Comparative Evaluation of Statistical Methods for Detecting Rater Bias in Ordinal Categorical Data

.

ABSTRACT

|  |
| --- |
| **Aims:** This study seeks to examine four statistical methods Modified McNemar Test, Single Binomial Test, Marginal Homogeneity Test, and Bias Index for identifying bias between two raters utilizing ordinal categorical data.  **Study design:**  This study employs a simulation-based comparative approach, using hypothetical 3×3 contingency tables to illustrate low, moderate, and high levels of agreement.  **Place and Duration of Study:**  Executed by a teaching member of a private university in the Philippines, from January 2024 to April 2025.  **Methodology:** Three simulated contingency tables representing varying agreement levels were analyzed using the four statistical methods. Test statistics and p-values were computed for each, and the Bias Index was calculated to assess directional bias. The outcomes were evaluated comparatively.  **Results:** Under low agreement conditions, all methods detected substantial bias, as reflected in high test statistics and low p-values. For moderate agreement, only the Marginal Homogeneity Test and Bias Index showed strong sensitivity. In the high agreement scenario, all approaches yielded non-significant results, and the Bias Index confirmed the absence of directional bias. Across all simulations, the Bias Index consistently produced interpretable and stable results.  **Conclusion:** While each method offers useful insights, the Bias Index emerged as the most robust and interpretable tool for assessing rater bias in ordinal data. A combined use of complementary statistical tests is recommended to determine both the direction and magnitude of bias in rater agreement studies. |

*Keywords:*

1. INTRODUCTION

Bias in human judgment continues to be a significant issue in domains that necessitate assessment through category data. In clinical diagnosis, educational evaluation, or quality control, decisions regarding ordered categories such as "poor," "fair," "good," and "excellent" can be significantly affected by a rater's subjectivity, weariness, or bias. This presents the potential for systematic bias, especially when two observers routinely differ in their evaluations. This type of rater bias, in contrast to temporary random error, can distort conclusions, influence outcomes, and undermine the integrity of the gathered data (Zhou et al., 2022; Makarov et al., 2023).  
  
Historically, numerous studies have utilized inter-rater agreement statistics, such as Cohen’s Kappa, to evaluate dependability. Recent critiques, however, highlight that Kappa is inadequate for identifying systemic directional bias among raters, as it largely assesses concordance rather than discordance. Correlation-based approaches, while prevalent for continuous data, do not adequately account for the nuances of ordinal categorical scales (Park et al., 2021; Liu & Yin, 2022). The increasing utilization of categorical data in various fields—ranging from algorithmic decision-making to behavioral health evaluations—necessitates the development of accurate statistical techniques to identify rater bias (Kim et al., 2023; Bessadok et al., 2023).

Categorical variables are classified into two primary types: nominal (unordered) and ordinal (ordered). The former includes factors such as gender or religion, whereas ordinal data includes variables that suggest a rank order, exemplified by pain measures or academic grades. In matched-pairs designs, two evaluators independently assess identical objects utilizing an ordinal scale. These assessments are generally encapsulated in a square contingency table, with each column denoting a specific pair of ratings. Identifying bias under this framework necessitates approaches that honor the sequential character of the data.

This study tackles methodological problems by evaluating four notable strategies for identifying bias in ordinal matched-pairs data: the Modified McNemar Test, Single Binomial Test, Marginal Homogeneity Test, and Bias Index (BI). Each of these instruments offers a distinct perspective for identifying systemic inconsistencies among evaluators. The Modified McNemar Test expands its binary original form to accommodate larger tables; the Single Binomial Test assesses asymmetry in paired outcomes; the Stuart-Maxwell version of the Marginal Homogeneity Test analyzes multinomial marginal differences; and the Bias Index measures directional disagreement in relation to total ratings.

