for all the selected farms (I-IT1, E-IT1, I-IT2, I-IT3, I_IT4, I_IT5, I_IT6, I-IT7, I-PT1). Legend: Δ = Δ index; B = *B* index; γ = $\gamma(AC_1)$ index; Γ = Γ index; π = π index; k_{PABAK} coincided with the related indexes: σ index, G index and S index; *k* = *k* index; *H* = *H* index; grey = bootstrap method; white = exact bootstrap method. The α index is not reported in the figure as it coincided with Cohen's *k*. The *J* index is not reported in the figure as it coincided with Hubert's Γ.

For each index, confidence intervals calculated using closed formulas of variance estimates and both the bootstrap and exact bootstrap methods are presented for the nine selected dairy goat farms in Table S1.

4. Discussion

From the results obtained in our study, it is evident that, when evaluating IOR, the choice of the agreement index is very subtle and becomes crucial in order to validate the method of evaluating welfare indicators. The paradoxical behavior of Cohen's *k*, Scott's π, and Krippendorff's α makes it difficult to use these indexes without a careful critical analysis of the results. For this reason, it is recommended to use other indexes that are not affected by the same paradox effect. For the evaluation of IOR in the case of dichotomous categorical indicators and the presence of two observers, Bangdiwala's *B* and Gwet's $\gamma(AC_1)$ were found to be the most appropriate indexes to be used.

When trying to find an adequate approach to evaluate the IOR of animal-based welfare indicators, it is very common to get lost within the array of different concepts and methods. Furthermore, it is common to find criticisms of different order for each method, which makes the selection even more difficult. In this work, we aimed to clear this on-going

discussion by focusing on dichotomous categorical animal-based welfare indicators in the presence of two observers. The literature shows the limitations of the method of calculating the agreement between observers by the proportion of concordant cases out of the total cases, without taking into account the concordance due to chance [40]. The same criticism involves the *H* index that, as expected, was unable to calculate the agreement by chance [41].

To evaluate the IOR, some authors used the approach based on the χ^2 test, calculated from a cross-classification table, or the approach based on correlation coefficients. However, both approaches appear unsuitable and, consequently, they were not implemented in this study. The χ^2 test measures the degree of independence between variables that does not necessarily coincide with concordance. In fact, the association measures calculate the deviation from chance contingencies between variables [4]. Therefore, the χ^2 statistic presents high values for any deviation from the association due to chance, both in case of agreement and in case of disagreement [40]. Similarly, the use of correlation coefficients that measure deviations from linearity is also discouraged because correlation and concordance are not the same [42]. According to Krippendorff [4], a valid index measures agreements or disagreements among multiple descriptions generated by a single coding procedure, regardless of who enacts the procedure.

Cohen [5] proposed the *k* index as an extension of Scott's π [19], which in defining the rate of agreement due to chance involves the knowledge of rate distributions for both observers. It assumes that the two observers are interchangeable, so that the marginal distributions are identical and hence the two indexes of Cohen and Scott are equivalent [40]. Although the *k* index is still the most widely used agreement index [43], in some circumstances where the concordance rate is very high, it shows unacceptably low values. Such a paradoxical behavior of Cohen's *k* is well studied in the literature [44,45]. To overcome this problem, Byrt et al. [46] proposed two diagnostics for *k* given by $BI = (n_{12} - n_{21})/N$ (bias index) and $PI = (n_{11} - n_{22})/N$ (prevalence index): BI is zero when the marginal distributions are equal and PI is zero when the categories are equally likely [47]. However, all this would make the reading of the value less immediate and the interpretation of the index more difficult. This is the reason why Byrt's diagnostics were not implemented in our study. Our results confirm the paradoxical behavior of the *k* index, as it showed unacceptably low values even in the presence of very high concordance rates. Landis and Koch [48] suggested different ranges of values for the *k* index: values higher than 0.74 indicate excellent agreement; values between 0.40 and 0.75 indicate a good agreement; and values less than 0.40 indicate a poor agreement. However, the same authors claimed that every categorization is arbitrary. In Table S1, where the concordance rate is 75%, *k* index (0.16) demonstrated a slight agreement according to the benchmarking proposed by Landis and Koch [48] and a marginal agreement according to the benchmarking of Fleiss [49]. This is also evident in Table 1 for farms I-IT1 (*k* index = 0.24; P_o = 77%), I-IT2 (*k* index = 0.27; P_o = 88%), and I-IT3 (*k* index = 0.55; P_o = 92%). Paradoxical behaviors are also evident in Vieira et al. [12], where a concordance rate of 92.42% corresponded to a mediocre value of the *k* index (0.44). For this reason, the *k* index cannot be considered adequate to analyse the IOR in the case of dichotomous categorical animal-based welfare indicators (such as the udder asymmetry evaluated in our study), for which the concordance between observers is presumed to be very high, even close to 100% in some cases [12]. More precisely, the paradox of the *k* index is twofold. The first paradox occurs when the marginal totals are highly unbalanced in a symmetrical way (e.g., farm E-IT1; Table S1), producing high values of P_e. The second paradox, not observed in our study but reported in the literature, appears when the marginal totals are asymmetrically balanced, producing values which cannot be high [44]. The k_M version proposed by Cohen [5] does not seem to avoid the two types of paradox [44]. Cicchetti and Feinstein [50] suggested tackling the paradox by adopting two indexes to account for the two paradoxes. We agree with Brennan and Prediger [51] that the indiscriminate use of the *k* index can be misleading and that other statistics may be more meaningful in some cases. Other authors [12] tried to overcome this paradox by presenting,

