



Inter-Rater Reliability in Social Sciences: A Comprehensive Review of Methods, Challenges, and Implications

Debadeepa Banerjee

Department of Psychology, Bethune College, Kolkata, India. e-mail: debadeepa@gmail.com

Ganesh Dutta

Department of Statistics, Basanti Devi College, Kolkata, India. e-mail: duttaganesh78@gmail.com (Corresponding Author)

ABSTRACT

Inter-rater reliability (IRR) is a crucial aspect of social science research, ensuring consistency and reproducibility across different evaluators. This review explores key statistical methods, including Bland-Altman Analysis, Intraclass Correlation Coefficient (ICC), Cohen's kappa, Fleiss' kappa, Lin's Concordance Correlation Coefficient (CCC), and Gwet's AC1, which are widely employed in social science studies. The paper highlights major challenges, such as subjective biases, variability in training, and contextual influences that affect evaluator agreement. Furthermore, the implications of unreliable ratings on research validity, policy-making, and practical applications are discussed. Addressing these challenges through standardization, structured training protocols, and advanced computational techniques is essential for enhancing the reliability and credibility of assessments in social sciences. This paper provides a comprehensive discussion of the aforementioned measures, accompanied by real-life examples, ensuring that social science researchers can effectively apply these methods in their own studies.

Keywords: Inter-rater reliability; Cohen's kappa; Intra-class correlation; Fleiss' kappa; Lin's concordance correlation coefficient.

1. Introduction

Agreement plays a crucial role in ensuring consistency across various domains. Whether it is a chef disagreeing on seasoning, doctors interpreting the same X-ray differently, or

teachers assigning varying grades to the same essay, discrepancies in judgment highlight the fundamental need for statistical agreement. Unlike correlation, which measures how two variables move together, agreement assesses whether their values truly match (cf. Altman and Bland (1983), Bland and Altman (1986)). High correlation does not necessarily imply agreement, as systematic differences can persist despite strong associations. For instance, if two blood pressure monitors consistently differ by 10 mmHg, a high correlation may still mask clinically significant discrepancies. In subjective evaluations, such as employee assessments or film reviews, variations among raters may be expected. However, in standardized settings like medical diagnoses or psychological testing, inconsistent ratings indicate a need for improved frameworks (cf. McGraw and Wong (1996)). Even measurement tools can be unreliable—a weighing scale that fluctuates for the same person raises concerns about accuracy (cf. Lin (1989)).

To assess agreement, various statistical methods have been developed. For nominal data, Cohen's Kappa (cf. Cohen (1960)) is widely used, with weighted extensions (cf. Landis and Koch (1977)) improving its applicability to ordinal categories. However, Cohen's Kappa has been critiqued for standardization issues, leading to extensions that accommodate multiple raters and mixed data types (cf. Eye et al. (2005), Shoukri (2003)).

For continuous data, Lin's CCC (cf. Lin (1989)) provides a robust measure of agreement, further refined for repeated measurements (cf. Chinchilli et al. (1996)) and mixed data types (cf. King and Chinchilli (2001)). Additional methods, such as the Total Deviation Index (cf. Lin (2000)) and coverage probability analysis (cf. Lin et al. (2002)), extend agreement assessment in laboratory and clinical settings. Multirater agreement on continuous scales requires specialized methods. Barnhart et al. (2007) developed ANOVA-based and modeling approaches for multiple raters. Refinements to agreement testing continue to emerge. Dutta and Sinha (2013) proposed a likelihood ratio test for improved hypothesis testing, and Pornpis et al. (2006) critically evaluated Cohen's Kappa, suggesting modifications to enhance its applicability.

This paper explores key agreement measures, including the Bland-Altman Analysis (Section 2), ICC (Section 3), Cohen's Kappa (Section 4), Fleiss' Kappa (Section 5), Lin's CCC (Section 6), and Gwet's AC1 (Section 7). The corresponding R code and output for the measures used in the example are provided in the Appendix. While traditional agreement measures remain widely used, recent advancements address their limitations. Future research should focus on refining agreement methodologies to better accommodate complex and high-dimensional data.

