Short Communication

An Exploratory, Data-informed Approach to Characterizing Weighting Choices for Cohen's Kappa in Ordinal Data

Bizhan Shabankhani[1], Mahsa Bakhshandeh[2], Mohammadmehdi Rezvani-Ahangarkolaei[3], Zeinab Babaei[4], Keihan Shabankhani[5*]

1. Department of Biostatistics, School of Health, Mazandaran University of Medical Sciences, Sari, Iran.
2. Department of Gynecology and Obstetrics, School of Medicine, Mazandaran University of Medical Sciences, Sari, Iran.
3. Health Sciences Research Center, Addiction Institute, Mazandaran University of Medical Sciences, Sari, Iran.
4. Department of Anesthesiology, School of Medicine, Babol University of Medical Sciences, Babol, Iran.
5. Department of Anesthesiology, School of Medicine, Mazandaran University of Medical Sciences, Sari, Iran.

Use your device to scan and read the article online

Citation Shabankhani B, Bakhshandeh M, Rezvani-Ahangarkolaei M, Babaei Z, Shabankhani K. An Exploratory, Data-informed Approach to Characterizing Weighting Choices for Cohen's Kappa in Ordinal Data. Iranian Journal of Health Sciences. 2026; 14(3):321-326. http://dx.doi.org/10.22034/ijhs.14.3.1207.2

http://dx.doi.org/10.22034/ijhs.14.3.1207.2

Article info:
Received: 28 Dec 2025
Revised: 17 Mar 2026
Accepted: 12 Apr 2026

ABSTRACT

Background and Purpose: Selecting an appropriate weighting scheme for Cohen's kappa in ordinal data depends on the structure and practical meaning of disagreements between categories. Although linear and quadratic weighting schemes are widely used, their selection is often guided by substantive considerations rather than a standardized statistical rule. This methodological short communication described an exploratory data-informed approach for characterizing ordinal association and category-level heterogeneity using Spearman's rank correlation and one-vs-rest receiver operating characteristic (ROC) analysis. The approach is intended to provide supplementary information for interpreting disagreement patterns rather than to establish a validated rule for selecting a specific kappa weighting scheme.

Materials and Methods: We described a two-component exploratory approach. First, Spearman's rank correlation coefficient was used as a descriptive measure of ordinal association between two raters and was interpreted separately from agreement statistics. Second, one-vs-rest ROC analysis was used to describe variation in category-level discrimination across ordinal categories. The difference between the maximum and minimum area under the curve (AUC) values was summarized using a descriptive heterogeneity index ($D=AUC_{max} - AUC_{min}$). The approach was illustrated using a simulated ordinal dataset. The index was used only to characterize the observed heterogeneity and was not assigned a validated cutoff or used as a definitive decision rule for selecting linear or quadratic weighting.

Results: In the illustrative example, the two raters demonstrated a moderately strong positive ordinal association (Spearman's P=0.737). One-vs-rest ROC analyses showed variation in AUC values across categories and between raters. The highest AUC was 0.833 and the lowest was 0.528, resulting in a descriptive heterogeneity index of D=0.305. This variation illustrates non-uniform category-level discrimination across the ordinal scale. The observed pattern may provide supplementary information when considering the characteristics of disagreement and the interpretation of alternative weighting schemes, but it does not establish that quadratic weighting is universally preferable.

*** Corresponding Author:**
Keihan Shabankhani, MD.
Address: *Department of Anesthesiology, School of Medicine, Mazandaran University of Medical Sciences, Sari, Iran.*
E-mail: *kshabankhani@gmail.com*

Copyright © 2026 The Author(s);
This is an open access article distributed under the terms of the Creative Commons Attribution License (CC-By-NC 4.0: https://creativecommons.org/licenses/by-nc/4.0/legalcode.en), which permits use, distribution, and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

Conclusion: This methodological short communication presented an exploratory approach for describing ordinal association and category-level heterogeneity when considering weighting choices for Cohen's kappa. Spearman's correlation and ROC/AUC analysis provide complementary descriptive information but do not directly measure inter-rater agreement or constitute a validated rule for selecting linear or quadratic weighting. The proposed D index is intended only as a descriptive summary of AUC heterogeneity. Larger simulation studies and independent datasets are required to determine whether these measures provide useful information for weighting decisions in different ordinal settings.

