Chilean Journal of Statistics
Volume 13, Number 2 (December 2022), Pages 165–186
DOI: 10.32372/chjs.13-02-03

Contingency Tables
Research Paper

Inter-rater agreement and adjusted overall degree of distinguishability for ordered categories

Ayfer Ezgi Yilmaz*

Department of Statistics, Hacettepe University, Ankara,Turkey

(Received: 29 January 2021 · Accepted in final form: 21 December 2021)

Abstract

In square contingency tables, weighted kappa and AC2 coefficients are used to summarize the degree of agreement between raters of an ordered square contingency table. In addition to investigate the agreement between raters, category distinguishability should be considered to check the reliability of the study. The overall degree of distinguishability is used for $R \times R$ tables. In some applications, the value of overall degree of distinguishability is calculated outside the defined range as negative values. Since overall degree of distinguishability is calculated by using all the category pairs, there occurs inflation on its value. In this study, adjusted overall degree of distinguishability is suggested to solve these two problems. Furthermore, interpretation of category distinguishability is outlined and benchmark scales for overall degree of distinguishability are developed. A simulation study is performed to compare the accuracy of the adjusted overall degree of distinguishability and the classical one. The tables to find the adjusted overall degrees of distinguishability for certain values of agreement coefficients are generated. The results are discussed over medical data sets.

Keywords: AC2 coefficient · Category distinguishability · Polynomial regression · Square contingency tables · Weighted kappa coefficient

Mathematics Subject Classification: 62H17.

1. Introduction

Categorical variables reflect the characteristics of the experimental unit such as gender, marital status, the severity of a disease, departments of a faculty, brand preference of the customers (Vélez and Marmolejo-Ramos, 2017). There are two types of categorical variables, nominal and ordinal. Ordinal variables are commonly used in the health sciences. The tables with ordered categories arise if two raters rate the same subjects or one rater rates the same subjects in two different time points. For these tables, analysis of agreement between the row and column classifications is important. Cohen's weighted kappa and AC2 coefficients are used to determine the level of agreement between the ordinal classifications of two raters (Cohen, 1968; Gwet, 2012).

Even though raters (observers) rate the subjects independently, a correlation occurs between their decisions. Two main components of the agreement are expressed as the marginal

*Corresponding author. Email: ezgiyilmaz@hacettepe.edu.tr

ISSN: 0718-7912 (print)/ISSN 0718-7920 (online)
© Chilean Statistical Society – Sociedad Chilena de Estadística
http://www.soche.cl/chjs

homogeneity and the category distinguishability. While marginal homogeneity represents the differences in the marginal distributions of raters, category distinguishability is the ability of raters to make a distinction between two categories (Darroch and McCloud, 1986; Valet and Mary, 2011).

In the agreement studies, it is necessary to determine if the ordered categories are distinguishable from one to another (Perkins and Becker, 2002). If the categories are indistinguishable, raters' perceptions differ. The main reasons for indistinguishability are the definition of categories and expertise of the rater. If the definition of categories is not clear, then different raters get them differently or the same rater does not distinguish the categories correctly. If raters are non-expert in their fields, then it may be difficult for them to distinguish the categories. The measure to calculate the distinguishability level is called degree of distinguishability (DD) (Darroch and McCloud, 1986).

Darroch and McCloud (1986) state that the values of DD range between 0 which refers to indistinguishability and 1 which refers to perfect distinguishability. In some applications, the value of DD is calculated outside the defined range as negative values. For example, the data used in Oh (2009) includes classifications of two trauma surgeons and two radiologists of 60 patients into four categories. For this example, DD between 2–3 and 3–4 categories are calculated as -0.43 and -0.67, respectively. Because the minimum value of DD is defined as negative, there is uncertainty about its interpretation. To solve this problem for 2×2 tables, Yilmaz and Saracbasi (2019) suggest to use the adjusted degree of distinguishability (ADD). In this article, the conditions where negative DD values occur are theoretically examined in Section 2.2.

