

MSA for Ordinal Scoring Data

Attributive Gage R&R and Generation of “Noisy Labels” for Machine Learning

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

2026-07-02

Table of contents

1 Methodological Setup of the Current Analysis	3
1.1 Project Description: Attributive Gage R&R Study (Ordinal Scales)	3
1.2 Scale & Weighting Setup	3
1.3 Rater & Trial Mapping	3
1.4 Suggested Draft for the Materials and Methods Section	3
2 Measurement System Analysis (MSA): Testing against Global Consensus	5
2.1 Visual Exploration: Systematic Bias (Variability Charts)	5
2.1.1 Global Appraiser Bias	5
2.1.2 Temporal Appraiser Bias	5
2.1.3 Severity-dependent Appraiser Bias	7
2.1.4 Full Variability Chart (Systematic Error)	7
2.2 Repeatability (Time- / Intra-Rater-Variation)	8
2.3 Reproducibility (Appraiser- / Inter-Rater-Variation)	8
2.3.1 Global System Reproducibility (All Raters)	8
2.3.2 Pairwise Reproducibility (Rater vs. Rater)	8
2.4 Accuracy (Agreement with Global Consensus)	9
2.4.1 Global Accuracy (All Trials Aggregated)	9
2.4.2 Temporal Accuracy (Separated by Trial)	9
3 Measurement System Analysis (MSA): Testing against Defined Standard Rater	11
3.1 Visual Exploration: Systematic Bias against Standard Rater	11
3.1.1 Global Appraiser Bias	11
3.1.2 Temporal Appraiser Bias	12
3.1.3 Severity-dependent Appraiser Bias	13
3.1.4 Full Variability Chart (Systematic Error against Standard)	13
3.2 Global Accuracy against Standard Rater (All Trials Aggregated)	14
3.3 Temporal Accuracy against Standard Rater (Separated by Trial)	14

4 Methodological Classification & Decision-Making Guide	16
4.1 Attributive Gage R&R: Repeatability vs. Reproducibility	16
4.2 Weighted Agreement (Cohen’s Kappa)	16
4.3 Gwet’s AC2 and the Prevalence Paradox	16
4.4 Automated Bias Detection: Kendall’s W vs. Absolute Agreement	16
4.5 The Global Consensus (Noisy Labels)	16
References	17

1 Methodological Setup of the Current Analysis

1.1 Project Description: Attributive Gage R&R Study (Ordinal Scales)

This Measurement System Analysis (MSA) evaluates the reliability of a visual scoring system. It partitions the measurement error into Repeatability (Intra-Rater consistency across trials) and Reproducibility (Inter-Rater agreement). The analysis utilizes weighted Cohen's Kappa and prevalence-adjusted Gwet's AC for ordinal data.

1.2 Scale & Weighting Setup

- **Scale Range:** Ordinal Scale from 1 to 5
- **Applied Penalty Weighting:** Custom
- **Interpretation Benchmark:** Mchugh
- **Confidence Intervals:** Percentile Bootstrapping with 100 Iterations
- **Control Limit (Base Tolerance):** ± 0.5 (scaled dynamically by $1/\sqrt{N}$)

1.3 Rater & Trial Mapping

The following structure was automatically detected from your configuration:

- **Rater_A** evaluated 3 trials: (A_Trial_1, A_Trial_2, A_Trial_3)
- **Rater_B** evaluated 3 trials: (B_Trial_1, B_Trial_2, B_Trial_3)
- **Rater_C** evaluated 3 trials: (C_Trial_1, C_Trial_2, C_Trial_3)

1.4 Suggested Draft for the Materials and Methods Section

To evaluate the reliability of the scoring system, an Attributive Gage R&R analysis was performed (AIAG-Work Group, 2010). The assessments of 3 independent appraisers, each evaluating the subjects across up to 3 trials (time points), were compared. The underlying variable was treated as an ordinal ranked scale with theoretical score levels ranging from 1 to 5. A custom weighting matrix was applied to penalize specific misclassifications according to their clinical or practical relevance. The measurement error was partitioned into Repeatability (Intra-rater consistency across multiple trials) and Reproducibility (Inter-rater agreement) (AIAG-Work Group, 2010). To isolate reproducibility from intra-rater noise, agreement between different raters was computed using purified rater medians. Absolute agreement was quantified utilizing Cohen's Kappa (Cohen, 1968) and prevalence-adjusted Gwet's AC2 Gwet (2014). Systematic bias (directional error) was visually explored using nested variability charts. To differentiate between random noise and systematic decalibration, a statistical control limit based on a base tolerance of ± 0.5 score units, adjusted by the sample size ($\pm 0.5/\sqrt{n}$), was applied. Furthermore, an automated bias detection algorithm evaluating Kendall's W (rank consistency) (Kendall & Smith, 1939) against unweighted Cohen's Kappa (absolute agreement) (Cohen, 1960) was implemented. Additionally, accuracy and bias were evaluated against a defined standard rater (Rater_A) acting as an assumed training baseline. 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 Mchugh (McHugh, 2012).

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) and `irrCAC` (Gwet, 2019): Core packages for calculating Kappa measures, Kendall's W, and Gwet's AC; `psych` (Revelle, 2026): Enables the use of complex, user-defined weighting matrices for weighted Kappa. 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).