Keywords: Cohen's Kappa, Weighted Kappa, Spearman's rank correlation, ROC curve, Area under the curve (AUC), Ordinal data, Inter-rater agreement

Introduction

Inter-rater reliability (IRR) is an important component of medical and health research because disagreement between raters can affect the interpretation of categorical and ordinal outcomes. Cohen's kappa is widely used to quantify agreement between two raters while accounting for agreement expected by chance. For ordinal variables, however, disagreements may differ in practical or clinical importance according to the distance between categories [1, 2]. Consequently, weighted versions of Cohen's kappa are commonly used when the ordinal structure of the outcome makes the magnitude of disagreement relevant [1, 3].

Two commonly used weighting schemes are linear and quadratic weighting. Linear weighting assigns penalties that increase proportionally with the distance between categories, whereas quadratic weighting assigns progressively greater penalties to more distant disagreements. The choice between these schemes is therefore related to how disagreement severity is understood within the specific measurement scale. Importantly, there is no universally applicable statistical rule for determining which weighting scheme is appropriate for every ordinal outcome; substantive knowledge of the scale remains important. Measures of association may provide additional information about the relationship between ordinal ratings. Spearman's rank correlation coefficient describes the degree to which the ordinal ordering of observations is similarly preserved between raters. However, correlation should not be interpreted as a measure of agreement, because strong association can coexist with systematic disagreement. Similarly, receiver operating characteristic (ROC) analysis primarily measures discrimination and does not directly quantify inter-rater agreement [4, 5].

Despite these limitations, category-specific ROC analysis may provide supplementary descriptive information when an appropriate external reference criterion is available. In particular, variation in one-vs-rest AUC values across ordinal categories may help characterize whether category-level discrimination is relatively homogeneous or heterogeneous. Such information may complement, but cannot replace, direct assessment of inter-rater agreement and consideration of the substantive meaning of the ordinal scale.

The present methodological note therefore introduced an exploratory data-informed approach that combines Spearman's rank correlation and category-specific ROC analysis to describe ordinal association and heterogeneity across categories. The proposed approach is intended as an illustrative and descriptive framework rather than a validated decision rule for selecting between linear and quadratic weighting. A simulated dataset was used solely to demonstrate how the proposed measures can be calculated and interpreted.

Materials and Methods

1. Cohen's kappa: Cohen's kappa measures agreement between two raters beyond the agreement expected by chance. In its unweighted form, all disagreements between categories receive the same treatment. Although unweighted kappa can be calculated for ordinal variables, it does not account for the magnitude or practical importance of disagreements between ordered categories [1, 2, 6].

2. Weighted kappa (Kw): Weighted kappa extends Cohen's kappa by assigning partial penalties to disagreements according to the distance between categories. This makes it more suitable for ordinal outcomes [7-9].

a) Linear weighted Kappa: Linear weighting assigns penalties that increase proportionally with the distance between categories. Under this scheme, moving

one category farther from exact agreement results in a proportional increase in the disagreement penalty. Linear weighting may be appropriate when the practical importance of disagreement is assumed to increase approximately proportionally across the ordinal scale [8-11].

b) Quadratic Weighted assigns increasingly greater penalties as the distance between categories increases. Consequently, disagreements between distant categories contribute substantially more to the overall disagreement than under linear weighting. Quadratic weighting may be informative when large category discrepancies are considered substantially more consequential than adjacent-category disagreements. However, the choice of quadratic weighting should be justified according to the substantive characteristics of the ordinal scale rather than determined solely by an empirical statistic [7, 8, 9, 12].

Exploratory use of Spearman's rank correlation

Spearman's rank correlation coefficient was used as a descriptive measure of ordinal association between the two raters. It quantifies the extent to which the relative ordering of observations is similarly preserved across raters. A higher coefficient indicates stronger monotonic association between the ratings.

Spearman's correlation does not directly measure inter-rater agreement and, therefore, should not be used as a stand-alone criterion for deciding whether unweighted or weighted kappa is appropriate. In particular, a high correlation may occur in the presence of systematic differences between raters. In the present framework, Spearman's coefficient was reported only to characterize the ordinal association between the ratings and to provide complementary information alongside direct agreement statistics.

Exploratory ROC analysis

To further characterize category-level patterns, one-vs-rest ROC analysis was performed separately for each ordinal category when an external reference classification was available. For each analysis, the target category was treated as the positive class and the remaining categories as the negative class. The resulting area under the curve (AUC) values describe the discriminatory separation associated with each category.

ROC/AUC analysis was not interpreted as a measure of inter-rater agreement and was not used as a validated criterion for determining the appropriate kappa weighting scheme. Rather, the analysis was used descriptively to examine whether category-level discrimination was relatively homogeneous or heterogeneous across the ordinal scale.

