Euroasia Journal of Mathematics, Engineering, Natural & Medical Sciences
International Indexed and Refereed
ISSN 2667-6702

Status of Weighted Agreement Statistics Between Two Raters in Ordinal Data Affected by Sample size and Number of Categories

Ordinal Verilerde İki Değerlendirici Arasındaki Uyum İstatistiklerinin Örneklem Büyüklüğünden ve Kategori Sayısından Etkilenme Durumları

Semra Erdoğan[1*], Damla Hazal Sucu[2]

[1]Assist. Prof. Dr., Mersin University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Mersin, Türkiye

[2] Res. Assist., Mersin University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Mersin, Türkiye

** Corresponding author: semraerdogann@gmail.com*

Geliş Tarihi / Received: 09.06.2023 *Araştırma Makalesi/Research Article*
Kabul Tarihi / Accepted: 18.07.2023 *DOI: 10.5281/zenodo.8239340*

ÖZET

Bu çalışmada amaç, sıralı ölçeklerde kullanılan değerlendiriciler arası ağırlıklandırılmış uyum istatistiklerini tanıtmak, sıralı ölçekler için kullanılan Spearman korelasyon katsayısı ile ağırlıklandırılmış uyum istatistiklerini karşılaştırmak ve örneklem büyüklüğünden, kategori sayısından etkilenme durumlarını ortaya koymaktır. Cohen's κ, Scott's π, Brennan-Prediger's (B-P), Gwet's AC2 and Krippendorff's Alpha are some of the common chance-corrected agreement measures to assess the agreement among two raters for ordinal outcome. The Pearson correlation, Spearman correlation and ICC are widely used for assessing reliability when ratings are on an interval scale. Both weighted agreement coefficients and correlation coefficients can be used to assess the reliability of ordinal rating scales. Bu amaçla, iki değerlendirici arasında ilişki yok iken, düşük, orta ve yüksek ilişki var iken farklı örneklem büyüklükleri ve kategori sayıları için veriler üretilmiş ve sözü edilen ağırlıklandırılmış uyum istatistikleri hesaplatılmıştır. Cohen's kappa, Scott's π ve Krippendorff Alpha katsayılarının korelasyon katsayıları ile benzer sonuçlar verdiği, B-P uyum istatistiğinde korelasyon katsayısı değerine çok yakın değerler aldığı söylenebilir. Ancak, Gwet's AC2 istatistiği, özellikle kategori sayısı 3 için, değerlendiriciler arasında ilişkinin olmadığı/düşük bir ilişkinin söz konusu olduğu durumlarda, korelasyon katsayısı değerinden farklılık gösterdiği ve değerlendiriciler arasında şansa bağlı da olsa orta düzeyde bir uyumdan bahsedilebileceği söylenebilir. Sıralı ölçeklerde iki değerlendirici arasındaki uyum araştırılırken, sadece kategori sayısı 3 olduğu durumlarda dikkat edilmesi ve uyum istatistiği olarak Gwet's AC2 uyum istatistiğinin kullanılması tavsiye edilmektedir. Bunun dışındaki diğer durumlarda, uyum ile ilişki kavramının birbirinin yerine gönül rahatlığı ile kullanılabileceği söylenebilir.

Anahtar Kelimeler: Brennan-Prediger, Gwet's AC2, Krippendorff's Alpha, Lineer ağırlıklı, Karesel ağırlıklı, Spearman korelasyon katsayısı.

ABSTRACT

The aim of this study is to introduce weighted inter-rater agreement statistics used in ordinal scales, compare weighted agreement statistics along with the Spearman correlation coefficient and reveal their status of being affected by the sample size and number of categories. For this purpose, data for different sample sizes and number of categories were produced for the cases when there is no relationship or when there is low, medium, or high relationship between the two raters, and the aforementioned weighted agreement statistics were calculated. It can be said that Cohen's kappa, Scott's π, and Krippendorff's alpha coefficients provide similar results with the correlation coefficients, and they get values very close to the correlation coefficient value in the B-P agreement statistics. However, it can be said that Gwet's AC2 statistic, especially for a category number of 3, differs from the correlation coefficient value in cases where there is no/low correlation between the raters, and a moderate level of agreement can be mentioned between the raters, albeit by chance. While investigating the agreement between two raters in ordinal scales, it is recommended to be careful in cases when only the number of categories is three and to use Gwet's AC2 agreement statistics. In other cases, it can be said that the concepts of agreement and relationship can be used interchangeably with peace of mind.

Keywords: Brennan-Prediger, Gwet's AC2, Krippendorff's Alpha, Linear weighted, Quadratic weighted, Spearman correlation coefficient.

1. INTRODUCTION

Inter-rater reliability, the agreement between measurements of more than one rater/method/device/scale, and intra-rater reliability, the agreement between measurements taken at different times by a single rater/method/device/scale, are important in many different disciplines (medicine, biology, psychology, psychology, and epidemiology, among others). In the field of medicine, there are many screening, imaging and diagnostic tests used to confirm that an individual is sick or healthy. Accurate and reliable diagnosis using these tests and later determining the most appropriate method to be applied for the treatment of the disease is very important for physicians and even for patients. In agreement studies, when evaluating the association of classifications of patient's disease or health status made by two or multiple raters, challenges arise when an ordered scale is used (Nelson & Edwards, 2018).