2. Bland-Altman Analysis

The Bland-Altman method (cf. Bland and Altman (1986)) is a widely used statistical approach for assessing agreement between two measurement techniques. It provides an intuitive graphical representation of measurement differences, helping to detect systematic bias and variability. By plotting differences against the means, it identifies proportional bias and measurement errors, making it particularly valuable in medical, biological, and industrial research.

A key advantage of this method is its clear visualization of agreement, even with small sample sizes. It effectively highlights systematic errors and proportional bias that correlation analysis might overlook. However, it assumes normally distributed differences, which may not always hold, and does not directly explain the sources of disagreement. Additionally, small sample sizes can lead to unreliable limits of agreement (LoA), requiring careful interpretation.

The Bland-Altman plot visualizes agreement by plotting the difference between two measurements $d = X_1 - X_2$ against their mean $\bar{X} = \frac{X_1 + X_2}{2}$. The mean difference (*bias*) is given by:

$$\bar{d} = \sum_{i=1}^n (X_{1i} - X_{2i})$$

with the 95% LoA calculated as $\bar{d} \pm 1.96 \times SD_d$, where SD_d represents the standard deviation of the measurement differences. A narrow LoA suggests strong agreement, while a wider LoA indicates greater variability. Outliers beyond these limits signal potential inconsistencies or systematic errors (cf. Giavarina (2015)). This method is widely applied in comparing medical instruments (cf. Bland and Altman (1999)), ensuring measurement reliability in environmental studies (cf. Giavarina (2015)), and validating sensors in industrial settings. Its simplicity and effectiveness make it a fundamental tool for assessing measurement agreement.

Example 2.1 Suppose we have systolic blood pressure (SBP) readings from two methods: a traditional mercury sphygmomanometer (X_1) and an automated digital monitor (X_2). The dataset presented in Table 1 is derived from the study conducted by Dewitte et al. (2002) and is used here for illustrative purposes. It does not represent the complete dataset from the study; rather, a subset has been selected to demonstrate the method effectively.

Table 1. Systolic Blood Pressure Readings from Two Methods

Patient	X_1 (Sphygmomanometer)	X_2 (Digital Monitor)
1	120	118
2	135	138
3	142	145
4	126	127
5	130	129
6	140	143
7	128	125
8	136	137
9	138	135
10	145	147

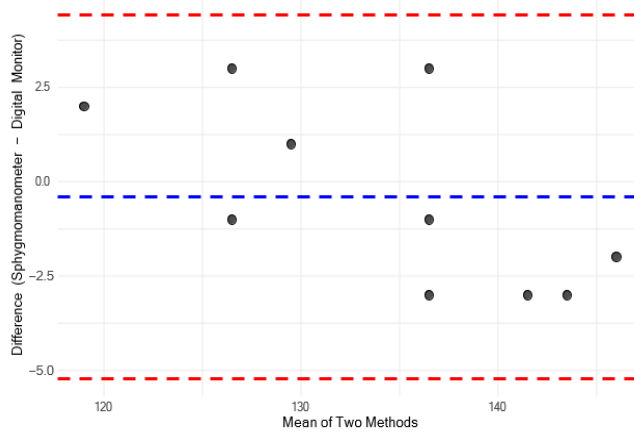


Figure 1. Bland-Altman Plot for Blood Pressure Measurement Methods

We analyze this data using the Bland-Altman plot in R, employing the R code provided in Appendix A.1. Figure 1 illustrates the agreement between X_1 and X_2 . The mean difference (bias) is -0.4 , with LoA $[-5.2, 4.4]$, indicating that 95% of differences fall within this range. The data points lie within these limits, showing strong consistency and no systematic bias or proportional trend. Thus, the digital monitor provides measurements comparable to the sphygmomanometer, with minor variations, making it suitable for clinical use.

3. Intraclass Correlation Coefficient (ICC)

The ICC quantifies the reliability and consistency of ratings when multiple raters evaluate the same subjects (cf. Shrout and Fleiss (1979)). It is computed using variance components from an ANOVA model:

$$ICC = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2} \quad (1)$$

where σ_b^2 and σ_w^2 represent between-subject and within-subject variance, respectively. Different ICC models exist based on rater types and measurement conditions.

ICC values range from 0 to 1, interpreted as follows: < 0.50 (poor reliability), $0.50 \leq ICC < 0.75$ (moderate reliability), $0.75 \leq ICC < 0.90$ (good reliability), and ≥ 0.90 (excellent reliability) (cf. Koo et al. (2016)). Higher ICC values indicate strong agreement, while lower values suggest inconsistency or measurement errors.