2 Measurement System Analysis (MSA): Testing against Global Consensus

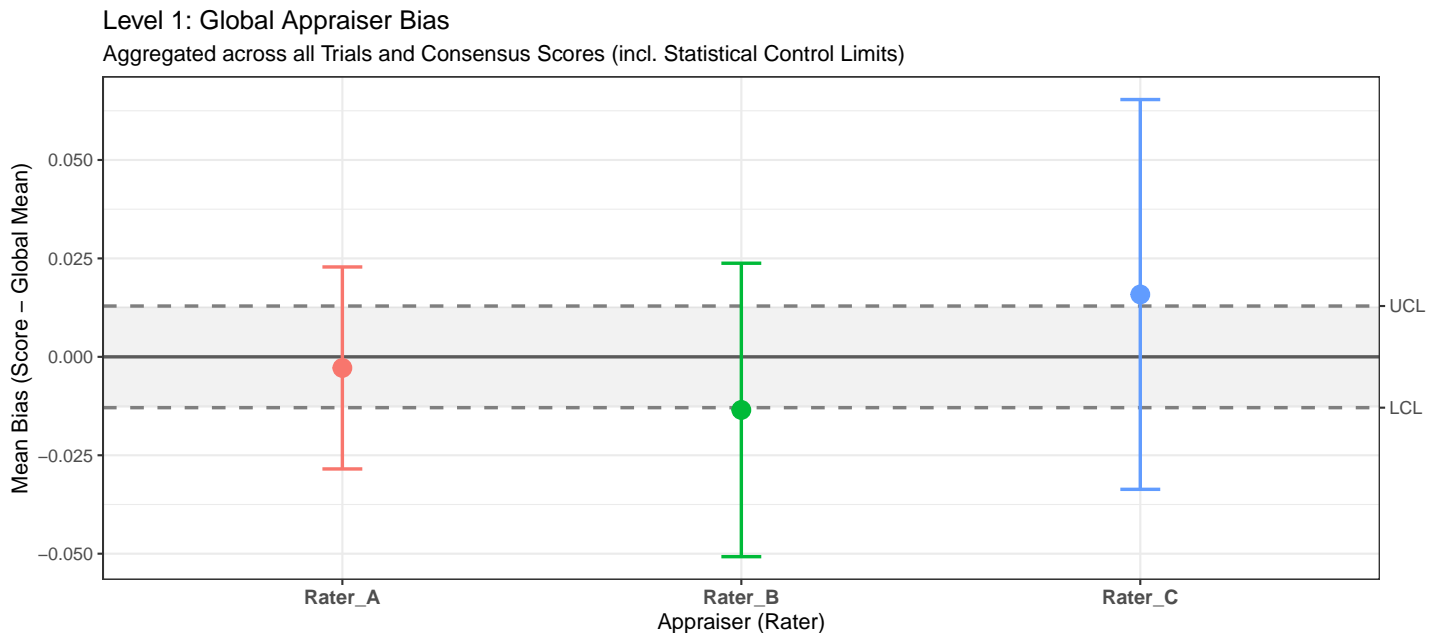
2.1 Visual Exploration: Systematic Bias (Variability Charts)

The following section utilizes a drill-down approach to decompose the systematic bias (directional error) of the measurement system. Bias is defined as the deviation of a single rating from the **Global Consensus Mean**. But be careful, this is **not a real Gold Standard**. A bias of 0 may indicate a perfect calibration. A positive value may imply a tendency to over-score (e.g., assessing too severely), whereas a negative value may indicate an under-scoring.

*Note: The gray shaded areas in the aggregate charts represent a kind of **Statistical Control Limit (UCL / LCL)**. The limit lines are derived from the given base tolerance (± 0.5) adjusted by the sample size according to the square root of N law ($\pm 0.5/\sqrt{n}$). Values outside this range may indicate a relevant systematic decalibration. However, caution is advised here as well, since there is no true gold standard for comparison. The appraiser with the smaller error bars may be more reliable.*

2.1.1 Global Appraiser Bias

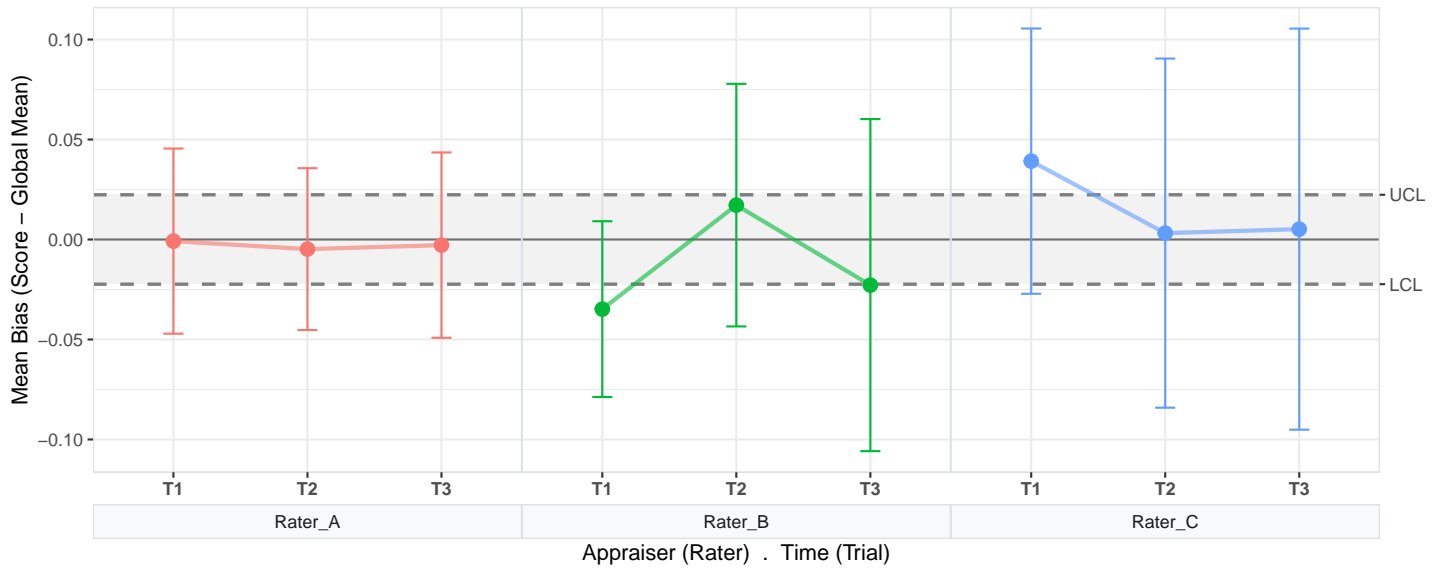
The most aggregated view. It highlights whether a specific appraiser exhibits a systematic strictness or leniency across all evaluated subjects and time points.