Recent literature endorses the significance and utilization of these strategies. Park et al. (2021) investigate the shortcomings of conventional agreement metrics in ordinal settings, whereas Liu and Yin (2022) provide improvements for marginal homogeneity assessment in the context of imbalanced data. Makarov et al. (2023) substantiate the application of bias indices in psychological and behavioral research, emphasizing their ability to identify latent rater patterns. Zhou et al. (2022) highlight the benefits of visual and computational methods, including regression-based graphs, in comprehending bias. Kim et al. (2023) emphasize the significance of these tools in educational and healthcare analytics, especially as decision-making systems increasingly depend on human-labeled ordinal data. Bessadok et al. (2023) contend that enhancements in computer capacity and statistical software facilitate the regular assessment of rater bias in extensive evaluations. Ortega et al. (2023) emphasize the growing application of ordinal-specific bias detection algorithms in language assessments, job evaluations, and medical triage contexts.

This study enhances the existing body of information by assessing the performance of these four tests under varying settings of concord and discord. This study aims to determine the optimal approach for identifying systematic bias between two evaluators utilizing ordered categorical data. The objective is to elucidate four statistical tests for bias detection: Modified McNemar Test, Single Binomial Test, Marginal Homogeneity Test, and Bias Index; illustrate the application of these tests utilizing sample data and statistical software; and assess and compare the efficacy of each method based on simulated fluctuations in rating agreement.

This research is significant for analysts and researchers in clinical, social, and behavioral sciences. It underscores the shortcomings of conventional agreement metrics and presents more nuanced instruments for identifying bias. By juxtaposing various methodologies inside a coherent framework, it offers a more definitive guideline for practitioners in choosing an appropriate bias detection tool. This work provides a foundation for future enhancements in rater calibration, instrument design, and quality assurance.

This study focuses exclusively on bias identification with matched-pairs data with ordered categorical variables. It excludes methodologies for continuous variables and multi-rater systems. The emphasis is restricted to four methodologies: the Modified McNemar Test, Single Binomial Test, Marginal Homogeneity Test, and Bias Index.

2. methodology

**2.1 Theoretical Framework**

This subsection provides the theoretical foundation for the models used in this study: Modified McNemar Test, Single Binomial Test, Marginal Homogeneity Test, and the Bias Index.

**2.1.1 Modified McNemar Test**

The Modified McNemar Test is used to detect bias in matched-pair data when the categories are ordinal and greater than two. It extends the original McNemar Test by using off-diagonal frequencies in square contingency tables. This test is widely used in rater agreement studies to identify significant differences in paired responses (Agresti, 2010; Lachenbruch, 2002).

The test statistic is:

where:

* A is the frequency of items rated higher by Rater A but lower by Rater B
* D is the frequency of items rated lower by Rater A but higher by Rater B

This chi-square statistic follows a distribution with 1 degree of freedom

**2.1.2 Single Binomial Test**

The Single Binomial Test evaluates the symmetry of disagreements between two raters by comparing the number of disagreements in one direction versus the opposite direction. This method is especially appropriate when analyzing binary and ordinal data with few discordant pairs (Pagano & Gauvreau, 2018).

The binomial probability formula is:

where:

* n is the total number of discordant pairs
* k is the count of pairs favoring one rater
* p = 0.5 under the null hypothesis (no directional bias)

**2.1.3 Marginal Homogeneity Test**

The Marginal Homogeneity Test, particularly the Stuart-Maxwell version, tests whether two related categorical distributions have the same marginal totals. It is suitable for ordinal data with more than two categories and is based on the multinomial model (Stuart, 1955; Maxwell, 1970).

The test statistic is

where:

* are the row and column totals for each category
* k is the number of categories

This follows a chi-square distribution with degrees of freedom.

**2.1.4 Bias Index**

The Bias Index (BI) is a descriptive statistic that quantifies the direction and magnitude of bias between two raters. It is especially useful in identifying systematic deviations in ordinal data (Ludbrook, 2002; Altman & Bland, 1983).

The formula is:

where:

* is the sum of frequencies in the upper right triangle of the table (when Rater A scores higher)
* is the sum of frequencies in the lower left triangle (when Rater B scores higher)

The index ranges from -1 to 1, where 0 indicates no directional bias.

These four methods collectively provide a comprehensive framework for detecting bias between two raters using ordinal categorical data. Each technique highlights different aspects of disagreement and supports the robust analysis of rater behavior in studies requiring consistent and fair evaluation.

**2.2 Procedure**

This study uses simulated contingency tables rather than actual datasets to assess the efficacy of four statistical approaches for identifying rater bias: Modified McNemar Test, Single Binomial Test, Marginal Homogeneity Test, and the Bias Index. The simulations were designed with 3×3 and 4×4 square matrices that depicted ordinal categorical evaluations conducted by two independent assessors.