ICC is widely applied in medical diagnostics, psychological assessments, education, and sports science. Its advantages include robust reliability assessment, flexibility, and various model choices. However, it assumes normality, depends on sample size, and is sensitive to outliers. Proper application ensures accurate and reproducible assessments.

Example 3.1 Consider a study that utilizes a small dataset adapted from Shrout and Fleiss (1979) to compute the ICC, with additional datasets available on platforms such as Figshare. This analysis provides valuable insights into the level of agreement between raters. The dataset consists of ratings assigned by three independent raters to six subjects, as presented in Table 2.

Table 2. Ratings by Three Raters (Source: Shrout and Fleiss (1979))

Subject	Rater 1	Rater 2	Rater 3
1	5.1	4.9	5.2
2	4.8	4.7	4.9
3	6.2	6.0	6.1
4	5.9	5.8	5.7
5	4.5	4.4	4.6
6	5.3	5.2	5.4

The computed ICC value of 0.982 indicates an exceptionally high level of agreement among raters, reflecting near-perfect consistency in their evaluations. The narrow confidence interval (0.928, 0.997) further underscores the reliability of this estimate, with minimal variability. Additionally, the extremely small p -value ($p = 2.68 \times 10^{-9}$)

confirms that the observed agreement is highly significant and unlikely to be due to chance. Such a high ICC demonstrates strong reliability, making these ratings suitable for practical applications. However, sustaining consistency over time necessitates periodic reliability assessments and rater calibration. This study, grounded in the methodology of Shrout and Fleiss (1979), provides a robust framework for assessing inter-rater reliability. The R code and output for ICC computation are presented in Appendix A.2.

4. Cohen's Kappa

Cohen's Kappa (κ) measures inter-rater agreement while accounting for chance agreement (cf. Cohen (1960)). It is widely used in categorical data classification, offering a more accurate assessment than simple percentage agreement. Mathematically, it is defined as:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (2)$$

where P_o is the observed agreement and P_e is the expected agreement by chance.

The value of κ ranges from -1 (systematic disagreement) to $+1$ (perfect agreement), with thresholds indicating slight ($\kappa \leq 0.20$), fair ($0.21 \leq \kappa \leq 0.40$), moderate ($0.41 \leq \kappa \leq 0.60$), substantial ($0.61 \leq \kappa \leq 0.80$), and almost perfect ($\kappa > 0.80$) agreement.

Cohen's Kappa is widely applied in medical diagnostics, survey research, and psychological testing, ensuring reliable classifications. However, it is sensitive to imbalanced data and assumes independent raters. Despite its limitations, it remains a fundamental tool for evaluating classification consistency across disciplines.

Example 4.1 To illustrate, Table 3 presents data from a study that examined the prevalence of depression among 200 primary care patients using two different assessment methods. The first method is based on self-reported information provided by the individual (proband-based diagnosis), while the second relies on information given by a close acquaintance, such as a family member or friend (informant-based diagnosis).

A straightforward approach to measuring agreement between the two methods is to calculate the proportion of cases where both resulted in the same diagnosis. In this case, the observed agreement is 65.5% $\left(\frac{65+66}{200} \times 100\%\right)$ meaning that the two methods concurred in 131 out of 200 cases. However, this measure does not account for the likelihood of agreement occurring purely by chance.

Table 3. Depression Diagnosis: Agreement between Proband and Informant

Proband	Informant:Not Depressed	Informant: Depressed	Total
Not Depressed	66	19	85
Depressed	50	65	115
Total	116	84	200

A more robust measure of agreement is Cohen’s Kappa (κ), which accounts for the probability of agreement occurring by chance. For this dataset, Cohen’s Kappa is computed as $\kappa = 0.33$, indicating a *fair* level of agreement. This suggests that while there is some consistency between the two assessment methods, discrepancies remain significant. To enhance diagnostic reliability, it is recommended to implement standardized assessment guidelines and provide comprehensive training for both self-reporting and informant evaluations. The R code for computing Cohen’s kappa, along with its output, is provided in Appendix A.3.