simultaneously, information on the overall agreement together with positive and negative agreement, and the prevalence of the indicator. However, even if this presents the reader with all the information for analysis, it puts an extra cognitive burden on whomever is analyzing the data, which can hinder its interpretation. For this reason, further research on the topic that assists in overcoming this drawback is needed.

The α index [30] assumes values very close to the k index [5], as they belong to the same approach. This is also confirmed by the results obtained in the current study, where the two indexes showed exactly the same values for all the nine considered farms. From our results, it seems that the α index suffers from the same paradoxical behavior as Cohen's k, as previously reported by Zhao [52] and Gwet [53].

From the analysis of our results, it appears evident that also the π index suffers the same paradoxical behavior seen for the k index, which represents an extension of π (see for example farms E-IT1, I-IT1, and I-IT2, where the values of the indexes are very far from P_o). In an interesting comparative publication of several indexes for 2 × 2 tables [18], both π index and k index produced very high distortions at extreme prevalence values and were shown to be the least well-performing indexes.

The k_{PABAK} does not show the paradox effects [47], as confirmed by the results obtained in our study. In the work of Ato et al. [18], the σ index [20] was considered as an unbiased index that presented an excellent behavior for 2 × 2 tables. The S index [28] also allows measuring the level of inter-rater agreement without incurring the paradoxes of the k index [54]. The G index has reasonably small biases for estimating the "true" IOR [21].

The Δ index [24] has also proven to be reliable in this study, confirming previous results obtained by Ato et al. [18].

The B index [33] showed the highest value among all considered indexes when the concordance rate attained the minimum value (75%) (farm E-IT1). However, at very high concordance rates, it gave the same values as the Δ index (farm I-IT7). If only one of the diagonal cells of the agreement table (Table S2) exhibits agreement, the B index equals P_o. In addition, the Bangdiwala's observer agreement chart (Table S1) represents an immediate and useful tool that does not suffer from the paradox effect [47] and is easily obtainable with the PROC FREQ of the SAS program [55] or by the "vcd" package of the R program [38].

The $\gamma(AC_1)$ index [21] is recommended [56,57], even if it is not widely adopted [17] because it is little known. In particular, it also equals P_o when the concordance is present in only one of the diagonal cells of the agreement table [47].

In order to provide confidence intervals, the bootstrap and exact bootstrap methods turned out to be simpler when compared to the implementation of closed variance formulas and, in particular, the exact bootstrap method is easily executable in SAS [35].

5. Conclusions

When evaluating dichotomous categorical animal-based welfare indicators, and particularly in the case of a high concordance rate, the optimal practice for measuring the IOR between two observers is the use of the B index [33] or the $\gamma(AC_1)$ index [21], as they are not affected by paradoxical behaviors. Both the bootstrap and exact bootstrap methods are easier to be executed when compared to closed formula of variance estimates and provide effective confidence intervals for all the considered agreement indexes, including B and $\gamma(AC_1)$. Our study also clearly demonstrates that the exact bootstrap is a valid method for the calculation of confidence intervals not only for the π index and k index, as already reported in the published literature, but for all the tested agreement indexes.

