

Concordance Analysis for Scoring Data

Annotated Report with Decision-Making Guidance based on Standard Methods

Kulig, Schäfer, Addo, Lange, Wilczek, Hensel, Jung

2026-07-03

Table of contents

1 Methodological Setup of the Current Analysis	3
1.1 Project Description: Give the Titel or your Project	3
1.2 Answers to the Questions Leading to the Setup	3
1.3 Applied Normalized Weighting Matrix	4
1.4 Applied Interpretation Benchmark	4
2 Descriptive Statistics (Groupwise/Pairwise) According to the Selected Methodological Setup	5
2.1 Concordance Tables as a Cross-Matrix & Grid with Percentage Agreement	5
2.2 Graphical Exploration: Bangdiwala's <i>B</i> Agreement Plot	7
2.3 Classwise Agreement	10
3 Results of the Statistical Analysis According to Chosen Methodological Setup	11
3.1 Concordance Analysis	11
3.2 Metric Equivalence-Tests	11
3.3 Bias-Analysis	12
4 Methodological Classification & Decision-Making Guide	13
4.1 Profile of the statistical methods used in concordance and agreement analysis	13
4.1.1 Percentage of agreement	13
4.1.2 Cohen's Kappa	13
4.1.3 Weighted Kappa	13
4.1.4 PABAK	13
4.1.5 Fleiss' Kappa	14
4.1.6 Krippendorff's Alpha	14
4.1.7 Gwet's <i>AC1</i> and <i>AC2</i>	14
4.1.8 Kendall's <i>W</i>	15
4.1.9 Intraclass Correlation Coefficient: $ICC_{(A,1)}$ and $ICC_{(C,1)}$	15
4.1.10 Bangdiwala's <i>B</i> Agreement Plot	15

4.1.11 Deming Regression	16
4.1.12 Bland-Altman Plot and Equivalence Analysis	16
4.2 Key points of the statistical analyses	16
4.2.1 The Prevalence/Bias Paradox	16
4.2.2 Relevance of Weighting (Penalty)	17
4.2.3 Experimental Feature: Asymmetric Confidence Intervals via Bootstrapping	17
4.2.4 Threshold Models: Transition from Dichotomous to Ordinal to Metric	17
4.2.5 Consistency vs. Agreement? What Does It Do?	18
4.2.6 Repeatability vs. Reproducibility	18
4.2.7 What to Do When There Are More Than 2 Raters or Time Points?	18

References

19

1 Methodological Setup of the Current Analysis

1.1 Project Description: Give the Titel or your Project

Give a Brief Description of your Project e.g.: This concordance analysis investigates the reliability of visual scoring systems within the context of agricultural engineering and precision livestock management. Human observation in ethology inevitably contains inherent noise and must be treated mathematically as providing ‘noisy labels’ rather than representing an absolute, perfect truth. This framework evaluates the inter-subjective agreement to validate these imperfect gold standards.

1.2 Answers to the Questions Leading to the Setup

The current analysis was automatically generated based on the following configurations:

- **Input Data File:** `Concordance_Data_Fixed_Standard.xlsx`
- **Evaluated Raters/Timepoints (3):** `Rater_A_Nom_4`, `Rater_B_Nom_4`, `Rater_C_Nom_4`
- **Scale Level:** Nominal (Score Levels: 4 | Healthy/Baseline = 0)
- **Applied Weighting Penalty:** Unweighted
- **Equivalence Margin (TOST):** ± 0.5 score units
- **ICC Calculation & Bias Recognition:** Active
- **Focus:** Agreement
- **Type:** Repeatability (“intra-rater”)

Suggested Draft for the Materials and Methods Section:

To evaluate the reliability of the scoring system, the assessments of 3 independent raters/time points were compared. The underlying variable was treated as an unranked nominal trait with 4 categories. The analysis was designed as a repeatability check (intra-rater reliability) to assess the stability of judgments across multiple time points. By focusing on strict absolute agreement, the framework evaluates whether assigned scores match exactly, penalizing any systematic severity shifts. To ensure robust statistical inference, asymmetric 95% confidence intervals for the concordance coefficients were calculated using percentile bootstrapping with 100 iterations. The magnitude of the resulting concordance coefficients was evaluated according to the benchmark proposed by Landis_koch.

The analyses in this framework are based on the statistical programming language R (R Core Team, 2026). To ensure comprehensive methodological coverage, the following libraries are integrated: `knitr` (Xie, 2025), `dplyr` (Wickham, François, et al., 2026), `tidyr` (Wickham et al., 2025), `ggplot2` (Wickham, Chang, et al., 2026): For data wrangling, advanced visualization (Variability Charts), and the dynamic generation of the results report via LaTeX/PDF; `readxl` (Wickham & Bryan, 2025) and `writexl` (Ooms & McNamara, 2024): For the reliable import and export of data from Microsoft Excel; `irr` (Gamer et al., 2019), `irrCAC` (Gwet, 2019): Core packages for calculating Kappa measures (Cohen, Fleiss) and Krippendorff’s Alpha, Kendall’s W, and ICC; `psych` (Revelle, 2026): Enables the use of complex, user-defined weighting matrices for weighted Kappa as well as certain aspects of the ICC. The `vegan` library (Oksanen et al., 2026) is used to implement Legendre’s approach (Legendre, 2005) as an extension of Kendall’s W. Bangdiwala’s B Agreement plot is generated using the `vcd` library (Meyer et al., 2002) and (Meyer et al., 2006). To generate the framework’s PDF output, RStudio (Posit Team, 2026) is required, along with the additional packages `yaml` (Stephens & Simonov, 2025), `rmarkdown` (Allaire, Xie, et al., 2026), `Quarto` (Allaire, Teague, et al., 2026) and `TinyTeX` (Xie, 2026).