5. Fleiss’ Kappa

Fleiss’ Kappa (κ) extends Cohen’s Kappa to measure agreement among multiple raters assigning categorical ratings (cf. Fleiss (1971)). Fleiss’s kappa is a generalization of Cohen’s kappa for more than two raters. It accounts for chance agreement, making it a robust inter-rater reliability metric. Mathematically, it is defined as:

$$\kappa = \frac{\overline{P_o} - \overline{P_e}}{1 - \overline{P_e}} \quad (3)$$

where $\overline{P_o}$ is the average observed agreement and $\overline{P_e}$ is the average expected agreement by chance over multiple raters. Its value ranges from -1 (systematic disagreement) to $+1$ (perfect agreement), with standard interpretation thresholds (cf. Landis and Koch (1977)).

Fleiss’ Kappa is widely applied in medical diagnosis, psychological assessments, educational research, and content analysis. It accommodates multiple raters and categorical classifications, making it more reliable than simple agreement percentages. However, it assumes all disagreements are equal, is sensitive to category prevalence, and may be unstable for small samples. Despite these limitations, it remains essential for evaluating classification consistency across disciplines.

Example 5.1 Diagnosing depression is often subjective, leading to variations in assessment among different clinicians. In this study, five psychiatrists independently evaluated 30 patients, classifying them into three categories: Mild, Moderate, or Severe. The ratings are presented in Table 4 (cf. Fleiss (1971)).

Table 4. Psychiatric Diagnoses by Five Raters (Source: Fleiss (1971))

Patient	Mild	Moderate	Severe	Patient	Mild	Moderate	Severe
1	3	2	0	16	3	2	0
2	1	4	0	17	0	3	2
3	2	2	1	18	2	2	1
4	0	5	0	19	1	3	1
5	1	3	1	20	0	4	1
6	2	3	0	21	1	3	1
7	3	2	0	22	2	3	0
8	2	1	2	23	3	1	1
9	0	4	1	24	4	0	1
10	1	2	2	25	1	4	0
11	3	1	1	26	0	5	0
12	4	0	1	27	2	3	0
13	0	5	0	28	1	2	2
14	2	3	0	29	3	2	0
15	1	2	2	30	2	3	0

Fleiss' Kappa was computed in R, yielding a value of $\kappa = -0.0724$, which suggests poor agreement among the psychiatrists. The negative value indicates that the observed agreement is lower than what would be expected by chance alone. This result highlights significant inconsistencies in diagnosis, likely due to subjective variability in assessment criteria. Given this low level of agreement, it is crucial to implement standardized diagnostic protocols and enhanced training to improve inter-rater reliability. Further research may explore structured diagnostic tools or consensus-driven methodologies to mitigate inconsistencies. The R script used for computation is provided in Appendix A.4, along with the output. The methodology follows the framework proposed by Landis and Koch (1977), with data sourced from Fleiss (1971).

6. Lin's Concordance Correlation Coefficient (CCC)

Lin's CCC (ρ_c) quantifies agreement between two continuous measurements by assessing both precision and accuracy, making it ideal for method comparison studies and reliability analysis (cf. Lin (1989)). Unlike Pearson's correlation, CCC accounts for both correlation and bias, providing a comprehensive measure of agreement. Mathematically, CCC is defined as:

$$\rho_c = \frac{2s_{xy}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2} \quad (4)$$

where s_{xy} is the covariance, s_x^2 and s_y^2 are variances, and $\bar{x}$, $\bar{y}$ are means of the two measurements. Alternatively:

$$\rho_c = r \times C_b \quad (5)$$

where r is Pearson's correlation, and C_b is a bias correction factor.

CCC ranges from -1 to 1 , with values above 0.90 indicating strong agreement and values below 0.50 suggesting poor agreement (cf. McBride (2005)). It is widely used in medical diagnostics, machine learning, environmental studies, and biological sciences.

While CCC effectively evaluates agreement, it is sensitive to outliers, assumes normality, and does not distinguish systematic bias from random error. Despite these limitations, it remains a robust tool for assessing measurement consistency across various fields.

Example 6.1 A study assessed the agreement between two methods of measuring eelgrass (*Zostera marina*) leaf area: direct measurement and digital image analysis. Lin's CCC was calculated to evaluate the consistency between these methods. The dataset consists of ten samples, as presented in Table 5.