The usefulness of ROC analysis in this context depends on the presence of a clinically or substantively meaningful external reference criterion. Therefore, the ROC results should be interpreted as supplementary information rather than as a replacement for direct assessment of agreement or substantive consideration of the ordinal measurement scale.

Descriptive heterogeneity index

For exploratory purposes, we defined a simple heterogeneity index:

$$D=AUC_{max} - AUC_{min}$$

Where, (AUC_{max}) a nd (AUC_{min}) represent the largest and smallest AUC values obtained across the one-vs-rest ROC analyses.

The D index summarizes the spread of category-specific AUC values. A larger value indicates greater variation in category-level discrimination, whereas a smaller value indicates more similar AUC values across the evaluated categories. In this study, D is used solely as a descriptive measure of heterogeneity. No cutoff value was proposed, and D was not used as a validated decision rule for selecting linear or quadratic weighting.

Illustrative example:

The proposed exploratory approach was demonstrated using a simulated dataset consisting of 36 observations rated independently by two raters on a three-level ordinal satisfaction scale: dissatisfied, neutral, and satisfied.

The simulated example was constructed solely to demonstrate the calculation and interpretation of the proposed descriptive measures. It was not intended to evaluate the validity, accuracy, sensitivity, specificity, or generalizability of the proposed approach.

Spearman's rank correlation was calculated to describe the ordinal association between the two raters. One-vs-rest ROC analyses were then performed for

Table 1. Category-specific area AUC values from the one-vs-rest ROC analyses

Groups (One-vs-Rest)	Rater 1	Rater 2
Dissatisfied vs others	0.833	0.528
Neutral vs others	0.75	0.639
Satisfied vs others	0.528	0.667

each satisfaction category using the specified external reference classification. The resulting AUC values were summarized using the D index.

Results

To explore category-specific patterns, one-vs-rest ROC analyses were performed for each category. The AUC values were as follows:

Two raters independently evaluated the satisfaction level of 36 patients using a three-level ordinal scale. Spearman's rank correlation coefficient between the two raters was calculated and found to be 0.737, indicating a strong positive monotonic association. Based on this result, the use of weighted Cohen's kappa was considered appropriate.

To explore category-specific patterns, one-vs-rest ROC analyses were performed for each category. The AUC values were as follows:

The AUC values varied across categories and between raters, indicating that category-level discrimination was not uniform across the ordinal scale. The largest AUC was 0.833 and the smallest was 0.528, yielding a D index of:

D=0.833 - 0.528=0.305

This value indicates a relatively heterogeneous distribution of category-specific AUC values within this illustrative dataset. The observed heterogeneity may be relevant when considering the practical interpretation of disagreements across the ordinal scale; however, it does not, by itself, establish that linear or quadratic weighting is superior. The example is intended only to demonstrate the calculation and descriptive interpretation of the proposed index.

The resulting AUC values are presented in Table 1.

Discussion

This methodological note presented an exploratory approach for describing ordinal association and category-level heterogeneity when considering weighting schemes for Cohen's kappa. The central premise was that the interpretation of disagreement in ordinal data should consider both the statistical structure of the ratings and the practical meaning of distances between categories. The proposed approach combined Spearman's rank correlation as a descriptive measure of ordinal association with category-specific ROC/AUC analysis as supplementary information about discrimination across the ordinal scale.

Importantly, neither Spearman's correlation nor ROC/AUC is a direct measure of inter-rater agreement. Spearman's correlation describes monotonic association between ratings, whereas ROC/AUC describes discrimination relative to an external reference criterion. Accordingly, these measures should not replace direct agreement statistics or substantive consideration of the ordinal scale. Their proposed role in this study is descriptive: they may help characterize patterns that can be considered alongside the observed agreement matrix and the clinical or practical meaning of the distances between categories.

Interpretation of the D index

The D index provides a simple summary of the spread of category-specific AUC values. In the illustrative example, D=0.305 indicated that the degree of category-level discrimination varied across the evaluated categories. This observation may be useful when describing the structure of an ordinal scale, particularly when disagreements across categories may have different practical implications.

However, the present study does not establish a direct relationship between the magnitude of D and the superiority of either linear or quadratic weighting. Therefore, no threshold was proposed. The D index should be regarded as an exploratory descriptive measure whose potential usefulness for weighting decisions remains to be evaluated.