1.3 Applied Normalized Weighting Matrix

No weighting matrix applied (categorical data or unweighted setup).

1.4 Applied Interpretation Benchmark

Based on the configuration, the **Landis_koch** benchmark is used to interpret the magnitude of the concordance coefficients in the following results:

Coefficient	Interpretation
< 0	Poor
0.01 - 0.20	Slight
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Substantial
0.81 - 1.00	Almost Perfect

The limits and their class assignments are derived from the following sources, depending on the setup: Landis & Koch (1977), Fleiss et al. (2003) and McHugh (2012). If you select the “custom” setting, you must determine the appropriateness of the limits yourself.

2 Descriptive Statistics (Groupwise/Pairwise) According to the Selected Methodological Setup

2.1 Concordance Tables as a Cross-Matrix & Grid with Percentage Agreement

The following cross-tabulations display the absolute number of assignments for all rater combinations based on the theoretical scale. Unoccupied categories are filled with “0” to map the complete scale matrix. The measure calculation is based on the actual scale.

Comparison: Rater_A_Nom_4 vs. Rater_B_Nom_4

Raw Percentage Agreement (without chance correction): 62 %

Pearson's Chi-squared test: $\chi^2 = 107.18$, p-value = <0.001

Cohen's Kappa (unweighted): 0.4984

Table Axis: Rows shows Score of = Rater_A_Nom_4 | Columns shows Score of = Rater_B_Nom_4

Score	healthy	preclinical	sick	serious
healthy	17	16	0	0
preclinical	3	9	2	0
sick	0	3	17	5
serious	0	0	9	19

Comparison: Rater_A_Nom_4 vs. Rater_C_Nom_4

Raw Percentage Agreement (without chance correction): 44 %

Pearson's Chi-squared test: $\chi^2 = 48.81$, p-value = <0.001

Cohen's Kappa (unweighted): 0.2670

Table Axis: Rows shows Score of = Rater_A_Nom_4 | Columns shows Score of = Rater_C_Nom_4

Score	healthy	preclinical	sick	serious
healthy	15	13	5	0
preclinical	3	7	4	0
sick	1	8	9	7
serious	0	4	11	13

Comparison: Rater_B_Nom_4 vs. Rater_C_Nom_4

Raw Percentage Agreement (without chance correction): 56 %

Pearson's Chi-squared test: $\chi^2 = 88.36$, p-value = <0.001

Cohen's Kappa (unweighted): 0.4080

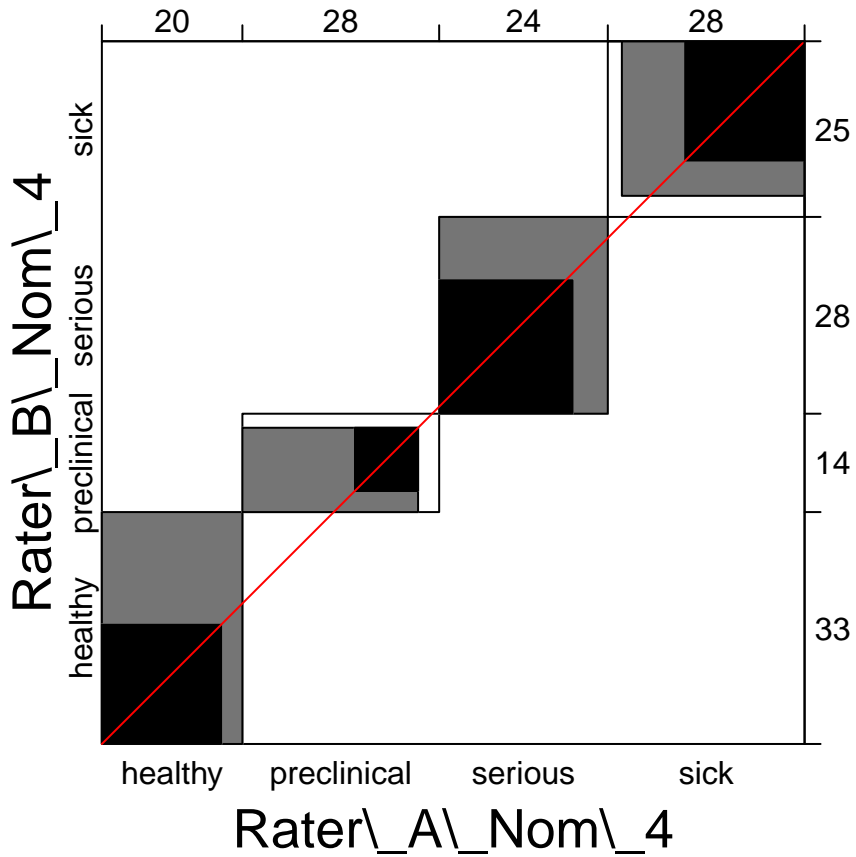
Table Axis: Rows shows Score of = Rater_B_Nom_4 | Columns shows Score of = Rater_C_Nom_4

Score	healthy	preclinical	sick	serious
healthy	13	7	0	0
preclinical	6	14	8	0
sick	0	11	13	4
serious	0	0	8	16