The biggest problem of the application of the agreement analyses is to determine the statistical method to be used. The statistical method to be applied varies according to the structure of the outcome variable, i.e., whether the outcome variable is continuous, discrete, binary, nominal, or ordinal, whether the outcome variable provides normality conditions, and depending on the number of raters, the number of diagnostic tests, and the number of categories in diagnostic tests (Lin *et al.*, 2012; Gwet, 2014; Kanık *et al.*, 2010). As a result, although there are different classifications in the literature, agreement statistics were classified according to the number of raters and scale type, as summarized in Table 1 (Lin *et al.*, 2012; Kanık *et al.*, 2012; Barnhart *et al.*, 2007; Lin, 2008; Haber *et al.*, 2005; Lin *et al.*, 2007; Haber & Barnhart, 2008; Tran *et al.*, 2020; Gwet, 2015).

Table 1. Inter-rater agreement measures summary

Type of data	Inter-rater Agreement Measures	Number of Raters
Binary/Nominal	Cohen's Kappa coefficient	Two raters
	Scott's Pi coefficient	Two raters
	G-index	Two raters
	Gwet's AC1 coefficient	Two and more than raters
	Krippendorff's Alpha Coefficient	Two and more than raters
	Fleiss Kappa coefficient	Two more than raters
	GEE	Two more than raters
Ordinal	Weighted Cohen's Kappa Coefficient	Two raters

	Weighted Scott's π Coefficient	Two raters
	Brennan-Prediger coefficient	Two raters
	Gwet's AC2 coefficient	Two and more than raters
	Krippendorff's Alpha Coefficient	Two and more than raters
	Kendall W coefficient	Two more than raters
Continuous	Concordance Correlation Coefficient	Two raters
	Bland &Altman Methods	Two raters
	Deming Regression Technique	Two raters
	Passing-Bablok Technique	Two raters
	Mountain Plot	Two and more than raters
	Intraclass Correlation Coefficient	Two and more than raters
	Krippendorff's Alpha Coefficient	Two and more than raters
	Fleiss Kappa coefficient	Two more than raters

It is observed that classical statistical methods such as Pearson correlation coefficient, regression analysis, independent sample t-test, Chi-square test or kappa statistics are widely used in many agreement studies (Bland & Altman, 2010). Stralen *et al.* (2012) revealed that systematic error is ignored when the Pearson correlation coefficient is used for testing the agreement between two continuous measurement methods, the effects of prevalence and bias effects were not removed as a result of the Cohen kappa coefficient being used in testing of the agreement between the two categorical measurement methods, and different weighting calculations for disagreement cells were ignored. However, the weighted kappa coefficient is widely used to measure agreement between ratings on ordinal scale, and the Pearson, Spearman and Intraclass correlation coefficients are commonly used to assess reliability when ratings are on an interval scale. Both kappa and correlation coefficients can be used to assess the reliability of ordinal rating scales (de Raadt *et al.*, 2021).

Therefore, ordinal data were produced for different correlation values (no relationship, low, medium, and high relationship) between the two raters, and weighted inter-rater agreement statistics were calculated and attempts were made to reveal their status of being effected by the sample size and number of categories. In addition, efforts were made to reveal whether the weighted agreement statistics calculated within the scope of this study are in agreement with the Spearman correlation coefficient.

2. MATERIAL and METHOD

Cohen's kappa, Scott's π, Brennan–Prediger's (B-P), Gwet's AC2, and Krippendorff's alpha are some of the common chance-corrected agreement measures to assess the agreement between two raters for ordinal data (Tran, 2020; Tran *et al.*, 2021; Cohen, 1968; Brennan & Prediger, 1981; Krippendorff, 2004). For these measures, Equation 1 denotes the main formula.

$$\kappa = \frac{P_a - P_e}{1 - P_e} \tag{1}$$

where P_a is the weighted proportion of observed agreement defined as

$$P_a = \sum_{i=1}^{q}\sum_{j=1}^{q} w_{ij} p_{ij} \tag{2}$$

Equation 2, for all agreement coefficients, except for Krippendoff's alpha, Equation 3 gives the weighted proportion of observed agreement for only Krippendorff's alpha (Tran *et al.*, 2020).

$$P_a = \left(1-\frac{1}{n\bar{r}}\right)p_{a0} + \frac{1}{n\bar{r}} \tag{3}$$

Where,

$$P_{a0} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n}\frac{r_{ik}\left(\bar{r}_{ik+}-1\right)}{\bar{r}\left(r_i-1\right)} \tag{4}$$

$$\bar{r} = \frac{1}{n}\sum_{i=1}^{n} r_i \tag{5}$$

$$\bar{r}_{ik+} = \sum_{l=1}^{q} w_{kl} r_{il} \tag{6}$$

r_{ik} is the number of raters who assigned the certain score x_k to subject i and $\bar{r}$ is the average number of raters per subject (Tran *et al.*, 2020). w_{ij} is the symmetrical agreement weight ranging between 0 and 1, for i,j = 1, 2….q, and $w_{ij} = 1$ if i = j. P_e is the weighted proportion of agreement expected by chance, which is different for each weighted inter-rater agreement measures. The formulas of P_e for each weighted inter-rater agreement measure are summarized in Table 3 (Tran *et al.*, 2020; Tran *et al.*, 2021; Cohen, 1968; Brennan & Prediger, 1981; Krippendorff, 2004).