2.1.2 Temporal Appraiser Bias

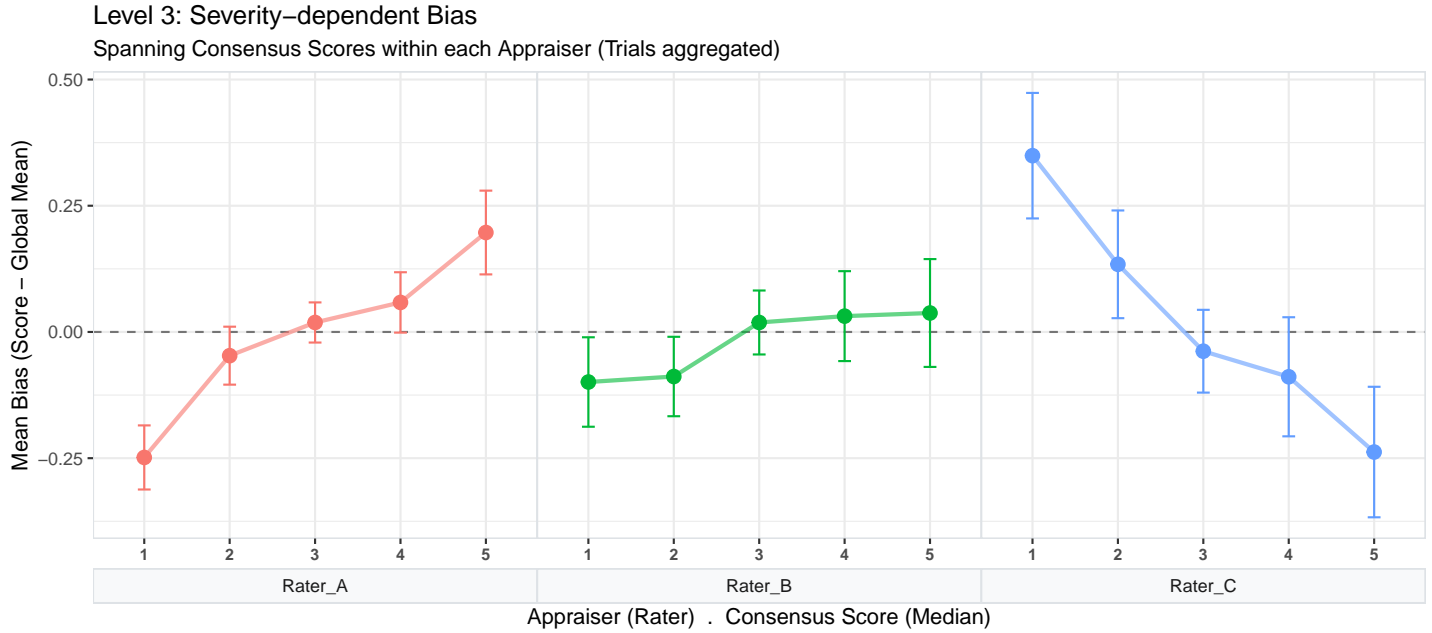
This view uncovers temporal instability. Relevant slopes between the different trials indicate cognitive **fatigue** if they deviate from the baseline, **learning** effects if they converge toward the baseline, or general **shifts** in the rater's scoring if the mean values fall outside the control limits.

Level 2: Temporal Appraiser Bias
 Spanning Trials within each Appraiser (incl. Statistical Control Limits)