2.2 Graphical Exploration: Bangdiwala's B Agreement Plot

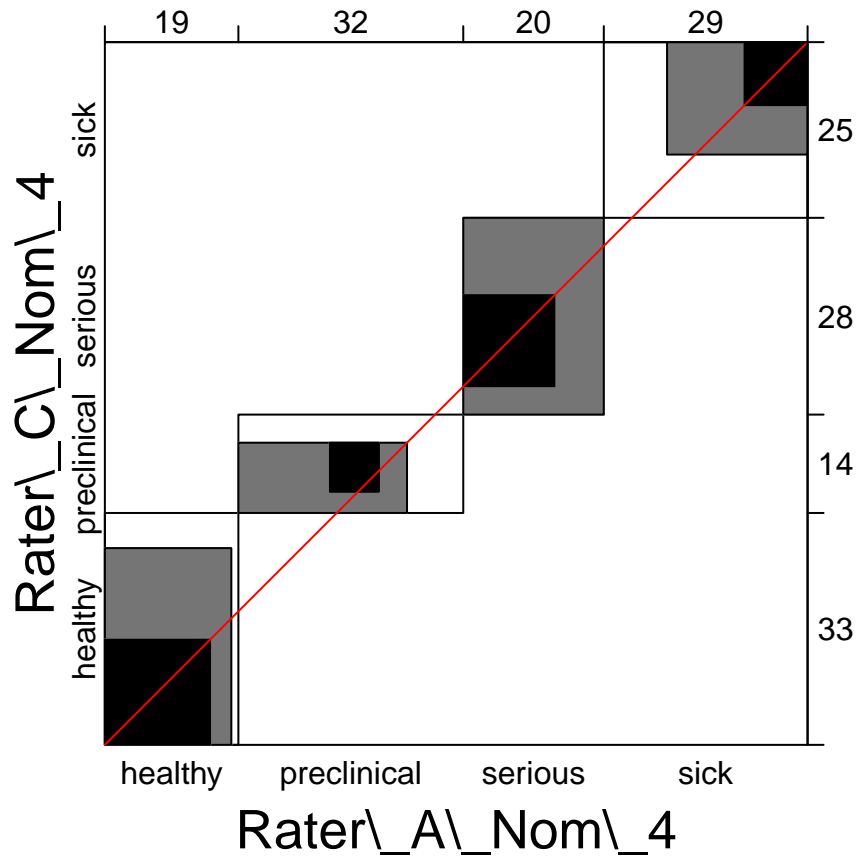
Agreement Plot: Rater_A_Nom_4 vs. Rater_B_Nom_4

Calculated Bangdiwala's B-Value: 0.4208



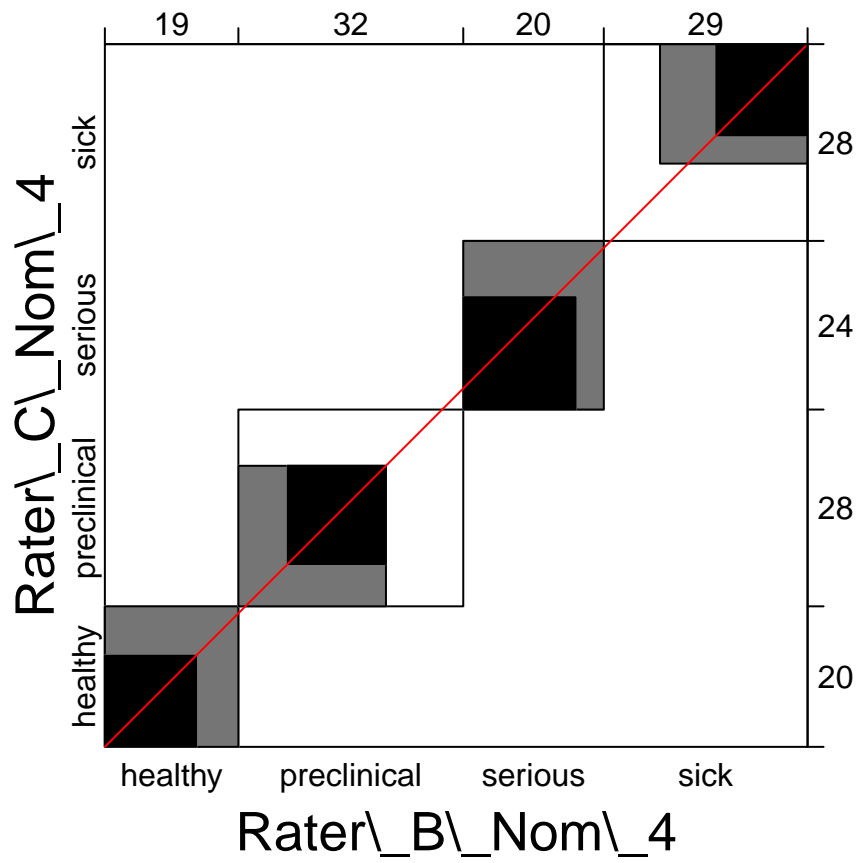
Agreement Plot: Rater_A_Nom_4 vs. Rater_C_Nom_4

Calculated Bangdiwala's B-Value: 0.222



Agreement Plot: Rater_B_Nom_4 vs. Rater_C_Nom_4

Calculated Bangdiwala's B-Value: 0.3076



2.3 Classwise Agreement

While global concordance measures describe the overall quality of the scoring, this evaluation highlights which specific scores carry the highest uncertainty.

Overall Classwise Agreement (All Raters)

Score Class	Kappa	p-Value	Evaluation
healthy	0.507	<0.001	Moderate
preclinical	0.211	<0.001	Fair
serious	0.561	<0.001	Moderate
sick	0.278	<0.001	Fair

Pairwise Classwise Agreement

Comparison: Rater_A_Nom_4 vs. Rater_B_Nom_4

Score Class	Kappa	p-Value	Evaluation
healthy	0.512	<0.001	Moderate
preclinical	0.277	0.006	Fair
serious	0.636	<0.001	Substantial
sick	0.512	<0.001	Moderate

Comparison: Rater_A_Nom_4 vs. Rater_C_Nom_4

Score Class	Kappa	p-Value	Evaluation
healthy	0.428	<0.001	Moderate
preclinical	0.097	0.334	Slight
serious	0.397	<0.001	Fair
sick	0.087	0.386	Slight

Comparison: Rater_B_Nom_4 vs. Rater_C_Nom_4

Score Class	Kappa	p-Value	Evaluation
healthy	0.586	<0.001	Moderate
preclinical	0.238	0.017	Fair
serious	0.650	<0.001	Substantial
sick	0.239	0.017	Fair

3 Results of the Statistical Analysis According to Chosen Methodological Setup

3.1 Concordance Analysis

The following table dynamically lists all theoretical variants evaluated by the framework logic. Methods that are mathematically unsupported or strongly discouraged given the current configuration are marked with **n/a**.