2.1.1. *Types of weighting*

When our outcome variable is categorical (binary/nominal), there is no weighting between categories in the agreement statistics used, i.e., all disagreements are considered equal. However, since there is a ranking between categories in ordinal data, it is necessary to rate the disagreement between the categories that is to be weighted, while calculating the agreement between the raters. In the literature, there are various weighting types, such as linear, quadratic, ordinal, radical, and ratio weights (Tran *et al.*, 2020, Vanbelle & Albert, 2009; Warrens, 2012). Among them, the most commonly used weighting type is the linear weighting proposed by Cicchetti & Allison (1971) and quadratic weighting proposed by Fleiss & Cohen (1973).

2.1.1.1. Linear weights: The formula used to calculate linear weight is as follows:

$$w_{kl} = \begin{cases} 1 - \dfrac{|k-l|}{q-1}, & k \neq l \\ 1\,, & k = l \end{cases} \tag{7}$$

where q is the number of categories and k, l = 1, 2, …, and q are the categories for the first and second rater (Tran *et al.*, 2020).

2.1.1.2. Quadratic weights: The formula used to calculate the quadratic weight is as follows:

$$w_{kl} = \begin{cases} 1 - \dfrac{(k-l)^2}{(q-1)^2}, & k \neq l \\ 1\,, & k = l \end{cases} \tag{8}$$

Quadratic weights are usually greater than linear weights (Tran *et al.*, 2020).

Table 3. Weighted Inter-Rater Agreement Coefficient and Definition

Measures	Definition
Weighted Cohen's Kappa (1968)	$P_e = \sum_{i=1}^{q}\sum_{j=1}^{q} w_{ij} p_{i+} p_{+j}$
Weighted Scott's π (1955)	$P_e = \sum_{i=1}^{q}\sum_{j=1}^{q} w_{ij} \pi_i \pi_j$ Where, $\pi_i = \dfrac{p_{i+} + p_{+i}}{2}$, $\pi\dot{j} = \dfrac{p_{j+} + p_{+j}}{2}$
Weighted Brennan–Prediger's (B-P) (1981)	$P_e = \dfrac{1}{q^2}\sum_{i=1}^{q}\sum_{j=1}^{q} w_{ij}$
Gwet's AC2 (2002)	$P_e = \dfrac{1}{q(q-1)}\left(\sum_{i=1}^{q}\sum_{j=1}^{q} w_{ij}\right)\sum_{i=1}^{q} \pi_i (1-\pi_i)$ Where, $\pi_i = \dfrac{p_{i+} + p_{+i}}{2}$
Krippendorff's Alpha (1970; 1978; 2004)	$P_e = \sum_{k=1}^{q}\sum_{l=1}^{q} w_{kl} \pi_k \pi_l$ Where, $\pi_k = \dfrac{1}{m}\sum_{i=1}^{m} r_{ik} \bar{r}$, $\pi_l = \dfrac{1}{m}\sum_{l=1}^{m} r_{il} \bar{r}$, m is the number of subjects rated by 2 raters.

2.2. Simulation Study

Data for a total of 150 combinations (5×5×3×2), i.e., for 5 different Spearman correlation coefficients (no correlation (0.1), low (0.3), medium (0.5; 0.6), and high (0.8)), 5 different sample sizes (30, 50, 100, 200, and 500), 3 different balanced categories (3, 4, and 5), and 2 different weighting methods (linear and quadratic), were produced and the weighted inter-rater agreement statistics under consideration were obtained with the help of the R package program. All processes were repeated 1000 times.

In addition, dendograms were drawn using the hierarchical clustering method to reveal the comparisons, similarities, and differences of the agreement coefficients addressed in this study. Thirty different dendograms of the agreement statistics were drawn for two different weighting methods, five different Spearman correlation coefficients, and three different categories. Among these dendograms, only the dendograms obtained for the 0.1 correlation coefficient and 3×3 category number belonging to the linear and quadratic weighting methods are discussed in this study and the results obtained from the other dendograms are summarized.

3. RESULTS and DISCUSSION

Linear weighting results of all agreement statistics in terms of different correlation coefficients, sample sizes, and category numbers are outlined in Table 4 and quadratic weighting results are summarized in Table 5.

When Cohen's kappa, Scott's π, and Krippendorff's alpha coefficients were examined, the agreement statistical values obtained according to the quadratic weighting method were equal to the Spearman correlation values, which were not affected by the sample size and the number of categories (Table 5). When the results obtained using the linear weighting method were examined, equal values, without being affected by sample size and number of categories, were obtained only for $\rho = 0.1$, and results obtained for other Spearman correlation values exhibited a deviation of 0.1 (Table 4).

When examined in terms of B-P statistics, it can be observed that the results were not affected by the sample size in any way. On examining the results obtained using the quadratic weighting method, a deviation of 0.1 was observed in the results obtained in the 3×3 category for correlation values ranging from 0.1 to 0.6, while values equal to the correlation value were obtained in all other categories (Table 5). On examining the results obtained using the linear weighting method, when the number of categories was three, a positive deviation of 0.1 was observed for correlation values between 0.1 and 0.5 and a negative deviation of 0.1 was observed for values above 0.5, and when the number of categories was 4, a negative deviation of 0.1 was observed for correlation values of 0.3 and above. Furthermore, when the number of categories was 5, a negative deviation of 0.2 was observed for correlation values of 0.5 and above (Table 4).