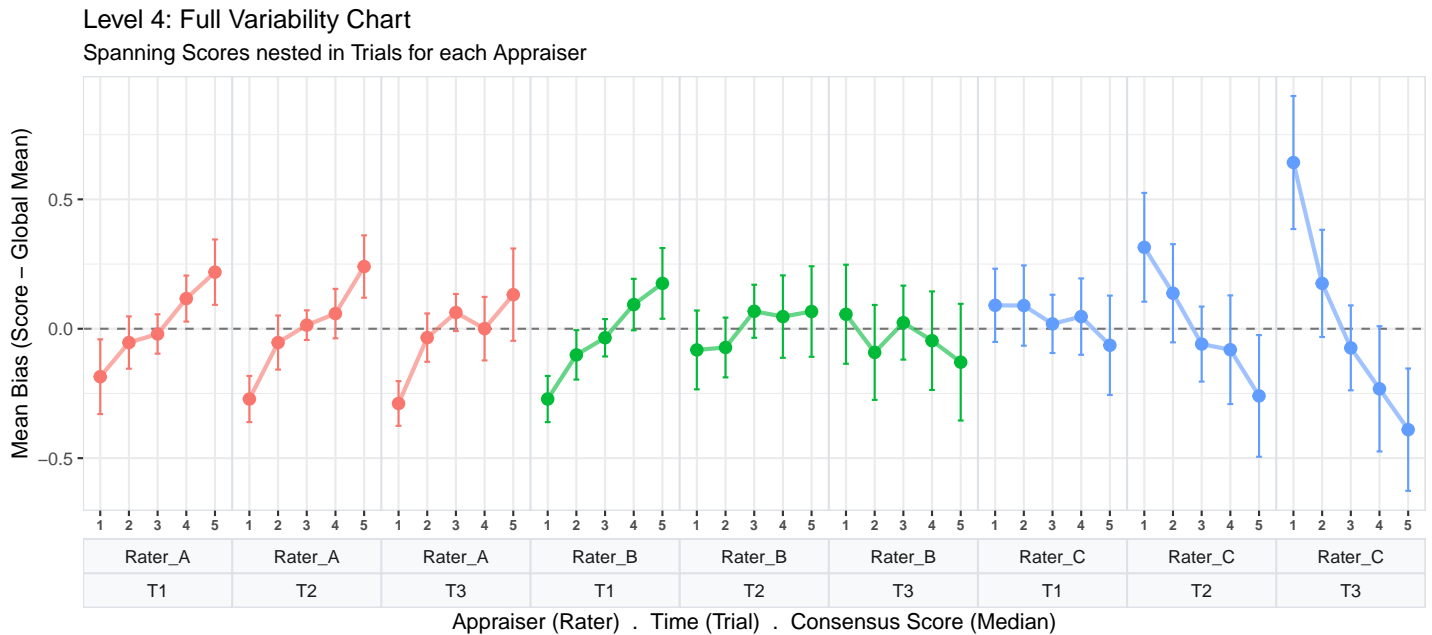
2.1.3 Severity-dependent Appraiser Bias

This plot decomposes the bias along the scale itself. It reveals whether an appraiser is perfectly calibrated for healthy subjects (e.g. Score 1) but struggles with correct classification at higher severity levels.



2.1.4 Full Variability Chart (Systematic Error)

The most detailed view. It combines all previous dimensions, visualizing the error profile for every specific combination of Appraiser, Time Point, and Severity.



2.2 Repeatability (Time- / Intra-Rater-Variation)

Repeatability measures the internal consistency of **each rater across their own multiple trials**. A low score here indicates that the rater is guessing or the scoring criteria are highly ambiguous.

Rater	Metric	Value	CI	Interpretation
Rater_A	Cohen's Kappa (w)	0.889	[0.8597; 0.9214]	Strong
Rater_A	Gwet's AC2	0.930	[0.9148; 0.9481]	Almost Perfect
Rater_B	Cohen's Kappa (w)	0.669	[0.6147; 0.7149]	Moderate
Rater_B	Gwet's AC2	0.728	[0.6862; 0.7628]	Moderate
Rater_C	Cohen's Kappa (w)	0.561	[0.5174; 0.6051]	Weak
Rater_C	Gwet's AC2	0.548	[0.5045; 0.5938]	Weak

2.3 Reproducibility (Appraiser- / Inter-Rater-Variation)

Reproducibility measures the agreement **between different raters**. To isolate this from intra-rater noise, this framework first calculates the median consensus for each rater across their trials, and then computes the agreement between these purified rater opinions.

2.3.1 Global System Reproducibility (All Raters)

Metric	Value	CI	Interpretation
Cohen's Kappa (w)	0.461	[0.3949; 0.5125]	Weak
Gwet's AC2	0.552	[0.5099; 0.5917]	Weak

2.3.2 Pairwise Reproducibility (Rater vs. Rater)

Comparison	Metric	Value	CI	Interpretation
Rater_A vs. Rater_B	Cohen's Kappa (w)	0.733	[0.6694; 0.7861]	Moderate
Rater_A vs. Rater_B	Gwet's AC2	0.818	[0.7757; 0.8603]	Strong
Rater_A vs. Rater_C	Cohen's Kappa (w)	0.373	[0.2944; 0.4326]	Minimal
Rater_A vs. Rater_C	Gwet's AC2	0.477	[0.4113; 0.5438]	Weak
Rater_B vs. Rater_C	Cohen's Kappa (w)	0.276	[0.2020; 0.3416]	Minimal
Rater_B vs. Rater_C	Gwet's AC2	0.346	[0.2775; 0.4197]	Minimal

2.4 Accuracy (Agreement with Global Consensus)

In the absence of a true, objective Gold Standard, the **Global Consensus Median** serves as the most robust estimate of the true underlying severity. This section evaluates the exact agreement (Accuracy) of each appraiser against this mathematical consensus, utilizing the specified penalty weightings.

2.4.1 Global Accuracy (All Trials Aggregated)

This perspective calculates an aggregated score (median) for each rater across their trials before comparing it to the Global Consensus. It filters out intra-rater noise to reveal the true baseline capability of each appraiser.

Rater	Metric	Value	CI	Interpretation
Rater_A	Cohen's Kappa (w)	0.960	[0.9333; 0.9775]	Almost Perfect
Rater_A	Gwet's AC2	0.976	[0.9612; 0.9850]	Almost Perfect
Rater_B	Cohen's Kappa (w)	0.760	[0.7102; 0.8091]	Moderate
Rater_B	Gwet's AC2	0.834	[0.7920; 0.8647]	Strong
Rater_C	Cohen's Kappa (w)	0.407	[0.3477; 0.4739]	Weak
Rater_C	Gwet's AC2	0.502	[0.4385; 0.5501]	Weak

Bias Analysis (Consistency vs. Absolute Agreement)

- **Rater_A:** No strong systematic bias: The appraiser shows both good consistency and acceptable absolute agreement. (*Kendall's W: 0.985* / *Kappa (unw.): 0.926*)
- **Rater_B:** No strong systematic bias: The appraiser shows both good consistency and acceptable absolute agreement. (*Kendall's W: 0.922* / *Kappa (unw.): 0.708*)
- **Rater_C:** Strong systematic bias detected: The appraiser is highly consistent in ranking but fails to hit the exact consensus level. (*Kendall's W: 0.788* / *Kappa (unw.): 0.291*)

2.4.2 Temporal Accuracy (Separated by Trial)

This perspective calculates the accuracy of each rater separately for every single time point (Trial). It reveals whether a rater's agreement with the team standard improves over time (e.g. through training effects) or deteriorates (e.g. due to fatigue).