Our results can be extended to any welfare assessment protocol (e.g., other species or different contexts of application) when two independent observers test dichotomous variables at the same time. Further studies are needed to find the best practice to assess IOR for other types of variables (e.g., trichotomic and four-level variables), also in the presence of more than two observers.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/ani11051445/s1, Table S1: Motivation examples, Table S2: Agreement table.

Author Contributions: Conceptualization, M.G., S.M., M.B., P.Q. and M.R.; methodology, M.G. and P.Q.; validation, M.G. and P.Q.; formal analysis, M.G.; investigation, M.B., P.Q. and A.C.L.V.; data curation, M.G.; writing—original draft preparation, M.G. and M.R.; writing—review and editing, S.M., M.B., P.Q., L.M.B., A.C.L.V., G.S. and M.R.; supervision, S.M., P.Q. and G.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki. Ethical approval was not necessary for this study, as no animal manipulation occurred.

Informed Consent Statement: Written informed consent has been obtained from the observers to publish this paper.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors would like to thank the farmers who allowed us to visit their farms. Fulvia Troja and Valentina Pitruzzella are also acknowledged for the application of the adapted AWIN protocol in extensive conditions.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

The considered agreement indexes are presented here in the chronological order by which they were developed.

Appendix A.1. π Index

$$\pi = \frac{P_o - P_e}{1 - P_e} \tag{A1}$$

where [19]:

P_o is the rate of observed concordance and represents the rate of concordant judgments of two independent observers who analyze the same dataset;
P_e is the rate of the expected agreement due to chance given by:

$$P_e = \sum_{i=1}^{M} p_i^2 \tag{A2}$$

where:

M is the number of categories;
p_i is the proportion of objects assigned to the i-th category.

This index varies from 0 (no agreement) to +1 (perfect agreement).

Appendix A.2. k and k_C Indexes

$$k = \frac{\sum P_{ii} - \sum P_{i.} P_{.i}}{1 - \sum P_{i.} P_{.i}} \tag{A3}$$

where [5]:

$\sum P_{ii}$ is the observed hit rate, denoted by P_o;
$\sum P_{i.} P_{.i}$ is the proportion of agreement due to chance, denoted by P_e. Hence, the formula can be summarized as:

$$k = \frac{P_o - P_e}{1 - P_e} \tag{A4}$$

The assumptions for k are the following [5,51]:

(a) The N objects categorized are independent;

(b) The categories are independent, mutually exclusive, and exhaustive;
(c) The assigners operate independently.

By examining the formula, it can be seen that Cohen [5] standardizes the difference between the observed agreement and the expected agreement, dividing it by the difference between the maximum value of k and the amount of agreement due to chance. In general, k assumes values between $-P_e/(1-P_e)$ and 1. The maximum value is reached only if the values outside the diagonal of the agreement table (Table S2) are zero and the marginal totals of the two observers are equal. However, when the marginal totals are asymmetric (as it happens very often), the maximum value of k will never be 1. To deal with this issue, Cohen [5] suggested the k maximum value:

$$k_M = \frac{P_{oM} - P_e}{1 - P_e} \tag{A5}$$

where:

P_{oM} is the maximum observed proportion, obtained by adding the minimum values of the individual marginal totals.

Cohen [5] estimated the k correct (k_C) dividing k by k_M:

$$k_C = \frac{k}{k_M} \tag{A6}$$

These indexes vary from 0 (no agreement) to +1 (perfect agreement).

Appendix A.3. k_{PABAK}

Many authors proposed an adjusted k given by:

$$k_{PABAK} = 2 * P_o - 1 \tag{A7}$$

where:

P_o is the concordance rate.

Despite being characterized by different formulas, the σ index [20], the G index [27], and the S index [28] are traced back to this criterion of correction.

This index varies from −1 (no agreement) to +1 (perfect agreement).

Appendix A.4. H Index

$$H = P_o = \frac{2 * C}{N_A + N_B} \tag{A8}$$

where [29]:

C is the number of concordant judgments;
N_A is the number of judgments of the observer A;
N_B is the number of judgments of the observer B.

