

Measuring Level of Agreement in Students' Academic Performance Comparing Kappa and Scott Pi Agreement Measures.

¹Onyenekwe Chukwuenyem Enoch, ²Adejumo Olusola Adebawale

¹Department of Statistics, Nnamdi Azikiwe University, PMB 5025, Awka, Nigeria.

²Department of Statistics, University of Ilorin, PMB 1515, Ilorin, Nigeria.

DOI: <https://doi.org/10.51583/IJLTEMAS.2025.1407000023>

Received: 05 July 2025; Accepted: 11 July 2025; Published: 01 Aug 2025

Abstract: Kappa and Scott Pi statistics are agreement measures that are used to measure the level of agreement between two raters. The academic performance of students used is the Cumulative Grade Point Average (CGPA) of year one and the final year. Year One is classified as rater one and that of the final year is classified as rater two. The Analysis was done to see the consistency in the students' performance, if the level of agreement is high, i.e. if there is agreement between their first-year result and that of the final year. A comparison of the Kappa Estimates and that of the Scot Pi was also done to determine the most efficient measure between the two. Kappa Agreement outperforms the Scot pi measures in terms of precision and efficiency. Also, there is no agreement between the Year One Performance and that of the Final Year

Keynotes: Kappa, Scott Pi, Proportion, Contingency table,

I. Introduction

In the realm of data analysis and research, agreement measures are crucial tools used to quantify the level of concordance or consistency between different observers and raters.

Agreement measures are relevant in psychology, medicine, and social sciences, where accurate and consistent data is essential for making informed decisions and drawing valid conclusions. By employing various statistical techniques and metrics, researchers and analysts can identify the degree of agreement or disagreement between assessments, ensuring that their findings are robust and trustworthy. In the field of statistical analysis and research, Kappa statistics are a widely used tool for assessing inter-rater agreement or the consistency between categorical variables. Developed by Jacob Cohen in 1960, the Kappa statistic quantifies the level of agreement between two or more raters or observers, taking into account the possibility of agreement occurring by chance.

The Kappa statistic is particularly valuable in evaluating the reliability of categorical data in studies where subjective judgments or classifications are involved. For instance, in medical diagnostics, Kappa can be used to measure how consistently different clinicians diagnose a condition. In content analysis, it helps in assessing the reliability of coding schemes applied by multiple researchers.

The Kappa value ranges from -1 to +1, where a value of +1 indicates perfect agreement, 0 indicates agreement no better than chance, and negative values suggest worse-than-chance agreement. By providing a more nuanced measure of agreement compared to simple percent agreement, the Kappa statistic offers a clearer picture of the reliability of categorical assessments and helps ensure that findings are both accurate and replicable.

Similarly, Scott Pi is a statistical measure used to assess the agreement level between raters beyond chance its values spreads from -1 to 1, with 0 depicting no agreement beyond chance. The primary purpose of Scott's pi is to measure the overall consistency of a rating system by measuring the degree of agreement between raters.

Both Cohen's Kappa and Scott's Pi measure the degree of agreement between two raters when evaluating categorical data (like grades or performance levels), while accounting for the agreement expected by chance.

II. Review of Related Literature

Evaluating the extent of agreement between two or between several raters is common in social, behavioral and medical sciences. To assess the extent to which a given characterization of a subject is reliable, it is clear that we must have a number of subjects classified more than once, that is, by more than one rater. The degree of agreement among the raters provides not more than an upper bound on the degree of accuracy present in the ratings. However, if agreement among the raters is good, then there is a possibility, but by no means a guarantee, that the ratings do in fact reflect the dimension they are purported to reflect. If their agreement is poor, on the other hand, then the usefulness of the ratings is severely limited, for it is meaningless to ask what is associated with the variable being rated when one cannot even trust those ratings to begin with. Agreement can also be termed a special case of association which reflects the extent to which observers classify a given subject identically into the same category (Adejumo 2002). Davide Chicco et al (2022) noted that Mathew correlation coefficient is more informative than the cohen's Kappa. Hence the need for the modification. Alberto M. Marchevsky et al (2020) opined that pathologists should forget about Cohen Kappa since it doesn't provide a more clinically applicable measure for interobserver measure. In order to assess the psychometric integrity of different ratings, we compute interrater agreement. It pertains to the extent to which the raters classify a given subject