No.	Method	Value	95 % CI	Interpretation	if $m > 2$	Weighting?
1	Percentage Agreement $p_o\%$	54.000 %	[0.4682; 0.6135]		pairwise avg.	not applicable
2	Cohen's Kappa	0.3911	[0.3035; 0.4753]	Fair	pairwise avg.	no
3	Weighted Kappa	n/a	n/a	Deactivated by Rule Engine (Variant 11 (Nominal, $m > 2$))	pairwise avg.	yes if $k > 2$
4a	PABAK 2×2	1.0000	[1.0000; 1.0000]	Almost Perfect	pairwise avg.	no
4b	PABAK $R \times C$	0.3867	[0.3177; 0.4800]	Fair	pairwise avg.	no
5	Fleiss' Kappa	0.3860	[0.2825; 0.4796]	Fair	native	no
6	Krippendorff's Alpha	0.3871	[0.2707; 0.4912]	Fair	native	yes if $k > 2$
7a	Gwet's $AC1$	0.3869	[0.2836; 0.4703]	Fair	native	no
7b	Gwet's $AC2$	n/a	n/a	Deactivated by Rule Engine (Variant 11 (Nominal, $m > 2$))	native	yes if $k > 2$
8	Kendall's W	n/a	n/a	Deactivated by Rule Engine (Variant 11 (Nominal, $m > 2$))	native	based on ordinal
9a	$ICC_{(C,1)}$	n/a	n/a	Deactivated by Rule Engine (Variant 11 (Nominal, $m > 2$))	native	based on ordinal
9b	$ICC_{(A,1)}$	n/a	n/a	Deactivated by Rule Engine (Variant 11 (Nominal, $m > 2$))	native	based on ordinal
10	Bangdiwala's B Agreement	0.3168	[0.2486; 0.4017]	Fair	native	only visual
11	Deming Regression	n/a	n/a	Deactivated by Rule Engine (Variant 11 (Nominal, $m > 2$))	pairwise avg.	not applicable
12	Bland-Altman, TOST	n/a	n/a	Deactivated by Rule Engine (Variant 11 (Nominal, $m > 2$))	pairwise avg.	not applicable

3.2 Metric Equivalence-Tests

Note: Metric equivalence tests (Deming Regression, Bland-Altman, TOST) are mathematically disabled for low-level categorical or non-numeric data.

3.3 Bias-Analysis

Overall System Bias (All Raters)

Pairwise Bias Analysis

Comparison: Rater_A_Nom_4 vs. Rater_B_Nom_4

Comparison: Rater_A_Nom_4 vs. Rater_C_Nom_4

Comparison: Rater_B_Nom_4 vs. Rater_C_Nom_4

4 Methodological Classification & Decision-Making Guide

4.1 Profile of the statistical methods used in concordance and agreement analysis

4.1.1 Percentage of agreement

1. Scale level: nominal / ordinal;
2. Number of raters: in its basic form $m = 2$ (for $m > 2$ usually as the mean of all pairwise comparisons);
3. Weighting: none (only exact matches count, main diagonal);
4. Classification: purely descriptive measure to display the absolute agreement rate;
5. Prerequisites & Limitations: major weakness is that chance is completely ignored. With only two categories (Yes/No), agreement by pure guessing would already be 50 %, which often makes $p_o\%$ alone seem too optimistic;
6. Critique & No-Go: must never be used in scientific papers as the sole measure of reliability. It serves as a necessary supplement to chance-corrected measures to make their results interpretable (especially in cases of paradoxes).

4.1.2 Cohen's Kappa

1. Scale level: nominal (also usable for ordinal data, but then ignores the ranking);
2. Number of raters: in its basic form $m = 2$ (for $m > 2$ usually as the mean of all pairwise comparisons);
3. Weighting: none; every error is considered equal;
4. Classification: primarily designed for absolute agreement in repeatability and reproducibility;
5. Prerequisites & Limitations: the observers must judge independently of each other. The biggest weakness is the extreme susceptibility to the prevalence paradox;
6. Critique & No-Go: for very rare or very common traits, the value drops massively, even if the raters show high agreement (p_o). In such a case, p_e mathematically approaches p_o , leading to a disproportionately small Kappa value. Main Source to Cite: Cohen (1960)

4.1.3 Weighted Kappa

1. Scale level: ordinal;
2. Number of raters: in its basic form $m = 2$ (for $m > 2$ usually as the mean of all pairwise comparisons);
3. Weighting: yes (linear, quadratic, or defined via individual matrices);
4. Classification: measures a weighted agreement. Used to determine reproducibility and repeatability for graded scores;
5. Prerequisites & Limitations: requires that distances between levels can be meaningfully interpreted. A quadratic weighting is mathematically similar to the *ICC*. The prevalence paradox described above also applies to weighted Kappa;
6. Critique & No-Go: application to purely nominal data without a ranking (e.g., different breeds), since a “distance” between categories is biologically meaningless here. Main Source to Cite: Cohen (1968), Cohen (1972)

4.1.4 PABAK

1. Scale level: nominal (mostly dichotomous, but expandable to k categories);
2. Number of raters: in its basic form $m = 2$ (for $m > 2$ usually as the mean of all pairwise comparisons);
3. Weighting: none;