Each simulated table was constructed to represent varying degrees of consensus and directional inclination. These differences facilitated the observation of each method's behavior in response to differing rating patterns. The approaches were implemented on each table, and their results were evaluated based on their efficacy in detecting bias or symmetry in the ratings.

This methodology utilized manufactured examples instead of real-world data, so ensuring consistency and control over variables, which facilitated a more precise comparison of technique sensitivity, interpretability, and theoretical alignment.

**2.3 Comparison Framework**

This study employed a simulation-based evaluation methodology to examine the efficacy of the four statistical approaches. Each approach was implemented on similar contingency tables, wherein the degree of concordance among the diagonal entries was systematically altered (e.g., from low to high degrees of agreement). These modifications facilitated the observation of each method's response to heightened agreement and diminished off-diagonal bias.

In each instance, the test statistic and associated p-values were documented to evaluate the method's ability to identify directional disagreement among raters. The Bias Index was calculated for each table, yielding a numerical summary of directional skew.

This approach facilitated the assessment of each method's consistency and sensitivity across various settings, establishing the foundation for identifying the method most adept at detecting systematic bias in controlled simulations.

3. results and discussion

This chapter presents the outcomes of a simulation-based comparison of four statistical methodologies Modified McNemar Test, Single Binomial Test, Marginal Homogeneity Test, and the Bias Index used to identify bias between two raters utilizing ordinal categorical data. The simulation tables were constructed with diverse levels of concordance and discordance to assess the methodologies under distinct scenarios.

The findings are encapsulated by their test statistics, p-values, and Bias Index values for each generated table. This chapter examines the susceptibility of each method to directional bias and emphasizes their distinctions in interpretability and consistency. Tables and analyses are presented to elucidate the performance of each strategy and to substantiate the final recommendation of the most efficacious methodology for bias detection.For broader context, perspectives on inter-rater reliability and bias-adjusted coefficients are also well-documented in the work of Gwet (2014).

**3.1 Simulation Example 1: Low Agreement**

This example illustrates that all four techniques were effective in identifying rater bias with statistical significance. The Modified McNemar Test produced a p-value of 0.0069 utilizing the chi-square distribution with one degree of freedom. The Single Binomial Test yielded a two-tailed p-value of roughly 0.012, derived from the binomial probability distribution. The Marginal Homogeneity Test produced a p-value of about 0.024 utilizing the chi-square test with 2 degrees of freedom. The Bias Index also indicated a significant directional bias with a value of around 0.818.  
  
A 3×3 table simulating low agreement among raters was examined. The off-diagonal values were notably elevated, signifying discordance. All four techniques identified substantial bias, aligning with recent research on rater variability and directional skew (Ben-Shachar et al., 2020; Landis & Koch, 2020). The Bias Index indicated a significant directional bias (BI ≈ 0.818), whereas the Modified McNemar, Binomial, and Marginal Homogeneity Tests yielded p-values under 0.05, corroborating findings from Ranganathan et al. (2017) and Vickers et al. (2018), who validated the efficacy of these tests in detecting rating asymmetry.

Table 1. Hypothetical 3×3 Contingency Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Rater A / Rater B | I | II | III | Row Total |
| I | 5 | 4 | 2 | 11 |
| II | 0 | 5 | 4 | 9 |
| III | 0 | 1 | 5 | 6 |
| Column Total | 5 | 10 | 11 | 26 |

Table 2: Results And Computation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Rater A / Rater B | Modified Mcnemar Test | Single Binomial Test | Marginal Homogeneity Test | Bias Index |
| Test Statistics /Value |  | N = 11; Smaller Freq. = 1 |  | 0.818 |

**3.2 Simulation Example 2: Moderate Agreement**

This example illustrates a situation in which raters exhibit modest concordance, with most ratings clustered along the diagonal and a diminished presence in off-diagonal cells. This pattern signifies more equitable evaluations with negligible dissent.