This index is expressed as a percentage and varies from 0 (no agreement) to 100% (perfect agreement).

Appendix A.5. α Index

The index aggregates disagreements observed on individual units and compares such an observed agreement to the agreement expected by chance:

$$\alpha = \frac{P_o - P_e}{1 - P_e} \tag{A9}$$

where [30]:

$$P_o = (n_{11} + n_{22})/n \tag{A10}$$

while

$$\hat{P}_e = \left(\frac{2m_1}{2n}\right)\left(\frac{2m_1 - 1}{2n - 1}\right) + \left(\frac{2m_2}{2n}\right)\left(\frac{2m_2 - 1}{2n - 1}\right) \tag{A11}$$

with

$$m_1 = \frac{n_{+1} + n_{1+}}{2} \text{ and } m_2 = \frac{n_{+2} + n_{2+}}{2} \tag{A12}$$

This index varies from 0 (no agreement) to +1 (perfect agreement).

Appendix A.6. Γ Index

This index [29] was proposed as a particular case of Kendall's generalized correlation coefficient [32,58]:

$$\Gamma = 1 - 4 * \frac{(n_{11} + n_{22})(n_{12} + n_{21})}{n(n-1)} \tag{A13}$$

When all the cells are equal, this index does not satisfy the reasonable requirement of being null [32].

This index varies from 0 (no agreement) to +1 (perfect agreement).

Appendix A.7. J Index

This index [32] was introduced as a correlation coefficient for nominal scale and represents a particular case of Kendall's general coefficient [58]:

$$J = \frac{[(n_{11} + n_{22}) - (n_{12} + n_{12})]^2}{n^2} \tag{A14}$$

The *J* index is closely related to Hubert's Γ, but it does not have the problem above-mentioned for Γ [59].

This index varies from 0 (no agreement) to +1 (perfect agreement).

Appendix A.8. B Index

The *B* index [33] summarizes the so-called "observer agreement chart", in which the degree of agreement is represented by the proportion of the two dark square areas shown in Table S1:

$$B = \frac{\sum_{i=1}^{k} a_{ii}^2}{\sum_{i=1}^{k} n_{i.} n_{.i}} \tag{A15}$$

where:

a_{ii}^2 is the square of the values of the concordant cells;
$n_{i.}$ is the total of *i*-th row;
$n_{.i}$ is the total of *i*-th column.

Being a proportion of areas, this index also varies from 0 (no agreement) to +1 (perfect agreement) [47].

Appendix A.9. Δ Index

Cohen's *k* may have a paradoxical behavior when marginal distributions are asymmetric [45]. Andrés and Marzo [24] proposed the so-called Δ index, which aims to be independent of marginal distributions. For 2 × 2 tables, these authors suggested an asymptotic approximation of the index, which can be used as a consistent measure of concordance [24]:

$$\Delta = p_{11} + p_{22} - 2\sqrt{p_{12}p_{21}} \tag{A16}$$

This index varies from 0 (no agreement) to +1 (perfect agreement).

Appendix A.10. $\gamma(AC_1)$ Index

To avoid the k paradoxical behavior, Gwet [21] proposed the coefficient of agreement $\gamma(AC_1)$ [21]:

$$\gamma(AC_1) = \frac{P_o - P_e^*}{1 - P_e^*} \in [-1, 1] \tag{A17}$$

where:

P_o is estimated with $P_o = (n_{++} + n_{--})/n$;
P_e^* is estimated as:

$$P_e^* = 2\pi_+(1 - \pi_+) \tag{A18}$$

with

$$\begin{gathered} \pi_+ = \left(p_{Oss1|0} + p_{Oss2|0}\right)/2, \\ p_{Oss1|0} = n_{Oss1|0}/n \text{ and } p_{Oss2|0} = n_{Oss2|0}/n. \end{gathered} \tag{A19}$$

This index varies from 0 (no agreement) to +1 (perfect agreement).

Appendix B

Here follows a description of the applied closed formulas of variance estimates.