identically into the same category. Kozlowski and Hat-trup (1992) noted that an interrater agreement index is designed to reference the interchangeability among raters; it addresses the extent to which raters make essentially the same ratings". Farzan Madadzadeh et al (2023) stated that Statistical methods have always been the solution to medical problems. Due to the problem of inconsistency in the diagnosis of dentists, statistical science has been provided for measuring the compatibility of diagnoses and the reliability of dentists. One of the most important statistical methods for examining the agreement between two experiments or diagnoses is Kappa statistics, which can be used in dental sciences.

Gerald Rau and Yu-Shan Shi (2021) stated that Cohen's kappa (κ) is often recommended for nominal data as a measure of inter-rater (inter-coder) agreement or reliability. In their work, they investigated which term is appropriate for genre analysis and what statistical measures are valid to measure it.

Many literatures has proposed different means of measuring the raters' agreement with different and under various practical situations, like the nominal, ordinal, and continuous scales. Kraemer (1988,1997), Verducci et al(1988), Barnhart and Williamson (2002), and others discussed basic issues in evaluating agreement and described a variety of ways of measuring it. Other authors like Fleiss(1971) , Landis and Koch (1977a,1977b), Schoulten (1982), and Shoukri have described rater agreement using generalizations of Cohen's kappa (Cohen,1960). Others like Jorgenson (1985), Darroch et al.(1986), Tanner and Young (1985a,1986b), Agrestic (1982), and Basu et al (2000) showed difficulties with summarizing pairwise agreement by a single measure, and instead proposed modelling the structure of agreement among raters. Two of the most influential studies in this area are those of Fleiss (1971) and Fleiss et al (1969), which contain the most popular results in use today. Fleiss et al (1969) provide large sample approximations of the variances of the k and weighted k statistics suggested by Cohen (1960, 1968), respectively in the case of two raters, while Fleiss extends the p -statistic to the case of multiple raters. Landis and Koch (1977) also gave an instructive discussion of inter-rater agreement among multiple observers. Agresti(2002) presents several modelling techniques for analysing rating data in addition to presenting a short account of the state of the art. Light (1971) introduced measures of agreement conditionally upon a specific classification category, and proposed a generalization of Cohen's k -coefficient to the case of multiple raters. Conger (1980) suggested an alternative multiple-rater agreement statistic obtained by averaging all pairwise overall and chance-corrected probabilities proposed by Cohen (1960). Conger (1980) also extended the notion of pairwise agreement to that of g -wise agreement where agreement occurs if g raters rather than two raters classify an object into the same category. Moss (2024) advances inter-rater reliability by introducing Fréchet variance-based weighting for g -rater comparison, establishing sound asymptotic inference, and recommending transformation-based confidence intervals. This enriches kappa measures for complex real-world settings involving more than two raters.

Ghafari and Bahariniya (2023) provided a comprehensive examination of the application of Kappa statistics as a measure of agreement in clinical and diagnostic settings, with a particular focus on dental examinations. Their work emphasized the importance of selecting appropriate agreement coefficients based on the number of raters and the nature of the data (nominal or ordinal). The authors reviewed commonly used Kappa variants, including Cohen's Kappa for two raters, Fleiss' Kappa for multiple raters, and the weighted Kappa for ordinal scales. They also highlighted practical challenges in interpreting Kappa values, particularly when prevalence and marginal distributions affect the coefficient (a phenomenon often referred to as the "Kappa paradox"). To support practitioners, the paper offered step-by-step guidance on implementing these statistics using software such as R, SPSS, and Stata. Overall, the study reinforced the value of Kappa statistics in assessing inter-rater reliability and recommended thoughtful selection of agreement measures depending on the specific research or clinical context.