Table 5. Comparison of Leaf Area Measurements (Source: Echavarría-Heras et al. (2014))

Leaf ID	Direct Measurement (cm ²)	Digital Image Analysis (cm ²)
1	35.5	36.0
2	42.3	41.8
3	28.7	29.1
4	33.9	34.2
5	40.1	39.8
6	37.4	37.9
7	45.2	44.7
8	31.8	32.0
9	29.5	29.9
10	38.6	38.2

The computed CCC value was 0.997 , indicating an almost perfect agreement between the two measurement techniques. Despite minor variations, digital image analysis proved to be a highly reliable alternative to direct measurement. Given its efficiency and

ease of use, researchers recommend adopting digital image analysis for large-scale studies to minimize manual errors and enhance data collection accuracy.

Key insights from the R output (Appendix A.5), with the corresponding R code also available in Appendix A.5, are as follows:

- $\rho_c = 0.9966$ confirms an extremely high level of agreement.
- Confidence Interval: (0.9918, 0.9986) ensures statistical robustness.
- Bias Correction Factor: $C_b = 0.9979$, indicating minimal systematic differences.
- Scale shift (0.9385) and location shift (0.0119) suggest negligible bias.

Had the CCC been significantly lower (e.g., < 0.90), additional calibration or methodological adjustments would have been necessary to enhance measurement consistency. Unlike Pearson’s correlation, which measures only association, CCC provides a more comprehensive assessment by capturing both association and agreement. With these results, the study reinforced the robustness of digital image analysis as a reliable tool for future research applications.

7. Gwet’s AC1

Gwet’s AC1 addresses the limitations of Cohen’s Kappa, particularly its sensitivity to category prevalence, ensuring a more stable measure of agreement (cf. Gwet (2008)). It is especially useful when data distributions are imbalanced. Mathematically, AC1 is defined as:

$$ACI = \frac{P_o - P_e}{1 - P_e} \quad (6)$$

where $P_o = \sum_i P_{ii}$ represents the observed agreement and $P_e = \frac{1}{k} \sum_i P_{i+} P_{+i}$ is the expected agreement under independence.

AC1 ranges from -1 to 1 , with values above 0.81 indicating almost perfect agreement and values below 0.00 suggesting systematic disagreement (cf. Landis and Koch (1977)). Unlike Kappa, AC1 remains robust even when category distributions are skewed. AC1 is widely used in medical diagnostics, psychology, content analysis, and machine learning, where reliable classification is crucial. Though more stable than Kappa, it still depends on category distributions and should be interpreted alongside other agreement measures for a comprehensive reliability assessment.

Example 7.1 The dataset `cac.raw4raters`, available in the `irrCAC` R package, contains categorical ratings assigned by four raters to 12 subjects. Each rater independently classified subjects into predefined categories based on their observed characteristics. This

dataset, originally used in reliability studies, helps evaluate the consistency of multiple raters in qualitative assessments. The dataset structure is presented in Table 6. This study evaluates inter-rater agreement using Gwet’s AC1, which addresses the limitations of Cohen’s kappa. The ratings were assigned in an experiment where multiple raters independently assessed subjective attributes of individuals, such as behavioral tendencies or medical diagnoses, to measure agreement levels.

Table 6. Ratings by Four Raters (Source: Gwet (2008))

Subject	Rate r 1	Rate r 2	Rate r 3	Rate r 4	Subject	Rate r 1	Rate r 2	Rate r 3	Rate r 4
1	A	A	A	B	7	A	B	A	B
2	B	B	A	B	8	A	A	A	B
3	A	A	A	A	9	B	B	A	B
4	B	A	B	B	10	A	A	A	A
5	A	A	B	A	11	B	B	B	A
6	B	B	B	B	12	A	B	A	A

The analysis yields a substantial agreement with $AC1 = 0.775$, supported by a statistically significant p -value ($p = 0.0002$) (for R code and output, see Appendix A.6). While ratings are generally consistent, minor variability exists. Enhancing reliability through standardized training, structured rating scales, and multiple independent raters is recommended for future studies.

8. Conclusion: Agreement as the Foundation of Reliability

Statistical agreement is fundamental to ensuring that measurements—whether conducted by humans or automated systems—are both reliable and reproducible. Various statistical methods, such as Bland-Altman analysis, ICC, Cohen’s kappa, Fleiss’ kappa, Lin’s CCC, and Gwet’s AC1, provide robust frameworks for quantifying agreement and enhancing measurement reliability.