Appendix B.1. π Index

Scott [19] proposed the following formula for the variance of π index:

$$Var(\pi) = \left(\frac{1}{1 - p_e}\right)^2 \frac{p_o(1 - p_o)}{n - 1} \tag{A20}$$

Appendix B.2. k, k_C, and α Indexes

In order to determine whether k differs significantly from zero, Fleiss et al. [60] proposed a formula for an asymptotic approximation of the variance in the case of an $m \times m$ table. Under the hypothesis of the agreement occurring by chance, the asymptotic variance equals the exact variance proposed by Everitt [61] based on the hypergeometric distribution:

$$Var_0(k) = \frac{p_e + p_e^2 - \sum_{i=1}^{m} p_{i.}p_{.i}(p_{i.} + p_{.i})}{n(1 - p_e)^2} \tag{A21}$$

For large n, a simplified version of the Fleiss' formula

$$Var(k) = \frac{1}{n(1 - p_e)^2}\left\{\sum_{i=1}^{m} p_{ii}[1 - (p_{i.} + p_{.i})(1 - \hat{k})]^2 + (1 - \hat{k})^2 \sum_{i \neq j}^{m} p_{ij}(p_{i.} + p_{.j})^2 - [\hat{k} - p_e(1 - \hat{k})]^2\right\} \tag{A22}$$

is given by Altman et al. [62].

$$Var(k) = \frac{\hat{p}_o(1 - \hat{p}_o)}{n(1 - p_e)^2} \tag{A23}$$

This allowed us to build an asymptotic two-side $1 - \alpha$ confidence interval for k:

$$k \pm \sigma_{\hat{k}} * z_{1-\alpha/2} \tag{A24}$$

where $\sigma_{\hat{k}}$ is the standard error of $\hat{k}$ and $z_{1-\alpha/2}$ is the quantile of the standard normal distribution.

The same closed formulas of variance estimates used for k and k_C were also implemented for α index, as these indexes belong to the same approach [62].

Appendix B.3. k_{PABAK} Index

According to Scott [19], we calculated the following formula:

$$Var(k_{PABAK}) = 4\frac{p_o(1-p_o)}{n-1} \tag{A25}$$

Appendix B.4. H Index

Following Scott [19], we obtained the simple formula:

$$Var(H) = \frac{p_o(1-p_o)}{n-1} \tag{A26}$$

Appendix B.5. J and Γ Indexes

Janson and Vegelius [59] proposed a simple formula for the variance estimation of the J index:

$$Var(J) = \frac{2}{n^2} \tag{A27}$$

The same closed formula of variance estimates used for J was also implemented for Γ index, as these indexes belong to the same approach [59].

Appendix B.6. Δ Index

Andrés and Marzo [24] presented these formulas for standard deviation:

$$SE(\Delta_I) = \sqrt{\frac{(1-\Delta)(1+\Delta)}{n}} \tag{A28}$$

$$SE(\Delta_{II}) = \sqrt{\frac{(1-\Delta)}{n}\sum_{i=1}^{2}\frac{x_{ii}}{n_i}} \tag{A29}$$

Appendix B.7. $\gamma(AC_1)$ Index

Gwet [21] proposed a complex formula for the variance of $\gamma(AC_1)$ index:

$$\begin{aligned} Var(\hat{\gamma}) = \frac{1-f}{n(1-p_e)^2}\Bigg\{ & p_o(1-p_o) - 4(1-\hat{\gamma})\left(\frac{1}{q-1}\sum_{k=1}^{q} p_{kk}(1-\hat{\pi}_k) - p_o p_a\right) \\ & +4(1-\hat{\gamma})^2\left(\frac{1}{(q-1)^2}\sum_{k=1}^{q}\sum_{l=1}^{q} p_{kl}[1-(\hat{\pi}_k+\hat{\pi}_l)/2]^2 - p_e^2\right)\Bigg\} \end{aligned} \tag{A30}$$