When Gwet's AC2 statistic was examined, it was observed that the results were not affected by the sample size. On examining the results obtained using the linear and quadratic weighting methods, when the number of categories was 3, it was observed that an agreement of at least 0.50 was present even when the correlation values were very low, and when the number of categories was 4 or 5, it exhibited similar results with the correlation value (Tables 4 and 5).

In the study, the similarities of the weighted agreement statistics were also examined. Accordingly, on examining the results obtained using the linear weighting method, it was found that the B-P and Gwet's AC2 agreement statistics values were similar and differed from other agreement statistics for all correlation values when the number of categories was three. When the number of categories was four or five, it was noted that only Gwet's AC2 agreement statistic differed from all other agreement statistics. On examining the results using the quadratic weighting method, it has been determined that the Gwet's AC2 agreement statistical value differed from other agreement statistics, for all correlation values when the number of categories was three or four, and that this difference was even more apparent when the number of categories was four. When the number of categories was five, it was observed that all agreement statistics had a similar structure. In Figure 1and 2, the dendograms of the agreement statistics of the results obtained were displayed according to the linear and quadratic weighting methods only for the correlation value of 0.1 and the number of categories of 3.

Table 4. Linear weighting results of agreement statistics in terms of correlation coefficients, sample sizes, and number of categories.

Linear		Cohen's Kappa			Scott's π			Brennan Prediger (B-P)			Gwet's AC2			Krippendorff's Alpha		
ρ	n	3X3	4X4	5X5	3X3	4X4	5X5	3X3	4X4	5X5	3X3	4X4	5X5	3X3	4X4	5X5
0.1	30	0.080	0.071	0.062	0.053	0.044	0.035	0.245	0.087	0.060	0.406	0.180	0.090	0.069	0.060	0.051
	50	0.085	0.076	0.067	0.063	0.056	0.046	0.254	0.096	0.065	0.416	0.185	0.087	0.073	0.065	0.055
	100	0.078	0.070	0.061	0.061	0.053	0.044	0.252	0.091	0.061	0.415	0.178	0.076	0.065	0.058	0.049
	200	0.076	0.069	0.062	0.061	0.056	0.048	0.251	0.092	0.062	0.416	0.178	0.075	0.063	0.058	0.051
	500	0.079	0.071	0.064	0.066	0.059	0.051	0.253	0.093	0.063	0.418	0.178	0.073	0.067	0.060	0.052
0.3	30	0.231	0.208	0.188	0.209	0.186	0.166	0.372	0.226	0.191	0.505	0.305	0.218	0.222	0.200	0.180
	50	0.238	0.218	0.197	0.221	0.202	0.180	0.382	0.237	0.198	0.515	0.313	0.217	0.229	0.210	0.188
	100	0.232	0.211	0.191	0.218	0.198	0.177	0.379	0.231	0.192	0.514	0.305	0.205	0.222	0.202	0.181
	200	0.231	0.212	0.193	0.219	0.201	0.181	0.378	0.231	0.194	0.514	0.305	0.204	0.221	0.203	0.183
	500	0.235	0.215	0.195	0.224	0.205	0.184	0.380	0.234	0.195	0.516	0.305	0.204	0.225	0.206	0.185
0.5	30	0.393	0.358	0.328	0.377	0.342	0.311	0.509	0.377	0.335	0.613	0.442	0.358	0.388	0.353	0.322
	50	0.397	0.367	0.337	0.384	0.355	0.323	0.512	0.384	0.340	0.617	0.466	0.356	0.391	0.361	0.330
	100	0.394	0.362	0.332	0.384	0.352	0.321	0.512	0.379	0.345	0.619	0.440	0.346	0.387	0.355	0.325
	200	0.394	0.364	0.336	0.384	0.354	0.326	0.509	0.380	0.337	0.617	0.439	0.346	0.385	0.356	0.328
	500	0.396	0.366	0.338	0.387	0.358	0.329	0.511	0.381	0.338	0.618	0.439	0.345	0.388	0.359	0.330
0.6	30	0.478	0.438	0.406	0.465	0.425	0.391	0.578	0.456	0.413	0.667	0.512	0.434	0.474	0.435	0.401
	50	0.480	0.447	0.413	0.469	0.436	0.402	0.581	0.463	0.417	0.671	0.518	0.432	0.474	0.442	0.408
	100	0.478	0.443	0.410	0.469	0.434	0.400	0.579	0.459	0.413	0.672	0.512	0.423	0.471	0.437	0.403
	200	0.477	0.444	0.413	0.469	0.437	0.404	0.577	0.459	0.414	0.670	0.510	0.422	0.470	0.438	0.406
	500	0.480	0.447	0.414	0.472	0.439	0.407	0.579	0.460	0.415	0.671	0.510	0.421	0.473	0.440	0.407
0.8	30	0.652	0.620	0.584	0.644	0.612	0.575	0.720	0.635	0.592	0.779	0.673	0.607	0.650	0.618	0.582
	50	0.657	0.624	0.589	0.650	0.618	0.581	0.725	0.637	0.593	0.783	0.674	0.605	0.654	0.621	0.585
	100	0.656	0.622	0.587	0.650	0.617	0.581	0.724	0.634	0.590	0.784	0.669	0.597	0.652	0.618	0.583
	200	0.656	0.623	0.591	0.651	0.618	0.585	0.722	0.634	0.592	0.783	0.668	0.598	0.652	0.619	0.586
	500	0.659	0.626	0.592	0.654	0.621	0.587	0.724	0.635	0.592	0.785	0.669	0.597	0.655	0.621	0.587