In this study, we delve into these methods in detail, illustrating their application with real-life examples. Special emphasis is placed on the interpretation of these agreement measures and the recommendations derived from their computed values. These insights serve as valuable guidance for practitioners and researchers across diverse disciplines within the social sciences.

Across fields such as medical diagnostics, education, psychology, and manufacturing, agreement metrics play a crucial role in maintaining precision, minimizing errors, and ensuring fairness in decision-making. Table 7 provides a comprehensive summary of the key applications of various agreement measures. Selecting the appropriate agreement

measure ensures reliable assessments and better decision-making. For further exploration, Mukherjee, Sinha, and Chattopadhyay (2018) provide an in-depth discussion on statistical agreement methods and applications.

Acknowledgement

We express our sincere gratitude to Prof. Bikas Kumar Sinha, retired Professor of Statistics, Indian Statistical Institute, Kolkata, India, for his constant support, insightful comments, and invaluable guidance. His thoughtful suggestions and encouragement have greatly contributed to the development of this work.

References

- [1] Altman, D. G. and Bland, J. M. (1983). Measurement in medicine: The analysis of method comparison studies. *Journal of the Royal Statistical Society*, 32(3), 307–317.
- [2] Barnhart, H. X., Haber, M. J., and Lin, L. I. (2007). An overview on assessing agreement with continuous measurements. *Journal of Biopharmaceutical Statistics*, 17(4), 529–569.
- [3] Bland, J. M. and Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 1(8476), 307–310.
- [4] Bland, J. M. and Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, 8(2), 135–160.
- [5] Chinchilli, V. M. Martel J. K., Kumanyika, S, and Lloyd, T. (1996). A weighted concordance correlation coefficient for repeated measures designs. *Biometrics*, 52(1), 341–353.
- [6] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.
- [7] Dewitte, K., Fierens, C., Stöckl, D., and Thienpont, L. M. (2002). Application of the Bland-Altman plot for interpretation of method comparison studies: A critical investigation of its practice. *Clinical Chemistry*, 48(5), 799–801; author reply 801-802 .
- [8] Dutta, G. and Sinha, B. K. (2013). Some further aspects of assessment of agreement involving bivariate normal responses. *International Journal of Statistical Sciences*, 13, 1–19.
- [9] Echavarría-Heras, H., Leal-Ramírez, C., Villa-Diharce, E., and Castillo, O. (2014). Using the value of Lin’s concordance correlation coefficient as a criterion for efficient estimation of areas of leaves of eelgrass from noisy digital images, 9(1), 29.
- [10] Eye, A. von and Mun, E. Y. (2005). *Analyzing Rater Agreement: Manifest Variable Methods*. Psychology Press.

- [11] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382.
- [12] Giavarina, D. (2015). Understanding Bland Altman analysis. *Biochem Med (Zagreb)*, 25(2), 141–151.
- [13] 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, 29–48.
- [14] King, T. S. and Chinchilli, V. M. (2001). A generalized concordance correlation coefficient for continuous and categorical data. *Statistics in Medicine*, 20(14), 2131–2147.
- [15] Koo, T. K. and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163.
- [16] Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.
- [17] Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1), 255–268.
- [18] Lin, L. I. (2000). Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in Medicine*, 19(2), 255–270.
- [19] Lin, L., Hedayat, A. S., Sinha, B., K., and Yang, M. (2002). Statistical methods in assessing agreement: Models, issues, and tools. *Journal of the American Statistical Association*, 97(457), 257–270.
- [20] McBride, G. B. (2005). *A proposal for strength-of-agreement criteria for Lin's concordance correlation coefficient*. NIWA Client Report: HAM2005-062.
- [21] McGraw, K. O. and Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1), 30–46.
- [22] Mukherjee, S. P., Sinha, B. K., and Chattopadhyay, A. K. (2018). *Statistical Methods in Social Science Research*. Springer Nature.
- [23] Sinha, B. K., Yimprayoon, P., and Tiensuwan, M. (2006). Cohen's Kappa statistic: A critical appraisal and some modifications. *Calcutta Statistical Bulletin*, 58, 151–170.
- [23] Shoukri, M. M. (2003). *Measures of Interobserver Agreement and Reliability*. CRC Press.
- [24] Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86 (2), 420–428.