References

1. Battini, M.; Vieira, A.; Barbieri, S.; Ajuda, I.; Stilwell, G.; Mattiello, S. Invited review: Animal-based indicators for on-farm welfare assessment for dairy goats. *J. Dairy Sci.* **2014**, *97*, 6625–6648. [CrossRef]
2. Meagher, R.K. Observer ratings: Validity and value as a tool for animal welfare research. *Appl. Anim. Behav. Sci.* **2009**, *119*, 1–14. [CrossRef]
3. Kaufman, A.B.; Rosenthal, R. Can you believe my eyes? The importance of interobserver reliability statistics in observations of animal behavior. *Anim. Behav.* **2009**, *78*, 1487–1491. [CrossRef]
4. Krippendorff, K. Reliability in content analysis: Some common misconceptions and recommendations. *Hum. Commun. Res.* **2004**, *30*, 411–433. [CrossRef]
5. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
6. De Rosa, G.; Grasso, F.; Pacelli, C.; Napolitano, F.; Winckler, C. The welfare of dairy buffalo. *Ital. J. Anim. Sci.* **2009**, *8*, 103–116. [CrossRef]
7. Marasini, D.; Quatto, P.; Ripamonti, E. Assessing the inter-rater agreement for ordinal data through weighted indexes. *Stat. Methods Med. Res.* **2016**, *25*, 2611–2633. [CrossRef]
8. Katzenberger, K.; Rauch, E.; Erhard, M.; Reese, S.; Gauly, M. Inter-rater reliability of welfare outcome assessment by an expert and farmers of South Tyrolean dairy farming. *Ital. J. Anim. Sci.* **2020**, *19*, 1079–1090. [CrossRef]