Rater	Trial	Metric	Value	CI	Interpretation
Rater_A	T1	Cohen's Kappa (w)	0.900	[0.8617; 0.9346]	Almost Perfect
Rater_A	T1	Gwet's AC2	0.938	[0.9133; 0.9606]	Almost Perfect
Rater_A	T2	Cohen's Kappa (w)	0.940	[0.9137; 0.9608]	Almost Perfect
Rater_A	T2	Gwet's AC2	0.963	[0.9477; 0.9766]	Almost Perfect
Rater_A	T3	Cohen's Kappa (w)	0.910	[0.8784; 0.9405]	Almost Perfect
Rater_A	T3	Gwet's AC2	0.944	[0.9207; 0.9609]	Almost Perfect
Rater_B	T1	Cohen's Kappa (w)	0.886	[0.8435; 0.9186]	Strong
Rater_B	T1	Gwet's AC2	0.926	[0.9036; 0.9515]	Almost Perfect
Rater_B	T2	Cohen's Kappa (w)	0.722	[0.6640; 0.7710]	Moderate
Rater_B	T2	Gwet's AC2	0.808	[0.7654; 0.8419]	Strong
Rater_B	T3	Cohen's Kappa (w)	0.508	[0.4536; 0.5693]	Weak
Rater_B	T3	Gwet's AC2	0.603	[0.5238; 0.6657]	Moderate
Rater_C	T1	Cohen's Kappa (w)	0.557	[0.4879; 0.6237]	Weak
Rater_C	T1	Gwet's AC2	0.667	[0.6245; 0.7129]	Moderate
Rater_C	T2	Cohen's Kappa (w)	0.359	[0.3032; 0.4362]	Minimal
Rater_C	T2	Gwet's AC2	0.433	[0.3771; 0.5050]	Weak
Rater_C	T3	Cohen's Kappa (w)	0.242	[0.1852; 0.3001]	Minimal
Rater_C	T3	Gwet's AC2	0.297	[0.2224; 0.3671]	Minimal

Bias Analysis by Trial

- **Rater_A (T1):** No strong systematic bias: The appraiser shows both good consistency and acceptable absolute agreement. (*Kendall's W: 0.964 | Kappa (unw.): 0.873*)
- **Rater_A (T2):** No strong systematic bias: The appraiser shows both good consistency and acceptable absolute agreement. (*Kendall's W: 0.975 | Kappa (unw.): 0.899*)
- **Rater_A (T3):** No strong systematic bias: The appraiser shows both good consistency and acceptable absolute agreement. (*Kendall's W: 0.955 | Kappa (unw.): 0.865*)
- **Rater_B (T1):** No strong systematic bias: The appraiser shows both good consistency and acceptable absolute agreement. (*Kendall's W: 0.965 | Kappa (unw.): 0.863*)
- **Rater_B (T2):** No strong systematic bias: The appraiser shows both good consistency and acceptable absolute agreement. (*Kendall's W: 0.907 | Kappa (unw.): 0.659*)
- **Rater_B (T3):** No strong systematic bias: The appraiser shows both good consistency and acceptable absolute agreement. (*Kendall's W: 0.840 | Kappa (unw.): 0.418*)
- **Rater_C (T1):** No strong systematic bias: The appraiser shows both good consistency and acceptable absolute agreement. (*Kendall's W: 0.847 | Kappa (unw.): 0.472*)
- **Rater_C (T2):** Strong systematic bias detected: The appraiser is highly consistent in ranking but fails to hit the exact consensus level. (*Kendall's W: 0.770 | Kappa (unw.): 0.249*)
- **Rater_C (T3):** Strong systematic bias detected: The appraiser is highly consistent in ranking but fails to hit the exact consensus level. (*Kendall's W: 0.723 | Kappa (unw.): 0.129*)

3 Measurement System Analysis (MSA): Testing against Defined Standard Rater

This optional chapter explores a “what-if” scenario: What if we assume that one specific appraiser (** Rater_A **) represents the absolute gold standard? The median score of this chosen rater across their trials serves as the reference benchmark for all appraisers.

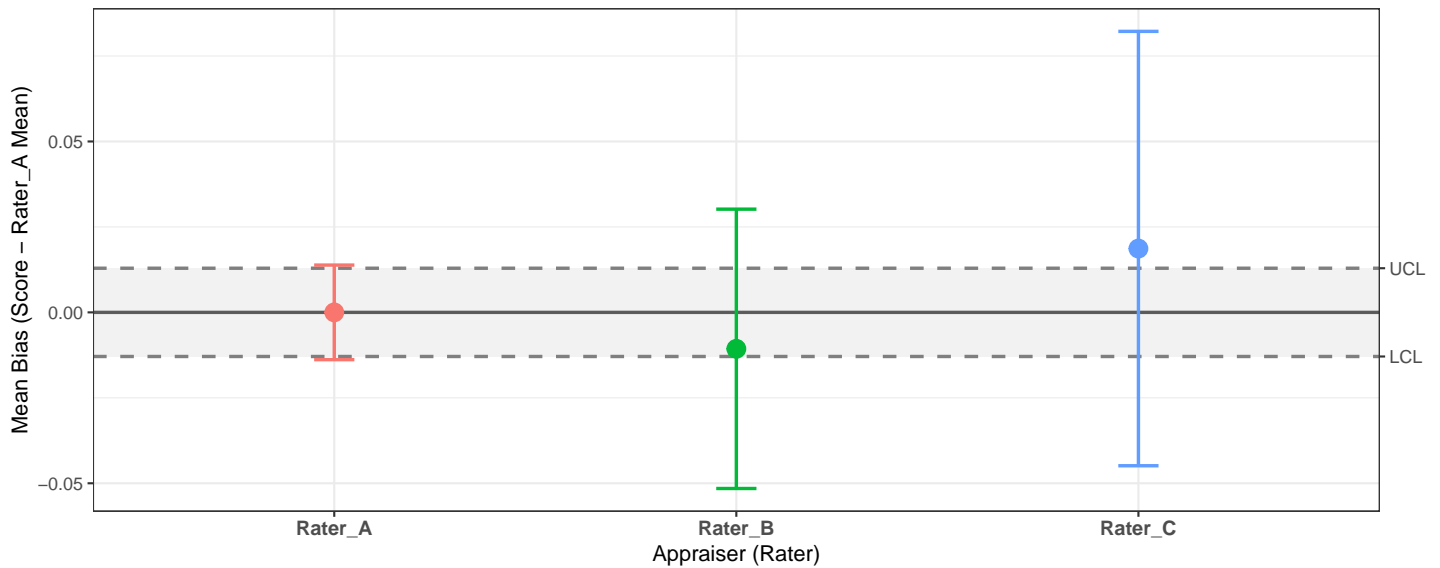
Note: In practical ethology and quality control, supposedly trained or senior personnel are not inherently or infallibly correct. This comparison should be treated strictly as an exploratory tool to visualize alignment with a defined training standard, rather than an absolute measure of ground truth.