A Appendix

A.1 R code generates the Bland-Altman plot

```
# Install necessary package if not installed
# install.packages("ggplot2")

# Load required library
library(ggplot2)

# Create data (Fictitious dataset for illustration)

# Calculate mean difference (bias)
bias <- mean(data$Diff) SD_diff <- sd(data$Diff)
LoA_upper <- bias + 1.96 * SD_diff LoA_lower <- bias - 1.96 * SD_diff

# Save plot to PDF
pdf("Bland_Altman_Plot.pdf", width = 7, height = 5)
# Generate Bland-Altman plot
ggplot(data, aes(x = Mean, y = Diff)) +

  geom_point(size = 3, alpha = 0.7) +
  geom_hline(yintercept = bias, linetype = "dashed", color = "blue", linewidth = 1) +
  geom_hline(yintercept = LoA_upper, linetype = "dashed", color = "red", linewidth = 1) +
  geom_hline(yintercept = LoA_lower, linetype = "dashed", color = "red", linewidth = 1) +
  labs(title = "Bland-Altman Plot: Blood Pressure Measurements",
       x = "Mean of Two Methods",
       y = "Difference (Sphygmomanometer - Digital Monitor)") +
  theme_minimal()

# Close the PDF device dev
.off()
```

Listing 1: R Code for Bland-Altman plot

A.2 R Code for ICC Calculation

The ICC is computed using the R programming language, as shown below:

```
# Load necessary package
install.packages("irr") # Install if not already
                             installed
library(irr)

# Define the dataset
ratings <- matrix(c(
  5.1, 4.9, 5.2,
  4.8, 4.7, 4.9,
  6.2, 6.0, 6.1,
  5.9, 5.8, 5.7,
  4.5, 4.4, 4.6,
  5.3, 5.2, 5.4
), nrow=6, byrow=TRUE)
# Compute ICC
icc_result <- icc(ratings, model="twoway", type="
  consistency", unit="single")
print(icc_result)
```

Listing 2: ICC Computation in R

The corresponding R output is:

```
Single Score Intraclass
Correlation Model: Two-Way
Type : Consistency
Subjects = 6
Raters = 3
ICC(C,1) = 0.982
F-Test, H0: r0 = 0 ; H1:
r0 > 0 F(5,10) = 167 , p
= 2.68e-09
95%-Confidence Interval for ICC Population Values: 0.928
< ICC < 0.997
```

A.3 Calculating Cohen's Kappa in R

The Cohen's Kappa statistic can be calculated in R using the psych package as shown in the code below:

```
# Load required package
install.packages("psych") # Install if not
                             already installed
library(psych)

# Define the ratings matrix (Fictitious Data)
ratings <- matrix(c(66, 19, 50, 65), nrow = 2, byrow =
  TRUE)
# Compute Cohens Kappa
kappa_result <- cohen.kappa(
  ratings) # Print the result
print(kappa_result)
```

Listing 3: R Code for Computing Cohen's Kappa

R Output: Cohen's Kappa Calculation

Executing the R script yields the following output:

The results indicate that the unweighted and weighted kappa values are identical, with an estimated agreement of $\kappa = 0.33$ and a confidence interval ranging from 0.20 to 0.45.

A4. Computation of Fleiss' Kappa in R

The following R code computes Fleiss' Kappa for the dataset

```
# Load necessary package
install.packages("irr") # Install if not already
                             installed
library(irr)

# Define the ratings matrix: Each row corresponds to a patient,
  and each column represents the count of ratings for each
  category (Mild, Moderate, Severe)
ratings <- matrix(c(
```

```

3, 2, 0, # Patient 1
1, 4, 0, # Patient 2
2, 2, 1, # Patient 3
0, 5, 0, # Patient 4
1, 3, 1, # Patient 5
2, 3, 0, # Patient 6
3, 2, 0, # Patient 7
2, 1, 2, # Patient 8
0, 4, 1, # Patient 9
1, 2, 2, # Patient 10
3, 1, 1, # Patient 11
4, 0, 1, # Patient 12
0, 5, 0, # Patient 13
2, 3, 0, # Patient 14
1, 2, 2, # Patient 15
3, 2, 0, # Patient 16
0, 3, 2, # Patient 17
2, 2, 1, # Patient 18
1, 3, 1, # Patient 19
0, 4, 1, # Patient 20
1, 3, 1, # Patient 21
2, 3, 0, # Patient 22
3, 1, 1, # Patient 23
4, 0, 1, # Patient 24
1, 4, 0, # Patient 25
0, 5, 0, # Patient 26
2, 3, 0, # Patient 27
1, 2, 2, # Patient 28
3, 2, 0, # Patient 29
2, 3, 0 # Patient 30
nrow=30, byrow = TRUE)