Zee, Mariani, Barisoni, Mahajan, & Gillespie (2024) present a **novel agreement coefficient** that systematically incorporates **rater-reported uncertainty** into the chance-correction framework. Rather than solely relying on marginal distributions to estimate chance agreement—as in Cohen's κ , Scott's π , or Gwet's AC1—their method uses empirical uncertainty data: any rating marked as uncertain is treated as potentially due to chance agreement. This leads to a more grounded estimate of true agreement, effectively removing "chance-induced" concordance from the observed probability and refining the denominator in standard agreement formulas

III. Methodology

Let us consider a reliability experiment where two raters referred to as raters A and B respectively must classify N subjects into one of two response categories. The two response categories labelled as 1 and 2 are assumed to be disjoint. (i.e. do not overlap).

Table 1 describes the outcome of this reliability experiment. It follows from that of all N subjects , A are classified by both raters into category 1, while B subjects were classified into category 2 by rater A and in category 1 by rater B. It should be noted that B_1 and B_2 represents the number of subjects that rater B classified into categories 1 and 2 respectively. A_1 and A_2 are also defined similarly.

Table 1

Rater B	Rater A		Total
	1	2	

1	A	B	B1=A+B
2	C	D	B2= C+D
Total	A1= A+C	A2= B+D	N

Scott's Pi Statistic.

Scott (1955) suggested computing the extent of agreement between raters A and B using the π - statistic PI, which is defined as

$$PI = \frac{p - e(\pi)}{1 - e(\pi)}, \quad 1$$

Where $P = \frac{(A+D)}{N}$ is the overall agreement propensity and $e(\pi)$ is given by

$$e(\pi) = \left\{ \frac{\frac{A1 + B1}{2}}{N} \right\}^2 + \left\{ \frac{\frac{A2 + B2}{2}}{N} \right\}^2 \quad 2$$

it should be noted that $e(\pi)$ designates the propensity for both raters to agree by chance without having the same assessment of a subject.. The P component of eqn 1 only involves subjects both raters have classified in the same category. It could be used as a naïve measure of the extent of agreement. However, there are reasons to believe that raters A and B would classify some subjects into the same category not for same reasons. Subjects classified in the same category for different reasons correspond to an agreement by chance. Hence Scott (1955) recommend eqn 2 as a measure of the chance agreement probability. Furthermore, it is worthy to state that Scott π involves a further assumption of homogenous rater marginal under chance.

Cohen's Kappa Statistic

This measures the agreement between two raters who each classify N items into mutually exclusive categories. Cohen (1960) criticized Scott's approach for computing chance agreement probability because it combines raters A and B classification data. The first term of the eqn 2 is obtained by averaging the proportions of subjects classified into category 1 by both raters and raising the average to the power of 2. This approach eliminates any difference that may exist in the rating pattern of both raters.

$$kappa = \frac{p - e(k)}{1 - e(K)} \quad 3$$

Where $p = \frac{A+D}{N}$ and $e(K) = \left[\frac{A1}{N} \right] \left[\frac{B1}{N} \right] + \left[\frac{A2}{N} \right] \left[\frac{B2}{N} \right]$ 4

The first term of the eqn 4 estimates the chance that both raters independently classify a subject into category 1. The second term on the other hand estimates the probability of independent classification of a subject into category 2. Unlike Scott's equation 2 , the terms in equation 4 are obtained by multiplying raters classification rates. Cohen's Kappa is the most widely used measure of agreement and possesses some good qualities.

Data Presentation

The academic performance of students used is the Cumulative Grade Point Average (CGPA) of the first year and the final year from the Department of Statistics, University of Ilorin, Kwara State. Year 1 is classified as rater one and that of the final year as rater two. Those with First Class and Second Class Upper Division were classified together, and Second Class Lower Division with Third Class were classified together.