Table 5. Quadratic weighting results of agreement statistics in terms of correlation coefficients, sample sizes, and number of categories

Quadratic		Cohen's Kappa			Scott's π			Brennan Prediger (B-P)			Gwet's AC2			Krippendorff's Alpha		
ρ	n	3X3	4X4	5X5	3X3	4X4	5X5	3X3	4X4	5X5	3X3	4X4	5X5	3X3	4X4	5X5
0.1	30	0.095	0.094	0.092	0.065	0.063	0.059	0.228	0.071	0.086	0.460	0.232	0.143	0.081	0.079	0.074
	50	0.103	0.104	0.101	0.078	0.079	0.074	0.239	0.086	0.098	0.472	0.238	0.139	0.087	0.088	0.084
	100	0.096	0.097	0.094	0.076	0.076	0.071	0.240	0.081	0.095	0.474	0.231	0.123	0.080	0.080	0.075
	200	0.093	0.095	0.096	0.074	0.076	0.075	0.238	0.080	0.096	0.475	0.230	0.119	0.077	0.079	0.077
	500	0.096	0.098	0.097	0.079	0.081	0.078	0.239	0.082	0.097	0.476	0.229	0.116	0.080	0.082	0.079
0.3	30	0.280	0.280	0.279	0.257	0.257	0.254	0.391	0.267	0.282	0.572	0.394	0.328	0.269	0.269	0.267
	50	0.288	0.292	0.290	0.269	0.273	0.269	0.401	0.281	0.290	0.583	0.402	0.324	0.277	0.280	0.276
	100	0.283	0.285	0.283	0.267	0.269	0.265	0.400	0.275	0.286	0.585	0.395	0.309	0.271	0.273	0.269
	200	0.281	0.285	0.286	0.266	0.270	0.270	0.397	0.274	0.287	0.584	0.392	0.305	0.268	0.272	0.272
	500	0.285	0.289	0.289	0.272	0.276	0.274	0.399	0.277	0.289	0.586	0.393	0.304	0.272	0.277	0.275
0.5	30	0.469	0.470	0.470	0.453	0.454	0.452	0.554	0.466	0.478	0.687	0.559	0.512	0.462	0.463	0.461
	50	0.473	0.479	0.477	0.460	0.466	0.463	0.559	0.473	0.481	0.692	0.562	0.506	0.465	0.471	0.468
	100	0.473	0.476	0.474	0.462	0.465	0.461	0.561	0.471	0.477	0.697	0.559	0.495	0.465	0.468	0.464
	200	0.471	0.476	0.478	0.460	0.466	0.466	0.557	0.469	0.479	0.694	0.555	0.493	0.462	0.467	0.468
	500	0.474	0.480	0.481	0.464	0.470	0.470	0.558	0.472	0.481	0.695	0.556	0.491	0.465	0.471	0.470
0.6	30	0.565	0.566	0.566	0.553	0.554	0.552	0.637	0.564	0.575	0.744	0.640	0.603	0.560	0.561	0.560
	50	0.568	0.574	0.572	0.557	0.563	0.560	0.640	0.570	0.576	0.749	0.644	0.597	0.562	0.568	0.565
	100	0.568	0.572	0.570	0.559	0.563	0.560	0.640	0.569	0.574	0.751	0.640	0.589	0.561	0.565	0.562
	200	0.566	0.573	0.574	0.557	0.565	0.564	0.637	0.568	0.575	0.749	0.638	0.586	0.558	0.566	0.565
	500	0.569	0.576	0.576	0.561	0.568	0.567	0.638	0.569	0.576	0.750	0.638	0.585	0.561	0.568	0.567
0.8	30	0.749	0.762	0.759	0.743	0.755	0.752	0.793	0.764	0.766	0.854	0.805	0.783	0.747	0.760	0.756
	50	0.755	0.764	0.762	0.750	0.758	0.756	0.798	0.764	0.767	0.859	0.804	0.779	0.752	0.761	0.758
	100	0.755	0.764	0.763	0.750	0.759	0.757	0.797	0.763	0.765	0.860	0.802	0.774	0.752	0.760	0.758
	200	0.755	0.765	0.766	0.750	0.761	0.761	0.796	0.763	0.767	0.859	0.802	0.774	0.751	0.761	0.762
	500	0.759	0.767	0.768	0.754	0.763	0.763	0.797	0.764	0.768	0.860	0.802	0.773	0.754	0.763	0.763