For the ordinal tables, DD is calculated for each 2×2 sub-table. The overall degree of distinguishability (ODD) is used to summarize the distinguishability of a table by a single value. Because ODD is an average of degrees of distinguishability calculated for all pairs, these negative values affect the value of ODD, as well. In this article, we suggest the adjusted overall degree of distinguishability (AODD) for $R \times R$ tables to overcome this problem. Furthermore, to the best of our knowledge, there is no information about the interpretation of ODD. It is aimed to assess the weighted kappa coefficient and AODD in square contingency tables together. A simulation study is performed to compare the accuracy of AODD with the classical one. Then, polynomial regression is used to model the weighted kappa coefficient, AC2, and AODD. Furthermore, benchmark scales for AODD are determined based on a simulation study. Inter-rater agreement coefficients and category distinguishability are reviewed in Section 2. The simulation study and the illustrative examples results are discussed in Section 3.

2. Inter-rater Agreement and DD

In this section we review the inter-rater agreement and the DD.

2.1 Inter-rater Agreement

There are numerous agreement coefficients for each table structure (nominal, ordinal or interval). The well-known agreement coefficients for ordinal categories are Cohen (1968) weighted kappa coefficient and Gwet (2012) AC2 coefficient. These coefficients are used to assess the agreement between ordinal classifications of two raters.

Consider two raters who classify the subjects into an R ordinal scale. Let n_{ij} denote the number of subjects placed in category i by the first rater and in category j by the second rater $(i, j = 1, 2, \ldots, R)$. The cell probabilities are $p_{ij} = n_{ij}/n$. The marginal row and column probabilities are $p_{i.} = n_{i.}/n$ and $p_{.j} = n_{.j}/n$, respectively. The observed agreement

and the proportion agreement expected by chance for kappa are

$$P_0 = \sum_{i=1}^{R}\sum_{j=1}^{R} w_{ij}p_{ij} \quad \text{and} \quad P_{e,\kappa} = \sum_{i=1}^{R}\sum_{j=1}^{R} w_{ij}p_{i.}p_{.j}. \tag{2.1}$$

The weighted kappa coefficient κ_w is

$$\kappa_w = \frac{P_0 - P_{e,\kappa}}{1 - P_{e,\kappa}}.$$

Since the proportion agreement expected by chance for AC2

$$P_{e,AC2} = \frac{w_T}{R(R-1)} \sum_{i=1}^{R} \pi_i(1-\pi_i), \tag{2.2}$$

where $w_T = \sum_{i=1}^{R}\sum_{j=1}^{R} w_{ij}$ and $\pi_i = (p_{i.} + p_{.i})/2$. The AC2 coefficient is given by

$$AC2 = \frac{P_0 - P_{e,AC2}}{1 - P_{e,AC2}},$$

where P_0 is defined in Equation (2.1). The w_{ij}'s in Equation (2.1) and (2.2) are the weights, $0 \leq w_{ij} \leq 1$. The coefficient allows each (i, j) cell to be weighted according to the degree of agreement between ith and jth categories (Shoukri, 2004). Different formulations of weighting schemes are compared by Tran et al. (2020) and it is concluded the accuracy of the coefficients is not sensitive to the used weights if the table of interest is not highly unbalanced and the true agreement is not that low. Thus, we only considered the linear weights $w_{ij} = 1 - |i - j|/(R - 1)$ of Cicchetti and Allison (1971).

2.2 Category Distinguishabilities

In the agreement studies with ordered scales, besides the level of agreement, category distinguishability is of interest (Becker, 1989). Category distinguishability is the ability of the raters to distinguish the categories. DD is used to measure the degree to which raters can distinguish between categories. In such studies, it is expected to have as many distinguishable categories as possible because indistinguishable categories may affect the level of agreement. If the definition of two categories is not clear or rater is not expert on his/her field, and raters cannot distinguish these categories well, then it would be difficult to categorize the subjects into a correct category. In that case, the choices they make between these two categories are expected by chance and these random classifications cause poor agreement.

Darroch and McCloud (1986) argue that kappa coefficient is not satisfactory as a measure of category distinguishability; and instead of kappa coefficient, the measure of DD can be used.

DD is suggested to investigate the ability of the raters to distinguish between two categories (Darroch and McCloud, 1986). Let any 2×2 sub-table of an $R \times R$ agreement table is shown in Table 1.

Table 1. Notation for a 2×2 sub-table.