4. Classification: agreement measure for repeatability and reproducibility, specialized for datasets with a skewed distribution (e.g., almost exclusively healthy animals);
5. Prerequisites & Limitations: calculates a theoretical Kappa assuming perfect uniform distribution (no bias, no prevalence distortion). It shows the potential agreement given an optimal trait distribution. **For multi-level nominal data ($R \times C$), it effectively evaluates binary exact matches against all mismatches while adjusting for the correct $1/k$ chance probability.**
6. Critique & No-Go: critics accuse PABAK of artificially “beautifying” the agreement, as real difficulties in detecting rare cases are ignored. It should therefore always be reported alongside the classic Kappa. Main Source to Cite: Byrt et al. (1993)

4.1.5 Fleiss’ Kappa

1. Scale level: nominal (also usable for ordinal data, but then ignores the ranking);
2. Number of raters: $m \geq 2$;
3. Weighting: none;
4. Classification: measure of agreement in reproducibility within larger teams;
5. Prerequisites & Limitations: requires a constant number of judgments per object. Like the classic Kappa, it is extremely prevalence-dependent (prevalence paradox) and ignores the distance of incorrect decisions in ordinal data;
6. Critique & No-Go: Fleiss’ Kappa should not be used in isolation if specific rater differences need to be uncovered. Here, the analysis should be supplemented by pairwise Kappas to avoid “blind spots” in averaging. It is primarily chosen to depict the overall agreement of a team; an exploratory review of pairwise agreements is (if feasible) still recommended. Main Source to Cite: Fleiss (1971), Fleiss et al. (2003)

4.1.6 Krippendorff’s Alpha

1. Scale level: any (nominal, ordinal, interval, ratio) due to the universal use of the distance function δ^2 ;
2. Number of raters: any ($m \geq 2$), the number of judgments per object does not have to be constant (excellent handling of missing values);
3. Weighting: yes, is a core component of the Krippendorff algorithm and is regulated via the distance function;
4. Classification: a comprehensive agreement measure even for incomplete and complex experimental designs. It can be used both for reproducibility (different raters) and for repeatability (one rater at ≥ 2 time points). It contains a correction factor for small samples (n), making it more precise than Fleiss’ Kappa;
5. Prerequisites & Limitations: a minimum variance of the sample must be given so that Alpha can measure the consistency of distinction. Alpha thus suffers from the same paradox as Kappa. In homogeneous groups (e.g., almost all cows are healthy), Alpha drops towards zero since there is hardly any “expected disagreement”. Any minimal real deviation is disproportionately penalized here. Like Fleiss, Alpha aggregates all judgments. A single poor rater is “smoothed out” in the total value but lowers it non-specifically;
6. Critique & No-Go: Krippendorff is often more conservative than Kappa. Using it as the “only truth” with homogeneous data is a no-go. A low Alpha with low observed disagreement (D_o) or high percentage agreement (p_o) is often an artifact of the sample, not the rater quality. Forgoing pairwise analyses when trying to identify “problem raters” in the team is also strongly discouraged. Main Source to Cite: Krippendorff (2018), Krippendorff (2004)

4.1.7 Gwet’s $AC1$ and $AC2$

1. Scale level: $AC1$ for nominal data, $AC2$ for ordinal or interval data;
2. Number of raters: any ($m \geq 2$);
3. Weighting: yes, for $AC2$ (analogous to weighted Kappa or Krippendorff);

4. Classification: agreement measure that protects the stability of the judgment against coincidental agreements in extreme distributions;
5. Prerequisites & Limitations: mathematically more robust than Kappa, Fleiss, and Krippendorff with skewed distributions. The main hurdle is its lower familiarity in classical literature, which is why it is often reported in addition to Kappa. Specialty: animal health monitoring and diagnosis of rare diseases, where the prevalence paradox would distort the results;
6. Critique & No-Go: Gwet's AC is considered a more "liberal" measure. Sole reporting without justification regarding the prevalence situation is a no-go. Criticism is often leveled at the model concept of "random guessing", which underlies the calculation of $e(\gamma)$. In reality, there will likely always be a gray area between total knowledge and total guessing. For ordinal data, $AC2$ should always be used with explicit mention of the weights (linear/quadratic). Main Source to Cite: Gwet (2008), Gwet (2014)

4.1.8 Kendall's W

1. Scale level: ordinal (rank data);
2. Number of raters: any ($m \geq 2$);
3. Weighting: not applicable (the measure is based on ranks);
4. Classification: measure of concordance (rank agreement) within groups or teams;
5. Prerequisites & Limitations: does not react to absolute level differences (bias). Requires a mathematical correction in the denominator for ties (identical scores), which is part of the standard implementation in software;
6. Critique & No-Go: a high W does not prove exact agreement, but only a synchronous ranking. W must not be used as the sole proof of diagnostic accuracy. In diagnostics, absolute agreement (Kappa/Gwet) is crucial. It is the ideal tool for root cause analysis (bias check) in a team. Main Source to Cite: Kendall & Smith (1939)

4.1.9 Intraclass Correlation Coefficient: $ICC_{(A,1)}$ and $ICC_{(C,1)}$

1. Scale level: metric or pseudo-metric (recommended for $k \geq 7$ levels; for $k = 5$ only with reservations);
2. Number of raters: any ($m \geq 2$);
3. Weighting: implicit (quadratic via analysis of variance);
4. Classification:
 - 4.1 Model choice: Two-Way Random for reproducibility (generalization to the population) and Two-Way Mixed for internal repeatability;
 - 4.2 Application: both consistency and absolute agreement; resulting in $ICC_{(C,1)}$ and $ICC_{(A,1)}$;
 - 4.3 Error analysis: comparing both values (ICC_C vs. ICC_A) serves as a systematic bias detector;
5. Prerequisites & Limitations: since the ICC is a ratio of variances, it strictly requires differences between objects. In homogeneous groups (e.g., 50 healthy cows with score 0), the ICC mathematically approaches zero or becomes undefined ($MS_{Items} = 0$), even at 100% agreement;
6. Critique & No-Go: unsuitable for narrow ordinal scales ($k < 5$) and homogeneous samples without variance; here, percentage agreement or Gwet's AC is essential as a corrective measure. Main Source to Cite: Fisher (1967), Fleiss & Cohen (1973)