```

```
# Compute Fleiss' Kappa
```

```
kappa_result <- kappam.fleiss(ratings) print(kappa_result)
kappa_result <- kappam.fleiss(ratings) print(kappa_result)
```

Listing 4: Fleiss' Kappa Computation in R

Executing the R script yields the following output:

Fleiss' Kappa for m Raters

```
Subjects = 30
Raters   = 3
Kappa    = -0.0724
z         = -1.35
p-value  = 0.176
```

A5. CCC in R

To assess the agreement between two measurement methods, we compute Lin's CCC using the epiR package in R. The following script performs the calculation:

```
# Load necessary package
install.packages("epiR") # Install if not already
                             installed
library(epiR)

# Define the dataset
direct_measurement <- c(35.5, 42.3, 28.7, 33.9, 40.1,
                        37.4, 45.2, 31.8, 29.5, 38.6)
digital_analysis  <- c(36.0, 41.8, 29.1, 34.2, 39.8, 37.9,
                        44.7, 32.0, 29.9, 38.2)

# Compute Lin's Concordance Correlation Coefficient (CCC)
ccc_result <- epi.ccc(direct_measurement, digital_analysis
                      , ci = "z-transform", conf.level = 0.95)

# Print the results
print(ccc_result)
```

Listing 5: CCC Computation in R

The corresponding R output is presented below:

```
$rho.c
      est      lower      upper
1 0.9966123 0.9918327 0.9985968
```

A6. Computation of Gwet's AC1: R Code and Output

To measure inter-rater reliability, Gwet's AC1 statistic is computed using the irr package in R. The following script performs the calculation:

```
# Install and load necessary package
install.packages("irr") # Install if not already
                             installed
library(irr)

# Define the ratings matrix
data <- matrix(c(
  "A", "A", "A", "B",
  "B", "B", "A", "B",
  "A", "A", "A", "A",
  "B", "A", "B", "B",
  "A", "A", "B", "A",
  "B", "B", "B", "B",
  "A", "B", "A", "B",
  "A", "A", "A", "B",
  "B", "B", "A", "B",
  "A", "A", "A", "A",
  "B", "B", "B", "A",
  "A", "B", "A", "A"
), nrow=12, byrow=TRUE)

# Compute Gwet's AC1
ac1_result <- gwet.ac1(data)
print(ac1_result)
```

Listing 6: Gwet's AC1 Computation in R

The computed Gwet's AC1 value from the R analysis is as follows:

Table 7: Comparison of Agreement Measures

Measure	Application	Field	Data Type
Bland–Altman	Agreement between two continuous measurements	Medical research, clinical trials	Continuous
ICC	Reliability among multiple raters	Psychology, medical imaging	Continuous
Cohen's Kappa	Agreement between two raters	Social sciences, medical diagnosis	Categorical
Fleiss' Kappa	Agreement among multiple raters	Surveys, multi-rater assessments	Categorical
Lin's CCC	Agreement between two continuous variables (adjusted for bias)	Epidemiology, genetics	Continuous
Gwet's AC1	Agreement in imbalanced categorical data	Clinical research, content analysis	Categorical

Table 8: Cohen's Kappa and Weighted Kappa Estimates

Metric	Lower	Estimate	Upper
Unweighted Kappa	0.20	0.33	0.45
Weighted Kappa	0.20	0.33	0.45
Number of subjects:	200		

Table 9: Gwet's AC1 Reliability Statistics

Statistic	Value
Agreement Coefficient (AC1)	0.775
Observed Agreement (P_o)	0.818
Expected Agreement (P_e)	0.190
Standard Error	0.1429
Confidence Interval	(0.461, 1.000)
p -value	0.0002