Rater 1	Rater 2 i	Rater 2 j	Total
i	n_{ii}	n_{ij}	$n_{i.}$
j	n_{ji}	n_{jj}	$n_{j.}$
Total	$n_{.i}$	$n_{.j}$	

The measure of category distinguishability is defined in terms of the following odds ratio.

$$\tau_{ij} = \frac{n_{ii}n_{jj}}{n_{ij}n_{ji}}, \quad i < j. \tag{2.3}$$

The DD of ith and jth categories is

$$\delta_{ij} = 1 - \tau_{ij}^{-1}, \tag{2.4}$$

where $0 \leq \delta_{ij} \leq 1$. When $\delta_{ij} \cong 1$, then there is a perfect distinguishability between these two categories. When $\delta_{ij} \cong 0$, then it is impossible to distinguish between these two categories and this is not a preferred situation in the studies.

The odds ratio measures the differences in the interpretation of two categories by two raters. If the odds ratio is equal to one as $(n_{ii}n_{jj} = n_{ij}n_{ji})$, then $\delta_{ij} = 0$ which represents the indistinguishable categories. In that case, the decision of the raters is effectively made by the toss of a suitably weighted coin (Darroch and McCloud, 1986). If the odds ratio is high enough as $(n_{ii}n_{jj} > n_{ij}n_{ji})$, then $\delta_{ij} \cong 1$ which represents the perfect distinguishable categories. In that case, the raters make their choices clearly, not by chance.

The ODD is suggested to summarize the category distinguishability of an $R \times R$ table by a single value. As δ_{ij} is defined as in Equation (2.4), the ODD is given in Equation (2.5).

$$\delta = \frac{1}{\binom{R}{2}} \sum_{i<j} \delta_{ij}, \tag{2.5}$$

where $0 \leq \delta \leq 1$. $\delta \cong 1$ if and only if all pairs of categories are completely distinguishable, and $\delta \cong 0$ if and only if all pairs of categories are completely indistinguishable.

The ranges of DD and ODD are defined between 0 and 1. In the applications, the value of DD may be calculated outside the defined range as negative. To solve this problem, Yilmaz and Saracbasi (2019) suggest to use ADD. The ADD between ith and $(i+1)$st categories is

$$\delta^a_{i(i+1)} = \begin{cases} 1 - \tau^{-1}_{i(i+1)} & \text{if } \tau_{i(i+1)} \geq 1, \\ 1 - \tau_{i(i+1)} & \text{if } \tau_{i(i+1)} < 1, \end{cases} \tag{2.6}$$

where $0 \leq \delta^a_{i(i+1)} \leq 1$, $i = 1, 2, \ldots, (R-1)$. The odds ratio for square contingency tables is

$$\tau_{i(i+1)} = \frac{n_{ii}\ n_{(i+1)(i+1)}}{n_{i(i+1)}\ n_{(i+1)i}}.$$

If the table contains sampling zeros, the odds ratio either equal to 0 or ∞. Hence, the degree of distinguishability is calculated as $-\infty$ or 1, respectively. To solve this problem, the odds

ratio is calculated by adding 0.50 to each cell (Agresti, 2002)

$$\tau_{i(i+1)} = \frac{(n_{ii} + 0.5)(n_{(i+1)(i+1)} + 0.5)}{(n_{i(i+1)} + 0.5)(n_{(i+1)i} + 0.5)}.$$

To get the inequality for the sub-tables with negative distinguishabilities ($\delta_{ij} < 0$), Equation (2.3) and (2.4) can be used

$$\begin{aligned}\delta_{ij} = 1 - \frac{n_{ij}n_{ji}}{n_{ii}n_{jj}} &= \frac{n_{ii}n_{jj} - n_{ij}n_{ji}}{n_{ii}n_{jj}} < 0 \\ &= n_{ii}n_{jj} - n_{ij}n_{ji} < 0.\end{aligned}$$

Then, DD is negative only if $n_{ii}n_{jj} < n_{ij}n_{ji}$. By considering Table 1, let $A = n_{ii} + n_{jj}$ show the number of subjects that the two raters agree and $D = n_{ij} + n_{ji}$ show the number of subjects that the two raters disagree.

In this study, when DD is negative, we conduct a more detailed investigation of the frequency distribution of sub-tables. We consider three cases: (a) raters agree more than disagree; (b) raters disagree more than agree; (c) equal agreement and disagreement (Figure 1).