Limitations

Several limitations should be emphasized. First, the proposed approach was demonstrated using a single simulated dataset and was not designed to validate the performance of the proposed measures. Second, the relationship between AUC heterogeneity and the optimal weighting scheme for Cohen's kappa has not been established empirically. Third, ROC analysis requires an appropriate external reference criterion, and its interpretation may therefore depend on the clinical or substantive context in which it is applied. Fourth, Spearman's correlation measures ordinal association rather than agreement and should not be interpreted as a substitute for kappa or other agreement coefficients.

Accordingly, the proposed approach should be considered hypothesis-generating and exploratory. Future research should evaluate its behavior using extensive simulation studies with different numbers of categories, distributions of ratings, patterns of disagreement, prevalence structures, and predefined weighting assumptions. Independent empirical datasets should also be used to determine whether category-level heterogeneity measures provide information that is useful beyond conventional agreement statistics and substantive judgment.

Conclusion

Choosing between linear and quadratic weighting for Cohen's kappa depends on the structure of disagreements and the practical meaning of distances between ordinal categories. This methodological note presented an exploratory approach for describing ordinal association and category-level heterogeneity using Spearman's rank correlation and one-vs-rest ROC/AUC analysis. The proposed D index provides a simple descriptive summary of variation in category-specific AUC values. However, neither Spearman's correlation nor ROC/AUC directly measures inter-rater agreement, and no validated threshold or decision rule for selecting a weighting scheme was proposed. The approach should therefore be regarded as supplementary and exploratory. Larger simulation studies and independent datasets are required before its potential role in informing weighting choices can be established.

Ethical Considerations

Compliance with ethical guidelines

There were no ethical considerations to be considered in this research.

Funding

This research did not receive any grant from funding agencies in the public, commercial, or non-profit sectors.

Authors contributions

Methodology: Bizhan Shabankhani, Mahsa Bakhshandeh, Mohammadmehdi Rezvani-Ahangarkolaei, and Zeinab Babaei; Conceptualization: Bizhan Shabankhani; Scientific review and interpretation: Mahsa Bakhshandeh, Mohammadmehdi Rezvani-Ahangarkolaei, and Zeinab Babaei; Study design, data analysis, manuscript preparation, and supervision of the research process: Keihan Shabankhani; Final approval: All authors.

Conflict of interest

The authors declared no conflict of interest.

Acknowledgements

The authors would like to thank all individuals who contributed to the scientific discussion and development of this methodological study. No specific institutional or external assistance was received for the preparation of this manuscript.

References

[1] Gwet KL. Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Piedmont: Advanced Analytics, LLC; 2014. [Link]

[2] McHugh ML. Interrater reliability: The kappa statistic. Biochemia medica. 2012; 22(3):276-82. [DOI:10.11613/BM.2012.031] [PMID]

[3] Feng GC. Factors affecting intercoder reliability: A monte carlo experiment. Quality & Quantity. 2013; 47(5):2959-82. [DOI:10.1007/s11135-012-9745-9]

[4] Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: An open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011; 12(1):77. [DOI:10.1186/1471-2105-12-77] [PMID]

[5] Hallgren KA. Computing inter-rater reliability for observational data: An overview and tutorial. Tutorials in Quantitative Methods for Psychology. 2012; 8(1):23-34. [DOI:10.20982/tqmp.08.1.p023] [PMID] [PMCID]

[6] Shabankhani B. Assessing the inter-rater reliability for nominal, categorical and ordinal data in medical sciences. Archives of Pharmacy Practice. 2020; 11(4-2020):144-8. [Link]

[7] Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions. Hoboken: John Wiley & Sons; 2013. [Link]

[8] Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: Measures of agreement. Perspectives in Clinical Research. 2017; 8(4):187-91. [DOI:10.4103/picr.PICR_123_17] [PMID]

[9] Vanbelle S. A new interpretation of the weighted kappa coefficients. Psychometrika. 2016; 81(2):399-410. [DOI:10.1007/s11336-014-9439-4] [PMID]

[10] Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples. BMC Medical Research Methodology. 2013; 13(1):61. [DOI:10.1186/1471-2288-13-61] [PMID] [PMCID]

[11] Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. Journal of Clinical Epidemiology. 1990; 43(6):551-8. [DOI:10.1016/0895-4356(90)90159-M] [PMID] [PMCID]

[12] Li M, Gao Q, Yu T. Kappa statistic considerations in evaluating inter-rater reliability between two raters: Which, when and context matters. BMC Cancer. 2023; 23(1):799. [DOI:10.1186/s12885-023-11325-z]