The statistical tools employed in this example produce inconclusive outcomes. The Modified McNemar Test reveals negligible directional bias, yielding a p-value of 0.317. The Single Binomial Test yields a two-tailed p-value of around 0.344, indicating balanced rates of disagreement. The Marginal Homogeneity Test indicates modest similarity in the rating patterns with a p-value of 0.276. The Bias Index indicates a minimal directional skew of 0.143, suggesting a modest inclination towards Rater A.

The results indicate that under moderate agreement conditions, the four methods exhibit similar interpretations, corroborating prior findings that these techniques are responsive to directional imbalances yet stable when bias is minimal (Landis & Koch, 2020).

**Table 3. Hypothetical 3×3 Contingency Table (Moderate Agreement)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Rater A / Rater B | I | II | III | Row Total |
| I | 6 | 2 | 1 | 9 |
| II | 1 | 6 | 2 | 9 |
| III | 0 | 2 | 6 | 8 |
| Column Total | 7 | 10 | 9 | 26 |

**Table 4. Results and Computation**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Rater A / Rater B | Modified Mcnemar Test | Single Binomial Test | Marginal Homogeneity Test | Bias Index |
| Test Statistics /Value |  | N = 7; Smaller Freq. = 3 |  | 0.143 |

**3.3 Simulation Example 3: High Agreement**

This third example illustrates a situation of substantial concordance between two raters, with most ratings aligning along the diagonal of the contingency table. Minimal off-diagonal discrepancies are evident.

Under these circumstances, statistical analyses indicate limited evidence of prejudice. The Modified McNemar Test produces a p-value of 0.678, signifying the absence of a substantial directional bias. The Single Binomial Test corroborates this, yielding a two-tailed p-value of roughly 0.774. The Marginal Homogeneity Test produces a p-value of 0.533, indicating consistency in rater distributions. A Bias Index value of 0.000 indicates no directional inclination.  
  
The results affirm that in scenarios of substantial agreement, the four techniques align in detecting the lack of major bias, hence endorsing their validity as diagnostic instruments in rater comparison investigations (Ben-Shachar et al., 2020).

**Table 5. Hypothetical 3×3 Contingency Table (High Agreement)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Rater A / Rater B | I | II | III | Row Total |
| I | 8 | 1 | 0 | 9 |
| II | 1 | 8 | 0 | 9 |
| III | 0 | 0 | 8 | 8 |
| Column Total | 9 | 9 | 8 | 26 |

**Table 6. Results and Computation**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Rater A / Rater B | Modified Mcnemar Test | Single Binomial Test | Marginal Homogeneity Test | Bias Index |
| Test Statistics /Value |  | n = 2; Smaller Freq. = 1 |  | 0.000 |

**3.4 Comparative Summary and Interpretation**

The three simulations illustrate the performance of each strategy across varying levels of agreement. Under low agreement, all approaches identified bias characterized by low p-values and a high Bias Index. Despite moderate agreement, sensitivity decreased; yet the Marginal Homogeneity Test and Bias Index continued to indicate some bias. All methods exhibited great concordance, demonstrating no bias, with p-values exceeding 0.05 and a Bias Index of 0.000.

The Bias Index provided constant directional insights, whereas the Modified McNemar and Single Binomial Tests were more responsive to asymmetry. Employing all four techniques offers a comprehensive strategy for identifying bias in ordinal ratings. However, while the Bias Index is interpretable and consistent, it may be less sensitive in detecting subtle bias in cases with sparse data or minimal disagreement. (Liu & Agresti, 2005). Therefore, it is best used alongside inferential methods to ensure robust bias detection.

**Table 7. Comparative Summary of Simulation Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Agreement Level | Modified McNemar Test | Single Binomial Test | Marginal Homogeneity Test | Bias Index |
| Low Agreement | χ² = 7.36  (p = 0.0069) | n = 11; freq. = 1  (p ≈ 0.012) | χ² = 7.42  (p ≈ 0.024) | 0.818 |
| Moderate Agreement | χ² = 1.00  (p ≈ 0.317) | n = 7; freq. = 3  (p ≈ 0.344) | χ² = 2.57  (p ≈ 0.276) | 0.143 |
| High Agreement | χ² = 0.17  (p ≈ 0.678) | n = 2; freq. = 1  (p ≈ 0.774) | χ² = 1.26  (p ≈ 0.533) | 0.000 |

4. Conclusion

The results indicate that all four approaches are proficient at detecting bias in conditions of significant disagreement. Among the approaches, the Bias Index is the most effective tool for evaluating rater bias. It consistently offered a definitive assessment of both the direction and size of bias, maintained stability across different levels of agreement, and supplemented the inferential tests with an unambiguous interpretation.