9. Czycholl, I.; Klingbeil, P.; Krieter, J. Interobserver reliability of the animal welfare indicators welfare assessment protocol for horses. *J. Equine Vet. Sci.* **2019**, *75*, 112–121. [CrossRef]
10. Czycholl, I.; Menke, S.; Straßburg, C.; Krieter, J. Reliability of different behavioral tests for growing pigs on-farm. *Appl. Anim. Behav. Sci.* **2019**, *213*, 65–73. [CrossRef]
11. Pfeifer, M.; Eggemann, L.; Kransmann, J.; Schmitt, A.O.; Hessel, E.F. Inter- and intra-observer reliability of animal welfare indicators for the on-farm self-assessment of fattening pigs. *Animal* **2019**, *13*, 1712–1720. [CrossRef] [PubMed]
12. Vieira, A.; Battini, M.; Can, E.; Mattiello, S.; Stilwell, G. Inter-observer reliability of animal-based welfare indicators included in the Animal Welfare Indicators welfare assessment protocol for dairy goats. *Animal* **2018**, *12*, 1942–1949. [CrossRef]
13. De Rosa, G.; Grasso, F.; Winckler, C.; Bilancione, A.; Pacelli, C.; Masucci, F.; Napolitano, F. Application of the Welfare Quality protocol to dairy buffalo farms: Prevalence and reliability of selected measures. *J. Dairy Sci.* **2015**, *98*, 6886–6896. [CrossRef] [PubMed]
14. Mullan, S.; Edwards, S.A.; Butterworth, A.; Whay, H.R.; Main, D.C.J. Inter-observer reliability testing of pig welfare outcome measures proposed for inclusion within farm assurance schemes. *Vet. J.* **2011**, *190*, e100–e109. [CrossRef] [PubMed]
15. Mattiello, S.; Battini, M.; De Rosa, G.; Napolitano, F.; Dwyer, C. How Can We Assess Positive Welfare in Ruminants? *Animals* **2019**, *9*, 758. [CrossRef]
16. Spigarelli, C.; Zuliani, A.; Battini, M.; Mattiello, S.; Bovolenta, S. Welfare Assessment on Pasture: A Review on Animal-Based Measures for Ruminants. *Animals* **2020**, *10*, 609. [CrossRef]
17. Walsh, P.; Thornton, J.; Asato, J.; Walker, N.; McCoy, G.; Baal, J.; Baal, J.; Mendoza, N.; Banimahd, F. Approaches to describing inter-rater reliability of the overall clinical appearance of febrile infants and toddlers in the emergency department. *PeerJ* **2014**, *2*, e651. [CrossRef]
18. Ato, M.; Lopez, J.J.; Benavente, A. A simulation study of rater agreement measures with 2x2 contingency tables. *Psicológica* **2011**, *32*, 385–402.
19. Scott, W.A. Reliability of content analysis: The case of nominal scale coding. *Public Opin. Q.* **1955**, *19*, 321–325. [CrossRef]
20. Bennett, E.M.; Alpert, R.; Goldstein, A.C. Communications through limited response questioning. *Public Opin. Q.* **1954**, *18*, 303–308. [CrossRef]
21. Gwet, K. Computing inter-rater reliability and its variance in presence of high agreement. *Br. J. Math. Stat. Psychol.* **2008**, *61*, 29–48. [CrossRef]
22. Tanner, M.A.; Young, M.A. Modeling agreement among raters. *J. Am. Stat. Assoc.* **1985**, *80*, 175–180. [CrossRef]
23. Aickin, M. Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's kappa. *Biometrics* **1990**, *46*, 293–302. [CrossRef]
24. Andrés, A.M.; Marzo, P.F. Delta: A new measure of agreement between two raters. *Br. J. Math. Stat. Psychol.* **2004**, *57*, 1–19. [CrossRef]
25. AWIN (Animal Welfare Indicators). AWIN Welfare Assessment Protocol for Goats. 2015. Available online: https://air.unimi.it/retrieve/handle/2434/269102/384790/AWINProtocolGoats.pdf (accessed on 3 May 2021).
26. Battini, M.; Stilwell, G.; Vieira, A.; Barbieri, S.; Canali, E.; Mattiello, S. On-farm welfare assessment protocol for adult dairy goats in intensive production systems. *Animals* **2015**, *5*, 934–950. [CrossRef] [PubMed]
27. Holley, J.W.; Guilford, J.P. A note on the G index of agreement. *Educ. Psychol. Meas.* **1964**, *34*, 749–753. [CrossRef]
28. Quatto, P. Un test di concordanza tra più esaminatori. *Statistica* **2004**, *64*, 145–151.
29. Holsti, O.R. *Content Analysis for the Social Sciences and Humanities*; Addison-Wesley: Reading, MA, USA, 1969; pp. 1–235.
30. Krippendorff, K. Estimating the reliability, systematic error and random error of interval data. *Educ. Psychol. Meas.* **1970**, *30*, 61–70. [CrossRef]
31. Hubert, L. Nominal scale response agreement as a generalized correlation. *Br. J. Math. Stat. Psychol.* **1977**, *30*, 98–103. [CrossRef]
32. Janson, S.; Vegelius, J. On the applicability of truncated component analysis based on correlation coefficients for nominal scales. *Appl. Psychol. Meas.* **1978**, *2*, 135–145. [CrossRef]
33. Bangdiwala, S.I. A graphical test for observer agreement. In Proceedings of the 45th International Statistical Institute Meeting, Amsterdam, The Netherlands, 12–22 August 1985; Bishop, Y.M.M., Fienberg, S.E., Holland, P.W., Eds.; SpringerLink: Berlin, Germany, 1985; pp. 307–308.
34. Efron, B. Bootstrap methods: Another look at the jackknife. *Ann. Stat.* **1979**, *7*, 1–26. [CrossRef]
35. Klar, N.; Lipsitz, S.R.; Parzen, M.; Leong, T. An exact bootstrap confidence interval for k in small samples. *J. R. Stat. Soc. Ser. D-Stat.* **2002**, *51*, 467–478. [CrossRef]
36. Kinsella, A. The 'exact' bootstrap approach to confidence intervals for the relative difference statistic. *J. R. Stat. Soc. Ser. D-Stat.* **1987**, *36*, 345–347, correction, **1988**, *37*, 97. [CrossRef]
37. Quatto, P.; Ripamonti, E. Raters: A Modification of Fleiss' Kappa in Case of Nominal and Ordinal Variables. R Package Version 2.0.1. 2014. Available online: https://CRAN.R-project.org/package=raters (accessed on 5 May 2021).
38. Meyer, D.; Zeileis, A.; Hornik, K. The Strucplot Framework: Visualizing Multi-Way contingency Table with vcd. *J. Stat. Softw.* **2006**, *17*, 1–48. [CrossRef]
39. S Original, from StatLib and by Tibshirani, R. R Port by Friedrich Leisch. Bootstrap: Functions for the Book "An Introduction to the Bootstrap". R Package Version 2019.6. 2019. Available online: https://CRAN.R-project.org/packages=bootstrap (accessed on 5 May 2021).