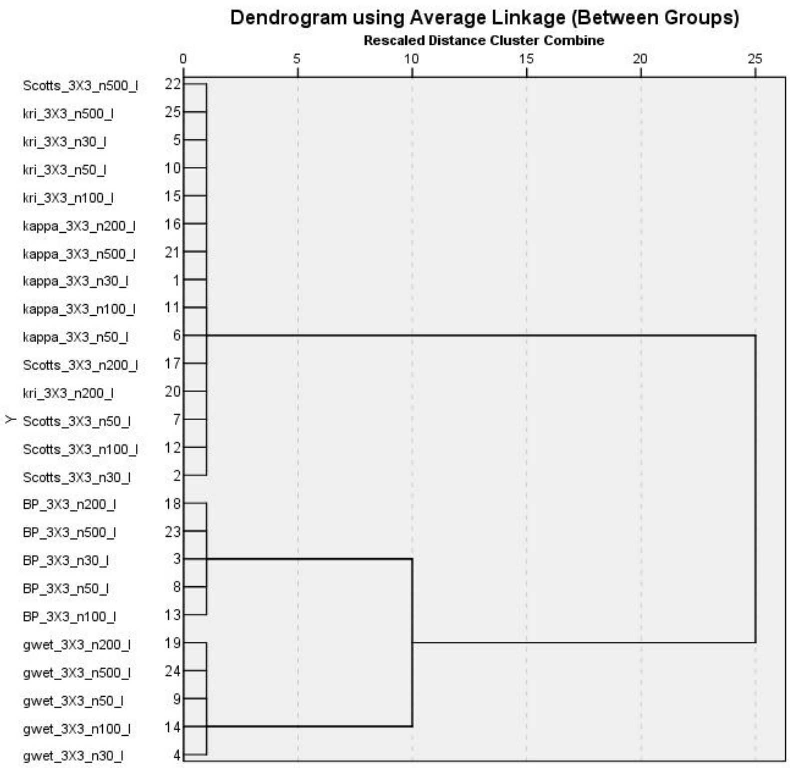

Figure 1. Linear weighting dendograms of all agreement statistics for ρ = 0.1 and 3×3 category number

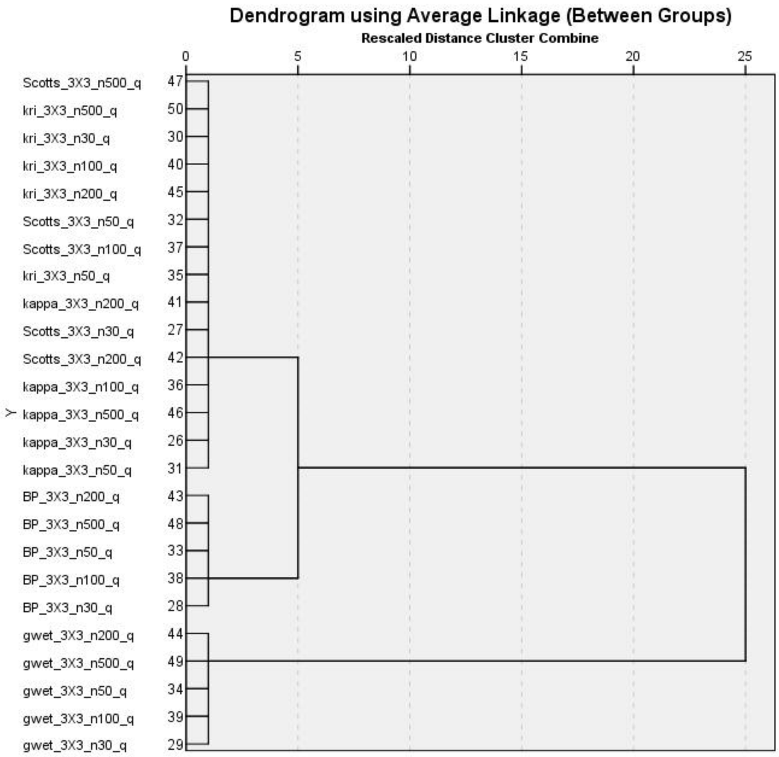

Figure 2. Quadratic weighting dendograms of all agreement statistics for ρ = 0.1 and 3×3 category