Table 2

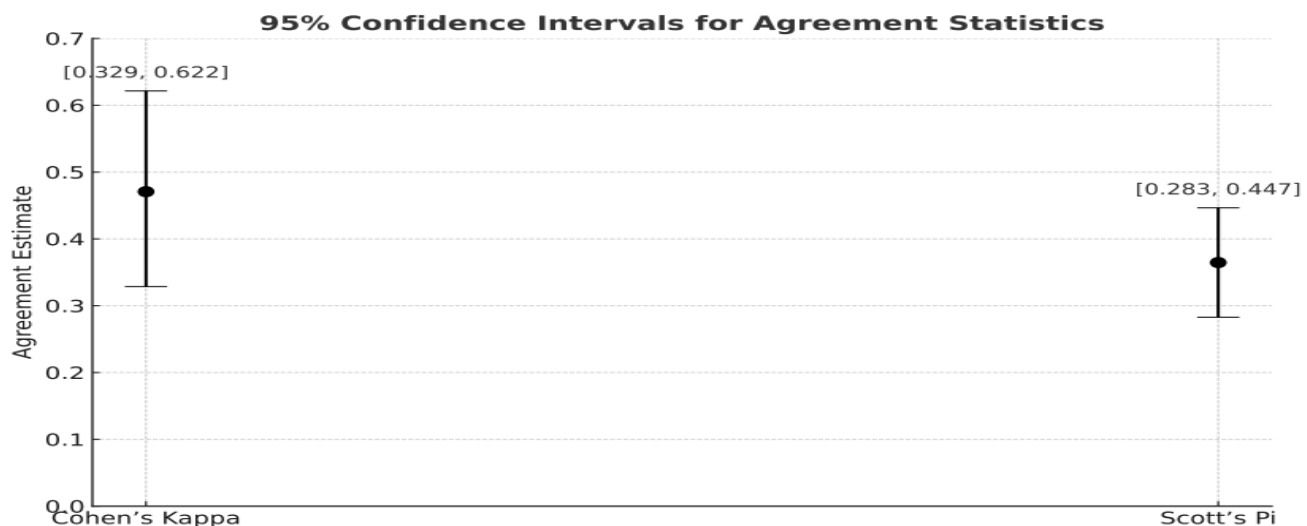
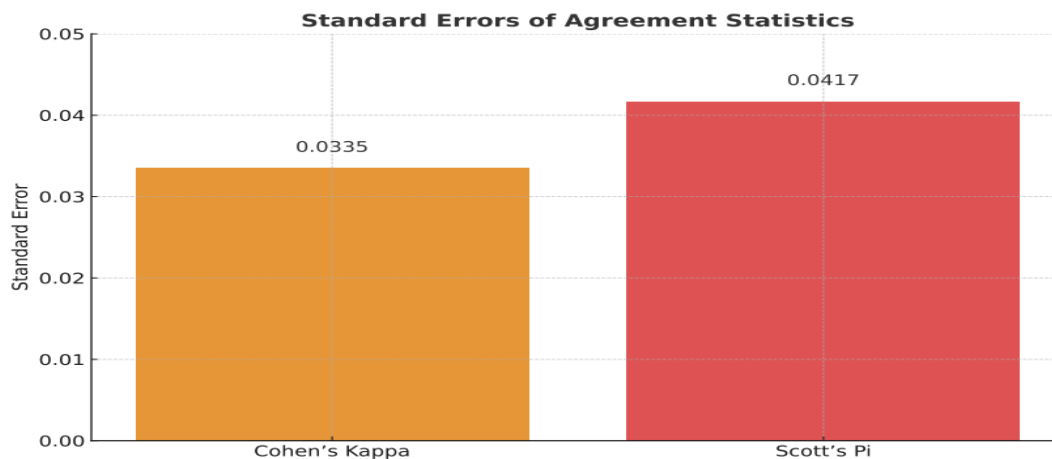
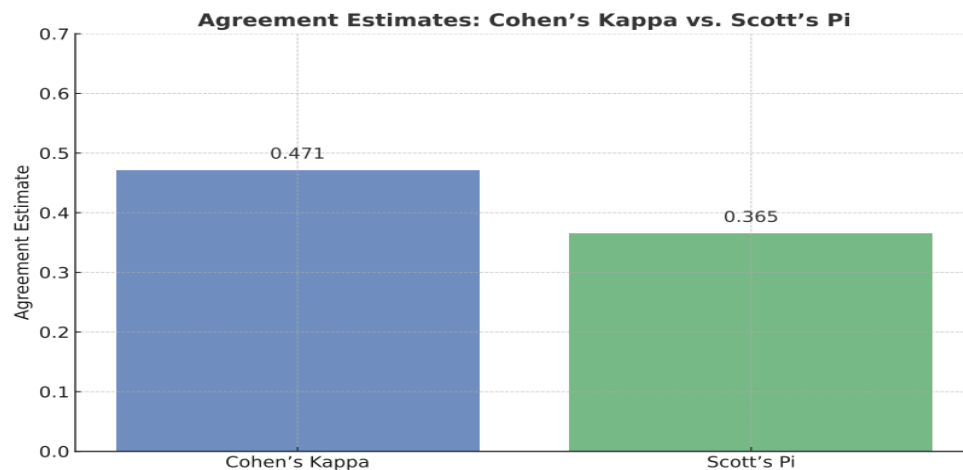
Rater B (Final Year)	Rater A(Year 1)	
	1(1 st class& 2 ¹)	2 (2 ² & # 3rd Class)
1 (1 st class& 2 ¹)	43	31
2(2 ² & # 3rd Class)	4	54