4.1.10 Bangdiwala's B Agreement Plot

1. Scale level: nominal / ordinal (weighting possible);
2. Number of raters: $m = 2$ (extensions for $m > 2$ exist but are graphically complex);
3. Weighting: possible (geometric representation of neighborhood agreements);

4. Classification: descriptive and visually exploratory measure, ideal for identifying “problem categories” and systematic deviations;
5. Prerequisites & Limitations: does not provide a p-value or a direct chance correction in the classic sense; the focus is on visual evidence;
6. Critique & No-Go: should not be used as a replacement, but as a qualitative supplement to chance-corrected measures (Kappa, Gwet, etc.) in scientific applications. Main Source to Cite: Bangdiwala & Shankar (2013), Munoz & Bangdiwala (1997), Stokes & Koch (2011)

4.1.11 Deming Regression

1. Scale level: metric or pseudo-metric (ordinal with $k \geq 8$);
2. Number of raters: exactly two ($m = 2$);
3. Weighting: not applicable;
4. Classification: a regression-based approach to evaluate both absolute agreement and systemic bias. Unlike ordinary least squares (OLS) regression, it accounts for measurement errors in both the x and y variables simultaneously;
5. Prerequisites & Limitations: assumes that both raters or methods are subject to measurement error, which is typical in ethological scoring. The interpretation of the intercept indicates a fixed/systematic bias, while the slope indicates proportional bias;
6. Critique & No-Go: completely unsuitable for low-level categorical or narrow ordinal data. A correlation coefficient derived from Deming regression is a measure of linear relationship, not strictly an agreement measure on its own, and must always be contextualized. Main Source to Cite: Deming (1943)

4.1.12 Bland-Altman Plot and Equivalence Analysis

1. Scale level: metric or pseudo-metric (ordinal with $k \geq 8$);
2. Number of raters: exactly two ($m = 2$);
3. Weighting: not applicable;
4. Classification: visual and statistical method to evaluate agreement by plotting differences between measurements against their mean. Often supplemented by Two One-Sided Tests (TOST) to prove practical equivalence within a predefined margin (ϵ);
5. Prerequisites & Limitations: the differences between measurements should be approximately normally distributed. The equivalence margin ϵ must be clinically or ethologically justified prior to the analysis;
6. Critique & No-Go: a high correlation does not imply good agreement; Bland-Altman is required to assess the actual magnitude of discrepancies. It fails to capture non-linear biases easily if not plotted, and should never be used for nominal or short ordinal scales. Main Source to Cite: Bland & Altman (1986), Schuirmann (1987), Walker & Nowacki (2011)

4.2 Key points of the statistical analyses

4.2.1 The Prevalence/Bias Paradox

In practice, the problem frequently arises that the herd prevalence of a trait is extremely asymmetric (e.g., 90% of the animals are healthy, 10% are lame). Classic measures like Cohen’s Kappa punish any minor deviation by the raters extremely harshly in such homogeneous datasets. Even with a percentage agreement of 95%, Kappa can drop close to 0. Therefore, this framework calculates more robust alternatives like the PABAK (Prevalence and Bias Adjusted Kappa) and Gwet’s *AC* in parallel to prevent severe misinterpretations of the measurement system’s reliability. Main Source to Cite: Feinstein & Cicchetti (1990), Cicchetti & Feinstein (1990).

4.2.2 Relevance of Weighting (Penalty)

Cohen's Kappa or Gwet's *AC1* are unweighted in their basic form. This means every deviation is considered equally wrong. For ordinal scales, unweighted calculations are legitimate if exact agreement is required. In practice, however, an error of Score 0 vs. a strongly deviating score is often more fatal than a deviation between two adjacent disease scores. Therefore, weightings should be applied: * **Linear**: Penalizes errors proportionally to the distance (Standard for ordinal scores). * **Quadratic**: Penalizes outliers disproportionately strong (similar to the *ICC*). * **Individual (Irregular)**: Allows specific misclassifications to be weighted clinically. Main Source to Cite: Cicchetti & Allison (1971), Fleiss & Cohen (1973)

4.2.3 Experimental Feature: Asymmetric Confidence Intervals via Bootstrapping

Neither PABAK nor Krippendorff's Alpha possess closed analytical formulas for the exact calculation of confidence intervals in their basic mathematical form. To close this scientific gap, this framework utilizes *percentile bootstrapping* (Standard: 100 iterations). Thousands of random new samples (with replacement) are drawn from the existing raw data. Bootstrapping is consistently applied as the primary or secondary CI-source for the following methods in this report:

- Methode 1: Percentage Agreement
- Methode 2: Cohen's Kappa
- Methode 3: Weighted Kappa
- Methode 4: PABAK 2x2 (4a) and PABAK RxC (4b)
- Methode 5: Fleiss' Kappa
- Methode 6: Krippendorff's Alpha
- Methode 7: Gwet's *AC1* (7a) and Gwet's *AC2* (7b)
- Methode 8: Kendall's *W*
- Methode 10: Bangdiwala's *B*
- Methode 11: Deming Orthogonal Regression
- Methode 12: Bland-Altman Mean Difference and TOST

The immense advantage of this iterative procedure is that the confidence interval turns out **naturally asymmetric**. It adapts exactly to the skewness of the data and prevents mathematically impossible boundaries (like values > 1.0).