40. Banerjee, M.; Capozzoli, M.; Mc Sweeney, L.; Sinha, D. Beyond kappa: A review of interrater agreement measures. *Can. J. Stat.-Rev. Can. Stat.* **1999**, *27*, 3–23. [CrossRef]
41. Wang, W. A Content Analysis of Reliability in Advertising Content Analysis Studies. Paper 1375. Master's Thesis, Department of Communication, East Tennessee State Univ., Johnson City, TN, USA, 2011. Available online: https://dc.etsu.edu/etd/1375 (accessed on 22 March 2021).
42. Lombard, M.; Snyder-Duch, J.; Bracken, C.C. Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Hum. Commun. Res.* **2002**, *28*, 587–604. [CrossRef]
43. Kuppens, S.; Holden, G.; Barker, K.; Rosenberg, G. A Kappa-related decision: K, Y, G, or AC1. *Soc. Work Res.* **2011**, *35*, 185–189. [CrossRef]
44. Feinstein, A.R.; Cicchetti, D.V. High agreement but low kappa: I. The problem of two paradoxes. *J. Clin. Epidemiol.* **1990**, *43*, 543–549. [CrossRef]
45. Lantz, C.A.; Nebenzahl, E. Behavior and interpretation of the κ statistics: Resolution of the two paradoxes. *J. Clin. Epidemiol.* **1996**, *49*, 431–434. [CrossRef]
46. Byrt, T.; Bishop, J.; Carli, J.B. Bias, prevalence and kappa. *J. Clin. Epidemiol.* **1993**, *46*, 423–429. [CrossRef]
47. Shankar, V.; Bangdiwala, S.I. Observer agreement paradoxes in 2 × 2 tables: Comparison of agreement measures. *BMC Med. Res. Methodol.* **2014**, *14*, 100. [CrossRef]
48. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. [CrossRef]
49. Fleiss, J.L. Measuring nominal scale agreement among many raters. *Psychol. Bull.* **1981**, *76*, 378–382. [CrossRef]
50. Cicchetti, D.V.; Feinstein, A.R. High agreement but low kappa: II. Resolving the paradoxes. *J. Clin. Epidemiol.* **1990**, *43*, 551–558. [CrossRef]
51. Brennan, R.L.; Prediger, D.J. Coefficient kappa: Some uses, misuses, and alternatives. *Educ. Psychol. Meas.* **1981**, *41*, 687–699. [CrossRef]
52. Zhao, X. When to Use Scott's π or Krippendorff's α, If Ever? Presented at the Annual Conference of Association for Education in Journalism and Mass Communication, St. Louis, MO, USA, 10–13 August 2011; Available online: https://repository.hkbu.edu.hk/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1002&context=coms_conf (accessed on 22 March 2021).
53. Gwet, K.L. On Krippendorff's Alpha Coefficient. Available online: http://www.bwgriffin.com/gsu/courses/edur9131/content/onkrippendorffalpha.pdf (accessed on 22 March 2021).
54. Falotico, R.; Quatto, P. On avoiding paradoxes in assessing inter-rater agreement. *Ital. J. Appl. Stat.* **2010**, *22*, 151–160.
55. Friendly, M. *Visualizing Categorical Data*; SAS Institute: Cary, NC, USA, 2000.
56. McCray, G. Assessing Inter-Rater Agreement for Nominal Judgement Variables. Presented at the Language Testing Forum, University of Lancaster, Nottingham, UK, 15–17 November 2013; Available online: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.725.8104&rep=rep1&type=pdf (accessed on 22 March 2021).
57. Wongpakaran, N.; Wongpakaran, T.; Wedding, D.; Gwet, K.L. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples. *BMC Med. Res. Methodol.* **2013**, *13*, 61. [CrossRef]
58. Kendall, M.G. *Rank Correlation Methods*; Hafner Publishing Co.: New York, NY, USA, 1955; pp. 1–196.
59. Janson, S.; Vegelius, J. The J-index as a measure of nominal scale response agreement. *Appl. Psychol. Meas.* **1982**, *6*, 111–121. [CrossRef]
60. Fleiss, J.L.; Cohen, J.; Everitt, B. Large-sample standard errors of kappa and weighted kappa. *Psychol. Bull.* **1969**, *72*, 323–327. [CrossRef]
61. Everitt, B.S. Moments of the statistics kappa and weighted kappa. *Br. J. Math. Stat. Psychol.* **1968**, *21*, 97–103. [CrossRef]
62. Altman, D.G. Statistics in medical journals: Some recent trends. *Stat. Med.* **2000**, *19*, 3275–3289. [CrossRef]