In our study, we attempted to reveal how the agreement statistics, which are used in ordinal data and in cases when there are two different raters, are affected by the sample size and the number of categories. The weighting of the agreement statistics used in the ordinal data was calculated using linear and quadratic weighting formulations. In the literature, there are various weighting types, such as linear, quadratic, ordinal, radical, and ratio weights. Linear, quadratic, radical, and ratio weightings are calculated over the scores in the ordinal categories. Among those weights, only the calculation of ordinal weights is done with ranks instead of actual values. Therefore, it is insensitive to the choices of scores (Tran *et al.*, 2020). In Warrens's (2012) study, presented several hypothetical examples in her study to show that the quadratic weighted kappa cannot distinguish between agreement tables with very different exact agreement values. Moreover, he states that the quadratic weighted kappa failure as a measure of agreement. Therefore, it recommends using linear weighted kappa instead of quadratic weighted kappa when a single agreement index needs to be reported for an ordinal scale (Warrens, 2012). Vanbella (2016) compared linear and quadratic kappa coefficients. Ideally, both weighted kappa coefficients should be reported. However, if a single coefficient of fit is to be used, in such a case, the linear kappa coefficient is more appropriate. Because this coefficient gives information about the distribution of the disagreement (Vanbella, 2016). Moreover, Vanbelle *et al.* (2016) and Warrens (2011) showed that linearly weighted kappa can be interpreted as a weighted average of the 2×2 tables' kappa's. Moradzadeh *et al.* (2017) showed that linear and quadratic weighted kappa can be computed as a function of unweighted kappa's. In our study, we only considered the agreement statistics in terms of linear and quadratic weighting methods. In Cohen's kappa, Scott's π, and Krippendorff's alpha coefficients, when calculated according to the linear weighting method, there were deviations compared to the correlation values, whereas values equal to the correlation value were obtained for all category numbers when calculated according to the quadratic weighting method. On examining the BP agreement statistics, when calculated according to the linear weighting method and when the number of categories is three, a deviation of 0.1 was observed in the positive direction, when there was no or very weak relationship between the raters, and in the negative direction, when there was high relationship. When the number of categories was four, a negative deviation of 0.1 was observed for correlation values of 0.3 and above. When the number of categories was five, a negative deviation of 0.2 was observed for values with a correlation value of 0.5 and above. When calculated according to the quadratic weighting method, a positive deviation of 0.1 was observed in cases where there was no or very weak correlation between the raters only when the number of categories was 3, while it was observed that it provided values equal to the correlation value in all other combinations. When the Gwet's AC2 agreement statistic was examined, when the number of categories was three, a moderate agreement was observed to be present in all cases when there is no or weak correlation between the raters, whether the calculation was made according to the linear or quadratic weighting method. When the number of categories was four or five, it has been determined that the results obtained from the quadratic weighting method had values closer to the correlation coefficient value compared to the results obtained from the linear weighting method.

In the literature, the concepts of relationship and agreement are widely used interchangeably. Especially in agreement studies, the use of correlation coefficients is frequently observed. Based on this, data for different correlation values between the two raters were produced and examined for whether the agreement statistics provide similar results in these cases. Agreement and correlation are used to indicate the strength of association between variables, but they are conceptive different and, hence, require the use of different statistical methods. If it is assumed that the variables measure the same structure, the agreement between these variables is assumed to measure different structures, but the relationship between these variables is also investigated. This important difference requires the use of different statistical methods. Statistical methods to be used vary depending on the distribution of the data. Therefore, it is necessary to be very careful when assessing agreement and correlation.

The presence of agreement suggests the existence of the relationship, but the reverse may not be true, i.e., if there is agreement between two variables, it means that there is a relationship already, but there may be a strong relationship even if there is no strong agreement between the results (Liu *et al.*, 2016). de Raadt *et al.* (2021) compared the correlation coefficients (Pearson, Spearman, and ICC) used in ordinal scales with the linear and quadratic kappa coefficients. According to the simulation results, the correlation coefficient values were shown to increase when there is high agreement between the two raters. Also, the difference between the correlation coefficient values is small only when the difference between the raters is small. Also, quadratic kappa is highly correlated with all correlation coefficients. If quadratic kappa is used instead of correlation coefficients, inter-rater reliability coefficients are inevitable that the results will be similar. In our simulation results, it can be said that Cohen’s kappa, Scott’s π, and Krippendorff’s alpha coefficients provided similar results with the inter-rater correlation coefficients, and values very close to the correlation coefficient value were obtained in the B-P agreement statistics. When the agreement between the two raters was examined, it can be said that the agreement statistics mentioned above can be used interchangeably, since they provided similar results to the Spearman correlation coefficient. However, as an agreement statistic, Gwet’s AC2 statistic differs from other agreement statistics. Particularly in cases when there are two raters, the number of categories is three, and there is no or low correlation between raters, a moderate level of agreement can be spoken of, albeit by chance.

Tran’s *et al.* (2020) compared weighted agreement statistics for the cases of balanced and unbalanced number of categories. According to their simulation results, all agreement statistics provided similar results in balanced cases, whereas Gwet’s AC2 and B-P agreement statistics provided better results in unbalanced cases. In addition, in cases when the inter-rater agreement was low, Cohen’s kappa, Scott’s π, and Krippendorff’s alpha coefficients performed better than Gwet’s AC2 and B-P agreement statistics, and B-P agreement statistics outperformed other agreement statistics when the inter-rater agreement was medium or high. In our simulation results, it can be said that Gwet’s AC2 agreement statistic differs from other agreement statistics when the relationship between raters was low and the number of categories was balanced and three, while in all the other cases, the correlation coefficient can be used instead of the agreement statistics.

4. CONCLUSION

In this study, when investigating the agreement between two raters in ordinal scales, it is recommended to be careful only in cases when the number of categories is three and to use Gwet’s AC2 agreement statistic as an agreement statistic in such cases. In other cases, it can be said that the concepts of agreement and relationship can be used interchangeably and easily. In addition, the Spearman correlation coefficient is available in all statistical package programs and is able to be implemented easily emerges as an alternative method for inter-rater agreement studies. It can also be said that the agreement statistics are not affected by the sample size.

Conflict of interest statement: There is no funding.

REFERENCES

Barnhart H.X., Haber M.J. & Lin L.I. (2007). An overview on assessing agreement with continuous measurements. Journal of Biopharmaceutical Statistics **17**(4):529-569. DOI: https://doi.org/10.1080/10543400701376480.