4.2.4 Threshold Models: Transition from Dichotomous to Ordinal to Metric

When moving from dichotomous to ordinal scales (e.g., 5-point scores), human observation acts as an imperfect gold standard. It contains inherent noise and must be treated mathematically as 'Noisy Labels' rather than an absolute truth. As the number of categories increases (e.g., $k > 7$), the scale approaches a metric continuum, requiring different statistical approaches like the Intraclass Correlation Coefficient (*ICC*). For narrower ordinal scales ($k \leq 4$), weighted categorical metrics (like Gwet's *AC2*) are strictly preferred over metric correlation techniques to avoid variance distortion. Main Source to Cite: Dealing with Narrow Ordinal Scales: Rhemtulla et al. (2012); Board Ordinal: Norman (2010), Sullivan & Artino (2013); Pseudo Metric: Finney & DiStefano (2006), Jensen et al. (2005), Lozano et al. (2008), Kvam et al. (2023)

4.2.5 Consistency vs. Agreement? What Does It Do?

Consistency measures whether raters rank objects in the same relative order (e.g., identifying differences systematically), effectively ignoring systematic offsets (bias). A rater who always scores exactly one point higher than a colleague will still show perfect consistency. Absolute agreement, however, requires raters to assign the exact same score, penalizing any level shifts or strictness differences. Comparing both mathematically reveals the exact nature of the measurement error.

4.2.6 Repeatability vs. Reproducibility

Repeatability (Intra-rater reliability) assesses the consistency of a single rater evaluating the same subjects multiple times under the same conditions. Reproducibility (Inter-rater reliability) evaluates the agreement between different raters evaluating the same subjects. Both are essential dimensions of a full Measurement Systems Analysis (MSA) to separate human cognitive drift from general scoring ambiguities. Main Source to Cite: AIAG-Work Group (2010)

4.2.7 What to Do When There Are More Than 2 Raters or Time Points?

For groupwise analyses ($m > 2$), pairwise metrics are insufficient to capture the overall system performance, as they only describe isolated fractions of the team. Omnibus measures like Fleiss' Kappa, Krippendorff's Alpha, and multi-rater Gwet's $AC1$ and $AC2$ are utilized to aggregate the overall agreement across the entire panel. Pairwise analysis should then be used selectively to trace back which specific rater combination causes the drop in the global coefficient.

References

- AIAG-Work Group. (2010). *Measurement systems analysis (MSA), reference manual* (4th ed.). Automotive Industry Action Group; Automotive Industry Action Group. <https://www.aiag.org/training-and-resources/manuals/details/MSA-4>
- Allaire, J. J., Teague, C., Xie, Y., Dervieux, C., & Woodhull, G. (2026). *Quarto* [Computer software]. Quarto. Zenodo. <https://doi.org/10.5281/zenodo.5960047>
- Allaire, J. J., Xie, Y., Dervieux, C., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., & Iannone, R. (2026). *Rmarkdown: Dynamic documents for r* [Computer software]. R Foundation for Statistical Computing. <https://cran.r-project.org/package=rmarkdown> <https://github.com/rstudio/rmarkdown>
- Bangdiwala, S. I., & Shankar, V. (2013). The agreement chart. *BMC Medical Research Methodology*, *13*(1), 97. <https://doi.org/10.1186/1471-2288-13-97>
- Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, *327*(8476), 307–310. [https://doi.org/10.1016/S0140-6736\(86\)90837-8](https://doi.org/10.1016/S0140-6736(86)90837-8)
- Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, *46*(5), 423–429. [https://doi.org/10.1016/0895-4356\(93\)90018-V](https://doi.org/10.1016/0895-4356(93)90018-V)
- Cicchetti, D. V., & Allison, T. (1971). A new procedure for assessing reliability of scoring EEG sleep recordings. *American Journal of EEG Technology*, *11*(3), 101–110. <https://doi.org/10.1080/00029238.1971.11080840>
- Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 551–558. [https://doi.org/10.1016/0895-4356\(90\)90159-M](https://doi.org/10.1016/0895-4356(90)90159-M)
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. <https://doi.org/10.1177/001316446002000104>
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220. <https://doi.org/10.1037/h0026256>
- Cohen, J. (1972). Weighted chi square: An extension of the kappa method. *Educational and Psychological Measurement*, *32*(1), 61–74. <https://doi.org/10.1177/001316447203200106>
- Deming, W. E. (1943). *Statistical adjustment of data*. J. Wiley & Sons, Incorporated. <https://books.google.de/books?id=r-I5AAAAMAAJ>
- Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 543–549. [https://doi.org/10.1016/0895-4356\(90\)90158-L](https://doi.org/10.1016/0895-4356(90)90158-L)
- Finney, S. J., & DiStefano, C. (2006). Nonnormal and categorical data in structural equation models. In G. R. Hancock & R. O. Mueller (Eds.), *A second course in structural equation modeling* (pp. 269–314). Information Age Publishing.
- Fisher, R. A. (1967). *Statistical methods for research workers* (13th ed.). Oliver; Boyd.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382. <https://doi.org/10.1037/h0031619>
- Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, *33*(3), 613–619. <https://doi.org/10.1177/001316447303300309>
- Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). John Wiley & Sons.
- Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). *Irr: Various coefficients of interrater reliability and agreement* [Computer software]. R Foundation for Statistical Computing. <https://doi.org/10.32614/CRAN.package.irr>
- Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 29–48. <https://doi.org/10.1348/000711006X126600>
- Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters* (4th ed.). Advanced Analytics, LLC.
- Gwet, K. L. (2019). *irrCAC: Computing chance-corrected agreement coefficients (CAC)* [Computer software]. R Foundation for Statistical Computing. <https://doi.org/10.32614/CRAN.package.irrCAC>
- Jensen, M. P., Martin, S. A., & Cheung, R. (2005). The meaning of pain relief in a clinical trial. *The Journal of*