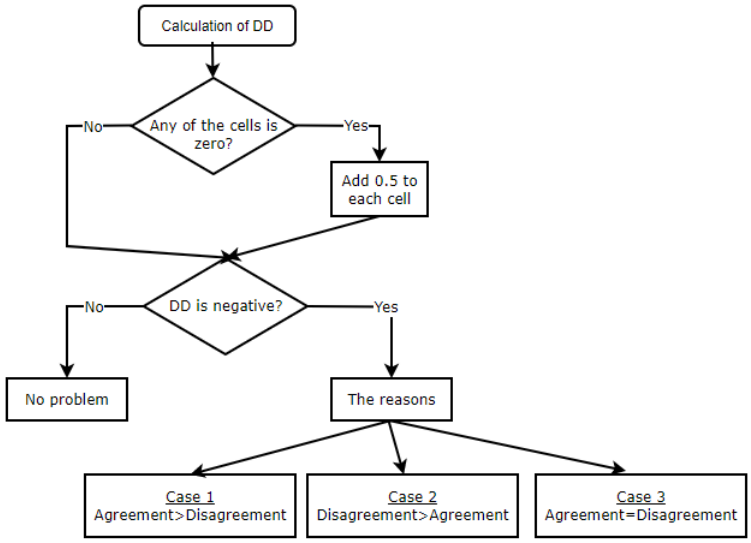

Figure 1. A flowchart showing the problems of DD calculations.

Case 1: Let there is negative distinguishability and the raters agree more than disagree. To solve the compound inequalities in Case 1, we need to consider the followings:

(1) When DD is negative, then ($n_{ii}n_{jj} < n_{ij}n_{ji}$).
(2) When the raters agree than disagree, then ($A > D$).
(3) When any cell frequency is zeros, 0.5 is added to all the cells. Thus, $n_{ii} > 0$, $n_{jj} > 0$, $n_{ij} > 0$, and $n_{ji} > 0$.

By considering these three inequalities above, the possible solutions are

- If $D > n_{ii}$, then $-n_{ii} + D < n_{jj} < n_{ij}n_{ji}/n_{ii}$.
- If $D < n_{ii}$, then $0 < n_{jj} < n_{ij}n_{ji}/n_{ii}$.

Case 2: Let there is negative distinguishability and the raters disagree more than agree. To solve the compound inequalities in Case 2, we need to consider the followings

(1) When DD is negative, then $n_{ii}n_{jj} < n_{ij}n_{ji}$.
(2) When the raters disagree than agree, then $(A < D)$.
(3) $n_{ii} > 0$, $n_{jj} > 0$, $n_{ij} > 0$, and $n_{ji} > 0$.

By considering these three inequalities above, the possible solutions are

- If $n_{ij} = n_{ii}$, then $0 < n_{jj} < n_{ji}$.
- If $n_{ji} = n_{ii}$, then $0 < n_{jj} < n_{ij}$.
- If $n_{ij} < n_{ii}$ and $n_{ji} < n_{ii}$, then $0 < n_{jj} < D - n_{ii}$.
- If $n_{ij} > n_{ii}$ and $n_{ji} > n_{ii}$, then $0 < n_{jj} < D - n_{ii}$.
- If $n_{ij} > n_{ii}$ and $n_{ii} > n_{ji}$, then $0 < n_{jj} < n_{ij}n_{ji}/n_{ii}$.
- If $n_{ji} > n_{ii}$ and $n_{ii} > n_{ij}$, then $0 < n_{jj} < n_{ij}n_{ji}/n_{ii}$.

Case 3: Let there is negative distinguishability and the number of subjects that two raters agree and the number of subjects that they disagree are equal. To solve the compound inequalities in Case 3, we need to consider the followings

(1) When DD is negative, then $n_{ii}n_{jj} < n_{ij}n_{ji}$.
(2) When there is equal agreement and disagreement, then $(A = D)$.
(3) $n_{ii} > 0$, $n_{jj} > 0$, $n_{ij} > 0$, and $n_{ji} > 0$.

By considering these three conditions above, the only possible solution is

- $n_{ii} < D$ and $n_{jj} = D - n_{ii}$, then $0 < n_{jj} < n_{ij}n_{ji}/n_{ii}$.

The negative values of DD may also affect the value of ODD because it is an average of the distinguishabilities between all possible combinations of the categories. When some of DD values are negative, ODD value is calculated less than its true value. In this paper, we proposed AODD to calculate the ADD.