IV. Results

R statistical package was used to analyse the data the below results gotten

Table 3 Comparison of Estimates for Kappa Agreement Measure and Scott Pi Agreement: Performance Metrics and Error Analysis

Method	Estimate	SE/SD	95%LI	95%UI				
Kappa	0.4707146	0.033500606	0.3286855	0.6215447				
Scott Pi	0.3646436	0.04173824	0.2828366	0.4464505				



Statistic	Cohen'S kappa	Scott's Pi	Explanation
Estimate	0.4707	0.3646	Moderate Jappa, Fair Pi Agreement
Standard Error(SE)	0.0335	0.0417	Kappa Has higher Precision
95% Confidence Interval	(0.3287,0.6215)	(0.2828,0.4465)	Kappa range includes substantial agreement
Agreement Level	Moderate	Fair	Kappa Indicates stronger Consistency
Assumes Equal Category Distribution	No	Yes	Pi may underestimate agreement
Adjust for rater bias	Yes	No	Kappa accounts for differences in rater use
Usefulness in Real -world settings	High	Moderate	Kappa better reflects practical conditions

V. Conclusion

Both Cohen's Kappa and Scott's Pi were calculated to assess inter-rater agreement in student performance ratings. Cohen's Kappa yielded an estimate of 0.4707, indicating moderate agreement, while Scott's Pi produced a lower estimate of 0.3646, suggesting only fair agreement.

Kappa also demonstrated higher precision, with a smaller standard error (0.0335 vs. 0.0417) and a wider confidence interval extending into the "substantial" agreement range ([0.3287, 0.6215]) compared to Pi's narrower range ([0.2828, 0.4465]).

Importantly, Cohen's Kappa adjusts for differences in rater behavior, whereas Scott's Pi assumes identical use of rating categories. This makes Kappa more suitable for real-world educational assessments where evaluators may not distribute ratings uniformly.

Therefore, Cohen's Kappa is a more efficient and informative measure for assessing agreement in this context.

Also in terms of the students' performance, there is no agreement in their performance. This implies that there is a difference in their performance for the different levels. The consistency is not strong. Students will need to be followed up to ensure consistency in their performances.