3.1 Visual Exploration: Systematic Bias against Standard Rater

3.1.1 Global Appraiser Bias

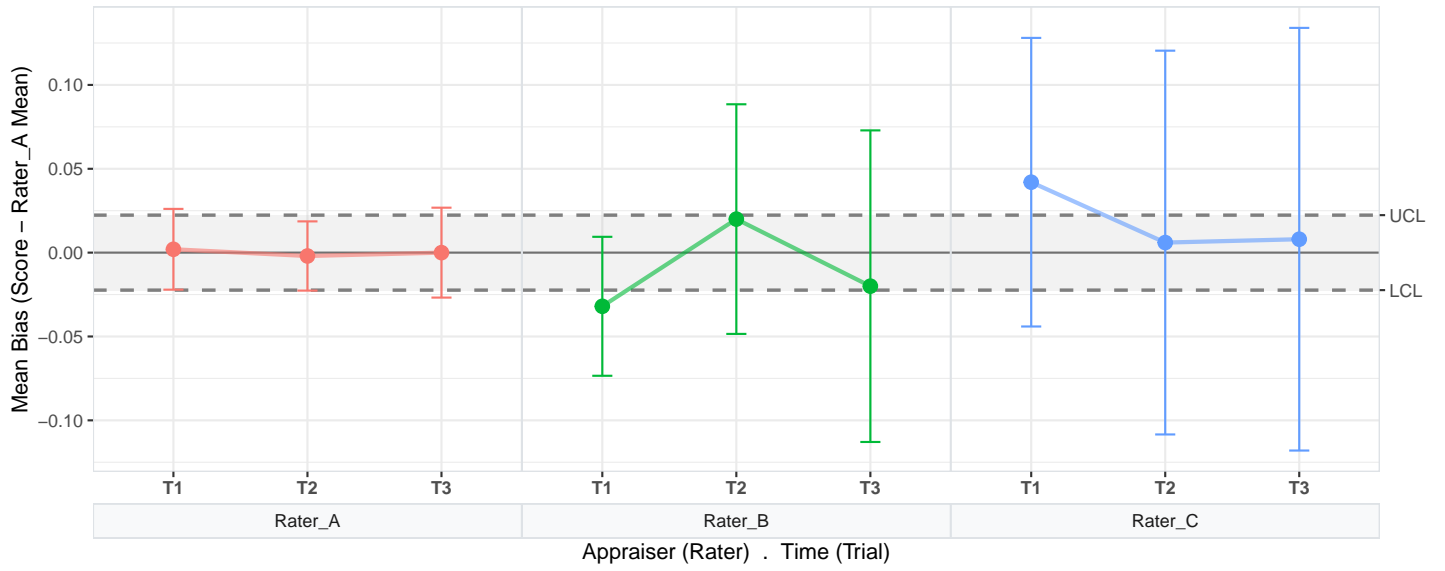
Level 1: Global Bias against Standard Rater

Standard Rater: Rater_A (Baseline represents mathematically 0 systematic bias)