We proposed AODD under two arguments. Firstly, the distinguishability between ith and jth categories is equal to the distinguishability between jth and ith categories. Secondly, DD only for the adjacent categories instead of all the pairs is used to calculate the overall degree of distinguishable because if categories (i) and $(i + 1)$ are distinguishable and categories $(i + 1)$ and $(i + 2)$ are distinguishable, then it is reasonable if categories (i) and $(i + 2)$ are distinguishable as well. Even though in the original formulation of DD, it is possible to calculate for each pair of categories, in real-life applications it is not preferable. Agresti (1988) discusses that the value of DD increases as the distance between the categories increases and calculated the DD's only for the adjacent categories. Valet et al. (2007) suggest the log-linear non-uniform association models and considered the distinguishability between two adjacent categories. The reason is that the association has an ordinal pattern and when the distance between the categories increases, the association (odds ratio) between them increases. Because the formulation of DD is directly related to the odds ratio, the ability of the raters to distinguish the categories also increases. To give an example, Valet et al. (2009) use the furrows between eyebrows, wrinkles above lips, nasolabial fold (NLF) scale for skin ageing signs (see Table 1 in Valet et al. (2009)). The data has six ordinal categories and the calculated DD values are

Categories	2	3	4	5	6
1	0.99	1.00	1.00	1.00	1.00
2		0.63	0.99	0.99	1.00
3			0.70	1.00	1.00
4				0.93	1.00
5					0.95

The results show that the DD values increase when the difference between the categories. If we calculate the DD values by using all the pairs, we calculate the ODD as 0.94. Because the values of non-adjacent categories are higher than the adjacent ones, the ODD value does not reflect the true distinguishability. However, when we calculate the DD values by using only the adjacent categories we calculate the ODD as 0.84. For this reason, it is sufficient to calculate DD only for the adjacent categories instead of all the pairs. The AODD is

$$\delta^a = \frac{1}{R-1}\sum_{i=1}^{R-1}\delta^a_{i(i+1)}. \tag{2.7}$$

Here, $\delta^a_{i(i+1)}$ can be calculated from Equation (2.6). The benchmark scale of AODD is derived and discussed in Section 3.4.

3. Simulation Study

A simulation study is performed to compare the proposed AODD with the classical one for square contingency tables with ordinal categories. It is also aimed to discuss the equivalence of weighted kappa coefficient and AODD, and to develop benchmarking scales for AODD.

3.1 Design of the Simulation Study

We use the method presented by Goktas and Isci (2011) to generate $R \times R$ contingency tables. We randomly generate two independent standard normal distributions (X and Y) considering the predetermined sample sizes. As ρ is the true correlation between X_1 and X_2 variables, we calculate a and b (see Goktas and Isci (2011) study for more detail).

$$a = \frac{\sqrt{1+\rho}+\sqrt{1-\rho}}{2} \quad \text{and} \quad b = \frac{\sqrt{1+\rho}-\sqrt{1-\rho}}{2}.$$

Then, $X_1 = aX + bY$ and $X_2 = bX + aY$ are generated. To get an $R \times R$ table, the range of generated X_1 and X_2 variables are divided into R sub-equal intervals and the cross-table is generated.

The sample size (n) is taken as 50, 100, 200, and 500. Number of categories (R) is taken as 3, 4, 5, and 6. Spearman's rank correlation coefficient (ρ) is considered as 0.20, 0.50, and 0.80 which refer to low, moderate, or strong relationship among the raters. The linearly weighted kappa coefficient, ODD, and AODD are calculated for each scenario.

The simulation software is developed in R version 3.6.1 by the author on a computer with an Intel® Core^{TM} i7 processor and 8GB RAM, running on the Windows version 8.1. We utilize stepAIC() function of **MASS** package to select the best fitted polynomial regression models (Venables et al, 2002) and **irrCAC** package to calculate kappa and AC2 coefficients (Gwet, 2019). The results are based on 50,000 replications. The computation time for each scenario is on average of about 755 seconds.

3.2 The comparison of overall and adjusted overall degrees of distinguishability

To compare the ODD and AODD, a simulation study is performed. Table 2 shows the minimum, maximum values, median, mean, and standard errors of ODD and AODD for different number of categories, different sample sizes, and different values of correlation.