References

- Adejumo A.O (2010). A data based approach of imputing missing ratings in raters agreement measurement for two raters. ICATOR Journal of mathematical Sciences. Vol 4,No 149-67
- Adejumo A.O, Heumann C, Toutenburg H.(2004) A review of agreement measure as a subset of association measure between raters, sonderforschungsbereich
- Albert, J. H., Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. J. Amer. Statist. Assoc. 88:669–679.
- Alberto M. Marchevsky, Ann E. Walts, Birgit I. Lissenberg-White, Erk Thunnissen (2020) Pathologists should probably forget about kappa. Percent agreement, diagnostic specificity and related metrics provide more clinically applicable measures of interobserver variability. [Annals of Diagnostic Pathology Volume 47](#), August 2020, 151561
- Banerjee, M., Capozzoli, M., McSweeney, L., Sinha, D. (1999). Beyond kappa: a review of interrater agreement measures. Can. J. Statist. 27:3–23.
- Barlow, W. (1996). Measurement of interrater agreement with adjustment for covariates. Biometrics 52:695–702.
- Basu, S., Nanerjee, M., Sen, A. (2000). Bayesian inference for kappa from single and multiple studies. Biometrics 56:577–582.
- Bloch, D. A., Kraemer, H. C. (1989). 2 × 2 kappa coefficients: measures of agreement or association. Biometrics 45:269–287.

9. Bond, M. E., Higgins, J. J. (2001). A note on 'a comparison of Bayes and maximum likelihood estimation of the intraclass correlation coefficient'. *Commun. Statist. Theor. Meth.* 30:371–380.
10. Chib, S., Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika* 85: 347–361.
11. Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational Psychology Measure.* 20:37–46. Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled or partial credit. *Psychol. Bull.* 70:213–220.
12. D. Chicco, M. J. Mattheijs, G. Jurman. (2021). MCC versus Cohen's Kappa and Brier score. .DOI 10.1109/ACCESS.2021.3084050, IEEE Access
13. Drasgow, F. (1988). Polychoric and polyserial correlations. In: Kotz, L., Johnson, N. L., eds. *Encyclopedia of Statistical Sciences*. Vol. 7. New York: Wiley, pp. 69–74.
14. [Farzan Madadzadeh](#), [Hesam Ghafari](#) and [Sajjad Bahariniya](#) (2023) Kappa Statistics: A Method of Measuring Agreement in Dental Examinations. *The Open Public Health Journal*
15. Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychol. Bull.* 76:378–382. Hale, C. A., Fleiss, J. L. (1993). Interval estimation under two study designs for kappa with binary classifications. *Biometrics* 49:523–534.
16. Jonas Moss (2024). Measures of Agreement with Multiple Raters: Fréchet Variances and Inference. [Psychometrika](#), [Volume 89](#), [Issue 2](#), June 2024, pp. 517 - 541
17. DOI: <https://doi.org/10.1007/s11336-023-09945-2>
18. Kraemer, H. C., Periyakoil, V. S., Noda, A. (2002). Kappa coefficients in medical research. *Statist. Med.* 21:2109–2129.
19. Lee, J. J., Tu, Z. N. (1994). A better confidence interval for kappa on measuring agreement between two raters with binary outcomes. *J. Computat. Graph. Statist.* 3:301–321.
20. Lipsitz, S. R., Laird, N. M., Brennan, T. A. (1994). Simple moment estimates of the kappa coefficient and its variance. *Appl. Statist.* 43:309–323.
21. Liu, C. (2001). Bayesian analysis of multivariate probit model: discussion of “the art of data augmentation” by Van Dyk and Meng. *J. Computat. Graph. Statist.* 10:75–81.
22. Nam, J. (2000). Interval estimation of the kappa coefficient with binary classification and an equal marginal probability model. *Biometrics* 56:583–585.
23. Nandram, B., Chen, M.-H. (1994). Accelerating Gibbs sampler convergence in the generalized linear models via a reparameterization. *J. Statist. Computat. Simul.* 81: 27–40.
24. Ochi, Y., Prentice, R. L. (1984). Likelihood inference in a correlated probit regression model. *Biometrika* 71:531–543.
25. Pan, W., Wall, M. M. (2002). Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Statist. Med.* 21:1429–1441.
26. Pearson, K. (1900). *Mathematical contributions to the theory of evolution*. vii. on the correlation of characters not quantitatively measurable.
27. Zee, M. D., Mariani, L., Barisoni, L., Mahajan, S., & Gillespie, M. (2024). A novel agreement statistic using data on uncertainty in ratings. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 72(5), 1293–1314. <https://doi.org/10.1093/jrssc/qnad055>