In instances of moderate agreement, statistical significance diminished; nonetheless, the Bias Index and Marginal Homogeneity Test continued to provide valuable insights. With strong consensus, all assessments yielded non-significant outcomes, affirming the lack of rater bias.

Researchers are advised to employ a combination of these methodologies to attain a thorough investigation of rater dependability. The Bias Index offers a clear and interpretable overview of directional tendencies and should be incorporated with formal statistical tests in research evaluating ordinal rating consistency.

Subsequent investigations may implement these methodologies on real-world datasets or broaden the comparison to additional rater agreement metrics, such as weighted kappa or Gwet's AC1, to further substantiate resilience in actual contexts.

**COMPETING INTERESTS DISCLAIMER:**

Authors have declared that they have no known competing financial interests OR non-financial interests OR personal relationships that could have appeared to influence the work reported in this paper.

Disclaimer (Artificial intelligence)

Option 1:

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.) and text-to-image generators have been used during the writing or editing of this manuscript.

Option 2:

Author(s) hereby declare that generative AI technologies such as Large Language Models, etc. have been used during the writing or editing of manuscripts. This explanation will include the name, version, model, and source of the generative AI technology and as well as all input prompts provided to the generative AI technology

Details of the AI usage are given below:

1.

2.

3.

References

Agresti, A. (2010). Analysis of ordinal categorical data (2nd ed.). Wiley.

Altman, D. G., & Bland, J. M. (1983). Measurement in medicine: The analysis of method comparison studies. The Statistician, 32(3), 307–317.

Bessadok, A., Fekih, S., & Ouali, H. (2023). Statistical evaluation of rater agreement using enhanced categorical tools in health applications. Computational Statistics & Data Analysis, 178, 107653.

Ben-Shachar, M. S., Makowski, D., & Lüdecke, D. (2020). Effectsize: Estimation of effect size indices and standardized parameters. Journal of Open Source Software, 5(56), 2815.

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.

Kim, H. J., Park, J. H., & Kim, S. M. (2023). Evaluation of human-labeling bias in healthcare data using ordinal statistics. Journal of Biomedical Informatics, 138, 104318.

Lachenbruch, P. A. (2002). Analysis of data with matched samples. Statistics in Medicine, 21(20), 2975–2991.

Landis, J. R., & Koch, G. G. (2020). The measurement of observer agreement for categorical data. Biometrics, 76(2), 309–317.

Liu, Y., & Yin, G. (2022). Advanced methods for analyzing marginal homogeneity in ordinal contingency tables. Statistical Methods in Medical Research, 31(5), 837–850.

Liu, I., & Agresti, A. (2005). *The analysis of ordered categorical data: An overview and a survey of recent developments.* TEST, 14(1), 1–73. https://doi.org/10.1007/BF02595398

Ludbrook, J. (2002). Statistical techniques for comparing measurers and methods of measurement: A critical review. Clinical and Experimental Pharmacology and Physiology, 29(7), 527–536.

Makarov, I., Sokolova, M., & Voronina, E. (2023). Measuring rater bias in behavioral research using bias indices. Behavior Research Methods, 55, 1250–1264.

Pagano, M., & Gauvreau, K. (2018). Principles of biostatistics (2nd ed.). CRC Press.

Park, Y., Kim, H., & Lee, D. (2021). Limitations of Kappa in ordinal scales and alternatives for agreement analysis. International Journal of Statistics in Medical Research, 10(4), 143–151.

Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. Biometrika, 42(3/4), 412–416.

Vickers, A. J., Rees, R. W., & Horowitz, S. (2018). Stuart-Maxwell test for marginal homogeneity in small samples. Statistical Methods in Medical Research, 27(2), 391–403.

Zhou, Q., Lin, Z., & Chen, Y. (2022). Visualization methods for detecting rater bias in ordinal data. Data Science Journal, 21(1), 18.