Bland J.M. & Altman D.G. (2010). Statistical methods for assessing agreement between two methods of clinical measurement. International Journal of Nursing Studies **47**(8):931-936.DOI: https://doi.org/10.1016/j.ijnurstu.2009.10.001.

Brennan R.L. & Prediger D.J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. Educational and Psychological Measurement **41**(3):687-699. DOI: https://doi.org/10.1177/001316448104100307.

Cicchetti D. & Allison T. (1971). A new procedure for assessing reliability of scoring EEG sleep recordings. The American Journal of EEG Technology **11**(3):101-109. DOI: https://doi.org/10.1080/00029238.1971.11080840.

Cohen J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin **70**(4):213-220. DOI: https://doi.org/10.1037/h0026256.

Fleiss J.L. & Cohen J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement **33**(3):613–619. DOI: https://doi.org/10.1177/001316447303300309

Gwet K.L. (2014). Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. 4 th edition, pp. 27-127, 185-302. Advanced Analytic, LLC, Gaithersburg, USA.

Gwet K.L. (2015.) Testing the difference of correlated agreement coefficients for statistical significance. Educational and Psychological Measurement **76**(4):609-637. DOI: 10.1177/0013164415596420.

Haber M. & Barnhart H.X. (2008). A general approach to evaluating agreement between two observers or methods of measurement. Statistical Methods in Medical Research **17**(2):151-169. DOI: 10.1177/0962280206075527.

Haber M., Barnhart H.X., Song J. & Gruden J (2005). Observer variability: A new approach in evaluating interobserver agreement. Journal of Data Science **3**(1):69-83.

Kanık E.A., Erdogan S. & Temel G.O. (2012). Agreement statistics impacts of prevalence between the two clinicians in binary diagnostic tests. Annals of Medical Research **19**(3):153-158. DOI: 10.7247/jiumf.19.3.5.

Kanık E.A., Orekici Temel G. & Ersöz Kaya I. (2010). Effect of sample size, the number of raters and the category levels of diagnostic test on Krippendorff alpha and the Fleiss kappa statistics for calculating inter rater agreement: A simulation study. Türkiye Klinikleri Journal of Biostatistics **2**(2):74-81. DOI:10.7247/jtomc.19.4.4.

Krippendorff K. (2004). Measuring the reliability of qualitative text analysis data. Quality and Quantity **38**(6):787-800. DOI: http://dx.doi.org/10.1007/s11135-004-8107-7.

Lin L. (2008). Overview of agreement statistics for medical devices. Journal of Biopharmaceutical Statistics **18**(1):126-144. DOI: 10.1080/10543400701668290.

Lin L., Hedayat A.S. & Wu W. (2007). A Unified approach for assessing agreement for continuous and categorical data. Journal of Biopharmaceutical Statistics **17**(4):629-652.

Lin L., Hedayet A.S. & Wu W. (2012). Statistical tools for measuring agreement. 1st edition, pp. 1-109. Springer, New York.

Liu J., Tang W., Chen G., Lu Y., Feng C. & Tu X.M. (2016). Correlation and agreement: overview and clarification of competing concepts and measures. Shanghai Archives of Psychiatry **28**(2):115-120. DOI: 10.11919/j.issn.1002-0829.216045.

Moradzadeh N., Ganjali M. & Baghfalaki T. (2017). Weighted Kappa as a function of unweighted kappas. Communications in Statistics-Simulation and Computation **46**(5):3769-3780. DOI:10.1080/03610918.2015.1105975

Nelson K.P. & Edwards D. (2018). A measure of association for ordered categorical data in population-based studies. Statistical Methods in Medical Research **27**(3):812-831. DOI: 10.1177/0962280216643347.

Raadt A., Warrens M., Bosker R. & Kiers H.A.L. (2021). A comparison of reliability coefficient for ordinal rating scales. Journal of Classification:1-25. DOI: https://doi.org/10.1007/s00357-021-09386-5.

Stralen K.J., Dekker F.W., Zoccali C. & Jager K.J. (2012). Measuring agreement, more complicated than it seems. Nephron Clinical Practice **120**(3):c162-c167. DOI: 10.1159/000337798.

Tran D., Dolgun A. & Demirhan H. (2020.) Weighted inter-rater agreement measures for ordinal outcomes. Communications in Statistics-Simulation and Computation **49**(4):989-1003. DOI: https://doi.org/10.1080/03610918.2018.1490428.

Tran Q.D., Dolgun A. & Demirhan H. (2021). The impact of gray zones on the accuracy of agreement measures for ordinal tables. BMC Medical Research Methodology **21**(1):1-9. DOI:10.1186/s12874-021-01248-3.

Vanbella S. (2016). A new interpretation of the weighted kappa coefficients. Psychometrika **81**(2):399-410. DOI: 10.1007/s11336-014-9439-4.

Vanbelle S. & Albert A. (2009). A note on the linearly weighted kappa coefficient for ordinal scales. Statistical Methodology **6**(2):157-163. DOI: https://doi.org/10.1016/j.stamet.2008.06.001.

Warrens M.J. (2012). Some paradoxical results for the quadratically weighted kappa. Psychometrika **77**(2):315-323. DOI: 10.1007/S11336-012-9258-4.