- Pain*, 6(6), 400–406. <https://doi.org/10.1016/j.jpain.2005.01.360>
- Kendall, M. G., & Smith, B. B. (1939). The problem of m rankings. *The Annals of Mathematical Statistics*, 10(3), 275–287. <https://doi.org/10.1214/aoms/1177732186>
- Krippendorff, K. (2004). Reliability in content analysis. *Human Communication Research*, 30(3), 411–433. <https://doi.org/10.1111/j.1468-2958.2004.tb00738.x>
- Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (4th ed., p. 472). Sage.
- Kvam, P. D., Marley, A. A. J., & Heathcote, A. (2023). A unified theory of discrete and continuous responding. *Psychological Review*, 130(2), 368–400. <https://doi.org/10.1037/rev0000378>
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159. <https://doi.org/10.2307/2529310>
- Legendre, P. (2005). Species associations: The kendall coefficient of concordance revisited. *Journal of Agricultural, Biological, and Environmental Statistics*, 10(2), 226–245. <https://doi.org/10.1198/108571105X46642>
- Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology*, 4(2), 73–79. <https://doi.org/10.1027/1614-2241.4.2.73>
- McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3), 276–282. <https://doi.org/10.11613/BM.2012.031>
- Meyer, D., Zeileis, A., & Hornik, K. (2006). The strucplot framework: Visualizing multi-way contingency tables with vcd. *Journal of Statistical Software*, 17(3), 1–48. <https://doi.org/10.18637/jss.v017.i03>
- Meyer, D., Zeileis, A., Hornik, K., & Friendly, M. (2002). *Vcd: Visualizing categorical data* [Computer software]. R Foundation for Statistical Computing. <https://doi.org/10.32614/CRAN.package.vcd>
- Munoz, S. R., & Bangdiwala, S. I. (1997). Interpretation of kappa and b statistics measures of agreement. *Journal of Applied Statistics*, 24(1), 105–112. <https://doi.org/10.1080/02664769723918>
- Norman, G. (2010). Likert scales, levels of measurement and the “laws” of statistics. *Advances in Health Sciences Education*, 15(5), 625–632. <https://doi.org/10.1007/s10459-010-9222-y>
- Oksanen, J., Simpson, G. L., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O’Hara, R. B., Solymos, P., Stevens, M. H. H., Szoecs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Borman, T., Carvalho, G., Chirico, M., De Caceres, M., ... Weedon, J. (2026). *Vegan: Community ecology package* [Computer software]. R Foundation for Statistical Computing. <https://doi.org/10.32614/CRAN.package.vegan>
- Ooms, J., & McNamara, J. (2024). *Writexl: Export data frames to excel “xlsx” format (version 1.5.1)* [Computer software]. R Foundation for Statistical Computing. <https://doi.org/10.32614/CRAN.package.writexl>
- Posit Team. (2026). *RStudio: Integrated development environment for r* [Computer software]. Posit Software, PBC. <http://www.posit.co/>
- R Core Team. (2026). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. <https://www.r-project.org/>
- Revelle, W. (2026). *Psych: Procedures for psychological, psychometric, and personality research* [Computer software]. R Foundation for Statistical Computing. <https://cran.r-project.org/package=psych>
- Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, 17(3), 354–373. <https://doi.org/10.1037/a0029315>
- Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6), 657–680. <https://doi.org/10.1007/BF01068419>
- Stephens, J., & Simonov, K. (2025). *Yaml: Methods to convert r data to YAML and back* [Computer software]. R Foundation for Statistical Computing. <https://doi.org/10.32614/CRAN.package.yaml>
- Stokes, M., & Koch, G. (2011). Up to speed with categorical data analysis. *SAS Global Forum 2011 - Statistics and Data Analysis*.
- Sullivan, G. M., & Artino, A. R. (2013). Analyzing and interpreting data from likert-type scales. *Journal of Graduate Medical Education*, 5(4), 541–542. <https://doi.org/10.4300/JGME-5-4-18>
- Walker, E., & Nowacki, A. S. (2011). Understanding equivalence and noninferiority testing. *Journal of General Internal Medicine*, 26(2), 192–196. <https://doi.org/10.1007/s11606-010-1513-8>
- Wickham, H., & Bryan, J. (2025). *Readxl: Read excel files* [Computer software]. R Foundation for Statistical

- Computing. <https://doi.org/10.32614/CRAN.package.readxl>
- Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D., & Brand, T. van den. (2026). *ggplot2: Create elegant data visualisations using the grammar of graphics* [Computer software]. R Foundation for Statistical Computing. <https://cran.r-project.org/package=ggplot2>
- Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2026). *Dplyr: A grammar of data manipulation* [Computer software]. R Foundation for Statistical Computing. <https://doi.org/10.32614/CRAN.package.dplyr>
- Wickham, H., Vaughan, D., & Girlich, M. (2025). *Tidyr: Tidy messy data* [Computer software]. R Foundation for Statistical Computing. <https://doi.org/10.32614/CRAN.package.tidyr>
- Xie, Y. (2025). *Knitr: A general-purpose package for dynamic report generation in r* [Computer software]. R Foundation for Statistical Computing. <https://yihui.org/knitr/> <https://cran.r-project.org/package=knitr>
- Xie, Y. (2026). *Tinytex: Helper functions to install and maintain TeX live, and compile LaTeX documents* [Computer software]. Comprehensive R Archive Network; R Foundation for Statistical Computing. <https://doi.org/10.32614/CRAN.package.tinytex>