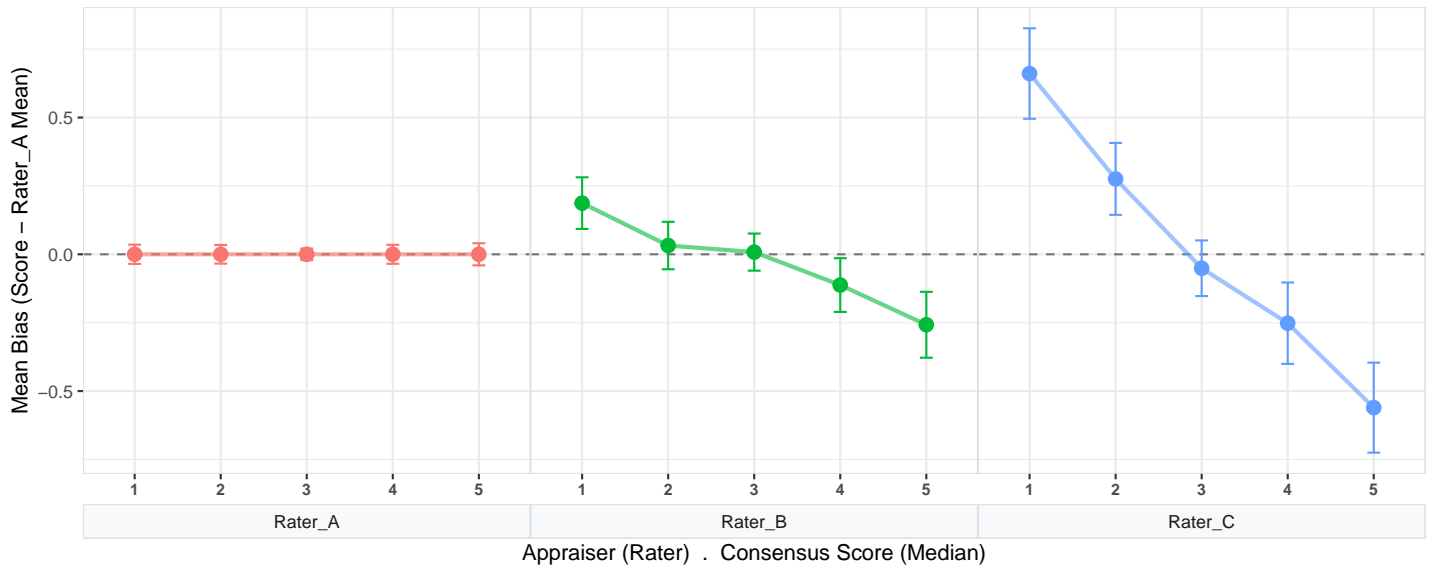
3.1.2 Temporal Appraiser Bias

Level 2: Temporal Bias against Standard Rater
Standard Rater: Rater_A



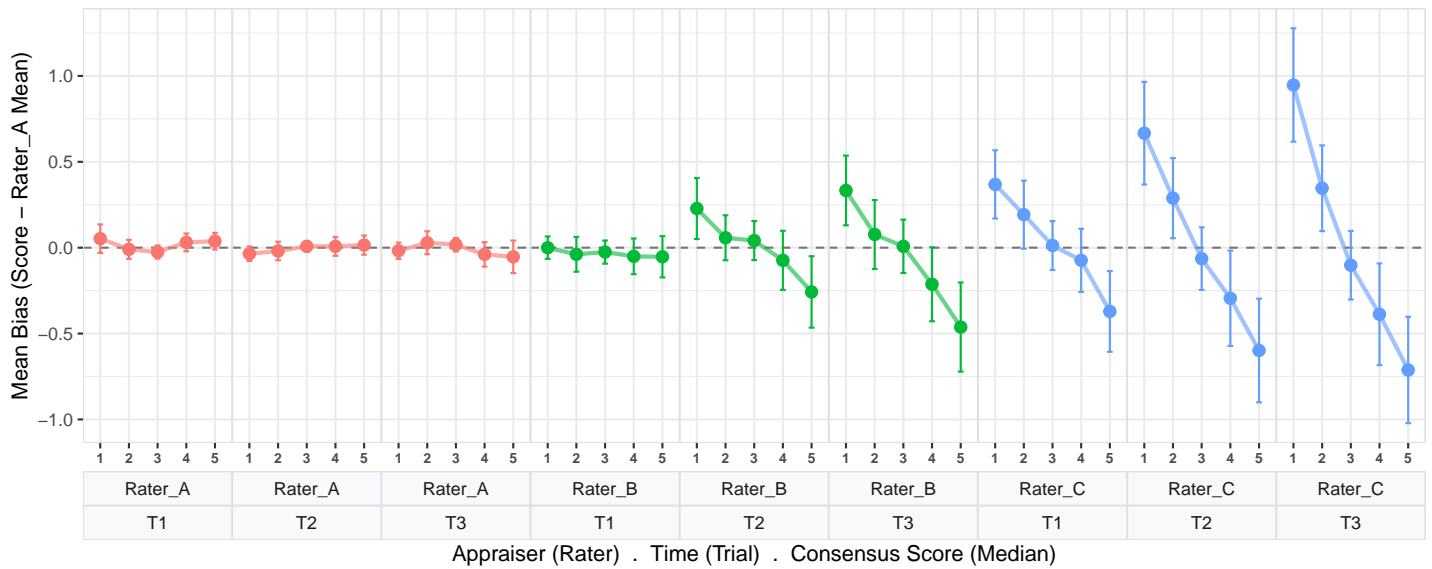
3.1.3 Severity-dependent Appraiser Bias

Level 3: Severity-dependent Bias against Standard Rater
 Standard Rater: Rater_A



3.1.4 Full Variability Chart (Systematic Error against Standard)

Level 4: Full Variability Chart
 Standard Rater: Rater_A



3.2 Global Accuracy against Standard Rater (All Trials Aggregated)

Rater	Metric	Value	CI	Interpretation
Rater_A	Cohen's Kappa (w)	1.000	[1.0000; 1.0000]	Almost Perfect
Rater_A	Gwet's AC2	1.000	[1.0000; 1.0000]	Almost Perfect
Rater_B	Cohen's Kappa (w)	0.733	[0.6817; 0.7845]	Moderate
Rater_B	Gwet's AC2	0.818	[0.7868; 0.8540]	Strong
Rater_C	Cohen's Kappa (w)	0.373	[0.2916; 0.4448]	Minimal
Rater_C	Gwet's AC2	0.477	[0.4160; 0.5553]	Weak

Bias Analysis against Standard Rater

- **Rater_A:** No strong systematic bias: The appraiser shows both good consistency and acceptable absolute agreement. (*Kendall's W: 1.000 / Kappa (unw.): 1.000*)
- **Rater_B:** No strong systematic bias: The appraiser shows both good consistency and acceptable absolute agreement. (*Kendall's W: 0.907 / Kappa (unw.): 0.652*)
- **Rater_C:** Strong systematic bias detected: The appraiser is highly consistent in ranking but fails to hit the exact consensus level. (*Kendall's W: 0.767 / Kappa (unw.): 0.248*)

3.3 Temporal Accuracy against Standard Rater (Separated by Trial)

Rater	Trial	Metric	Value	CI	Interpretation
Rater_A	T1	Cohen's Kappa (w)	0.933	[0.8979; 0.9702]	Almost Perfect
Rater_A	T1	Gwet's AC2	0.959	[0.9398; 0.9752]	Almost Perfect
Rater_A	T2	Cohen's Kappa (w)	0.976	[0.9583; 0.9896]	Almost Perfect
Rater_A	T2	Gwet's AC2	0.985	[0.9740; 0.9931]	Almost Perfect
Rater_A	T3	Cohen's Kappa (w)	0.922	[0.8817; 0.9554]	Almost Perfect
Rater_A	T3	Gwet's AC2	0.952	[0.9292; 0.9732]	Almost Perfect
Rater_B	T1	Cohen's Kappa (w)	0.880	[0.8446; 0.9216]	Strong
Rater_B	T1	Gwet's AC2	0.923	[0.8944; 0.9530]	Almost Perfect
Rater_B	T2	Cohen's Kappa (w)	0.686	[0.6226; 0.7552]	Moderate
Rater_B	T2	Gwet's AC2	0.785	[0.7460; 0.8234]	Moderate
Rater_B	T3	Cohen's Kappa (w)	0.484	[0.4269; 0.5425]	Weak
Rater_B	T3	Gwet's AC2	0.586	[0.5210; 0.6462]	Weak
Rater_C	T1	Cohen's Kappa (w)	0.539	[0.4727; 0.6013]	Weak
Rater_C	T1	Gwet's AC2	0.657	[0.6065; 0.7063]	Moderate
Rater_C	T2	Cohen's Kappa (w)	0.327	[0.2712; 0.4071]	Minimal
Rater_C	T2	Gwet's AC2	0.408	[0.3319; 0.4646]	Weak
Rater_C	T3	Cohen's Kappa (w)	0.226	[0.1680; 0.2939]	Minimal
Rater_C	T3	Gwet's AC2	0.285	[0.2142; 0.3407]	Minimal

Bias Analysis by Trial against Standard Rater

- **Rater_A (T1):** No strong systematic bias: The appraiser shows both good consistency and acceptable absolute agreement. (*Kendall's W: 0.974* | *Kappa (unw.): 0.932*)
- **Rater_A (T2):** No strong systematic bias: The appraiser shows both good consistency and acceptable absolute agreement. (*Kendall's W: 0.988* | *Kappa (unw.): 0.967*)
- **Rater_A (T3):** No strong systematic bias: The appraiser shows both good consistency and acceptable absolute agreement. (*Kendall's W: 0.965* | *Kappa (unw.): 0.910*)
- **Rater_B (T1):** No strong systematic bias: The appraiser shows both good consistency and acceptable absolute agreement. (*Kendall's W: 0.961* | *Kappa (unw.): 0.865*)
- **Rater_B (T2):** No strong systematic bias: The appraiser shows both good consistency and acceptable absolute agreement. (*Kendall's W: 0.890* | *Kappa (unw.): 0.602*)
- **Rater_B (T3):** Strong systematic bias detected: The appraiser is highly consistent in ranking but fails to hit the exact consensus level. (*Kendall's W: 0.822* | *Kappa (unw.): 0.379*)
- **Rater_C (T1):** No strong systematic bias: The appraiser shows both good consistency and acceptable absolute agreement. (*Kendall's W: 0.838* | *Kappa (unw.): 0.440*)
- **Rater_C (T2):** Strong systematic bias detected: The appraiser is highly consistent in ranking but fails to hit the exact consensus level. (*Kendall's W: 0.746* | *Kappa (unw.): 0.211*)
- **Rater_C (T3):** Strong systematic bias detected: The appraiser is highly consistent in ranking but fails to hit the exact consensus level. (*Kendall's W: 0.706* | *Kappa (unw.): 0.117*)

4 Methodological Classification & Decision-Making Guide

4.1 Attributive Gage R&R: Repeatability vs. Reproducibility

In visually assessed ethological or clinical data (e.g., lameness scoring), the human observer acts as the measurement instrument. An Attributive Gage R&R study systematically partitions the human measurement error into two distinct dimensions (AIAG-Work Group, 2010): * **Repeatability (Intra-Rater Reliability)**: The internal consistency of a single appraiser when evaluating the same subject multiple times. Low repeatability indicates ambiguity in the scoring criteria or cognitive fatigue. * **Reproducibility (Inter-Rater Reliability)**: The agreement between different appraisers. To isolate this metric from individual noise, this framework evaluates the purified rater medians against each other.

4.2 Weighted Agreement (Cohen's Kappa)

While the unweighted Cohen's Kappa evaluates strictly exact matches (Cohen, 1960), it is inadequate for ordinal scales. In livestock monitoring, confusing a severe pathology (Score 5) with a healthy state (Score 1) is substantially more detrimental than a minor one-point deviation. By applying linear or quadratic penalty matrices Fleiss & Cohen (1973), weighted Kappa appropriately accounts for the magnitude of disagreement, mirroring the practical severity of misclassifications.

4.3 Gwet's AC2 and the Prevalence Paradox

A well-documented limitation of Kappa is the *Prevalence Paradox* Cicchetti & Feinstein (1990). In highly homogeneous herds (e.g., 90% of animals are perfectly healthy), the mathematical expectation of chance agreement (p_e) inflates. Consequently, even minor deviations by the raters cause Cohen's Kappa to drop disproportionately close to zero, falsely suggesting a flawed measurement system. Gwet's AC2 Gwet (2014) stabilizes the chance-agreement probability under such skewed distributions, providing a more robust reliability metric for typical field data.

4.4 Automated Bias Detection: Kendall's W vs. Absolute Agreement

A common systemic error is a fixed rater bias (e.g., an appraiser consistently scoring one level too strictly). Kendall's Coefficient of Concordance (W) evaluates pure rank consistency (Kendall & Smith, 1939), thereby ignoring these fixed baseline shifts. By contrasting (Feinstein & Cicchetti (1990), Cicchetti & Feinstein (1990)) Kendall's W with unweighted Cohen's Kappa (which penalizes shifts), the framework distinguishes between erratic guessing (both metrics low) and a systematic calibration error (Kendall high, Kappa low).

4.5 The Global Consensus (Noisy Labels)

Because visual scoring lacks a true, objective gold standard, human evaluations must be treated as *Noisy Labels*. By computing the mathematical median across all appraisers and trials, individual cognitive noise is filtered out, establishing a robust **Global Consensus**. This consensus serves as the most reliable ground truth available for downstream Machine Learning applications or the training of diagnostic sensor systems.

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>
- 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>
- 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)
- 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>
- 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>
- 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>
- McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282. <https://doi.org/10.11613/BM.2012.031>
- 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>
- 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>
- 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>