Table 2. The descriptive statistics of ODD and AODD.

			δ					δ^a				
ρ	R	n	Min	Med	Max	Mean	S.E.	Min	Med	Max	Mean	S.E.
0.20	3	50	-4.14	0.4820	0.99	0.3591	0.0021	0.00	0.5744	0.99	0.5676	0.0009
		100	-4.01	0.4230	0.96	0.2551	0.0024	0.00	0.5484	0.96	0.5402	0.0008
		200	-6.14	0.4215	0.92	0.2524	0.0025	0.01	0.5054	0.92	0.4986	0.0007
		500	-7.69	0.4338	0.83	0.3617	0.0015	0.00	0.4559	0.83	0.4482	0.0006
	4	50	-58.09	0.0213	0.94	-0.3168	0.0056	0.00	0.4988	0.95	0.4965	0.0007
		100	-14.99	0.0788	0.88	-0.1354	0.0035	0.01	0.4251	0.88	0.4253	0.0006
		200	-11.43	0.1355	0.80	0.0220	0.0023	0.01	0.3489	0.80	0.3522	0.0006
		500	-2.89	0.1853	0.68	0.1548	0.0011	0.01	0.2661	0.68	0.2713	0.0005
	5	50	-43.41	-0.3168	0.95	-0.9310	0.0088	0.00	0.5466	0.95	0.5406	0.0006
		100	-55.72	-0.1302	0.90	-0.6567	0.0082	0.00	0.4869	0.94	0.4863	0.0006
		200	-30.20	-0.0130	0.83	-0.2652	0.0046	0.02	0.4074	0.89	0.4097	0.0005
		500	-7.12	0.0741	0.70	0.0136	0.0015	0.02	0.2999	0.71	0.3043	0.0004
	6	50	-37.73	-0.6933	0.94	-14.977	0.0114	0.01	0.5752	0.95	0.5704	0.0006
		100	-46.63	-0.3444	0.90	-0.9613	0.0094	0.04	0.5166	0.94	0.5157	0.0006
		200	-48.73	-0.1467	0.88	-0.4246	0.0050	0.06	0.4417	0.88	0.4432	0.0005
		500	-6.95	-0.0036	0.70	-0.0607	0.0015	0.03	0.3154	0.75	0.3187	0.0004
0.50	3	50	-2.76	0.7672	1.00	0.6768	0.0013	0.05	0.7692	1.00	0.7287	0.0008
		100	-3.16	0.7838	0.99	0.7110	0.0011	0.03	0.7842	0.99	0.7479	0.0006
		200	-1.85	0.7922	0.97	0.7585	0.0007	0.06	0.7923	0.97	0.7682	0.0005
		500	-1.44	0.7994	0.93	0.7898	0.0003	0.30	0.7994	0.93	0.7900	0.0003
	4	50	-26.73	0.3909	0.95	0.1911	0.0034	0.01	0.5435	0.95	0.5387	0.0007
		100	-6.68	0.4512	0.93	0.3645	0.0017	0.04	0.5053	0.93	0.5014	0.0007
		200	-5.98	0.4822	0.85	0.4507	0.0009	0.00	0.4932	0.85	0.4869	0.0006
		500	-0.13	0.5042	0.76	0.4937	0.0005	0.05	0.5044	0.76	0.4963	0.0004
	5	50	-31.48	0.0577	0.95	-0.4320	0.0071	0.01	0.5568	0.95	0.5518	0.0006
		100	-36.17	0.2179	0.89	-0.0558	0.0046	0.03	0.4939	0.91	0.4917	0.0006
		200	-14.91	0.2946	0.87	0.2056	0.0019	0.03	0.4284	0.87	0.4293	0.0006
		500	-1.69	0.3442	0.75	0.3198	0.0008	0.03	0.3782	0.75	0.3788	0.0005
	6	50	-28.78	-0.3556	0.95	-10.445	0.0098	0.00	0.5714	0.95	0.5674	0.0006
		100	-41.40	-0.0528	0.88	-0.4928	0.0073	0.05	0.5051	0.92	0.5044	0.0005
		200	-14.48	0.0888	0.83	-0.0702	0.0032	0.01	0.4261	0.85	0.4281	0.0005
		500	-3.66	0.1724	0.72	0.1311	0.0012	0.04	0.3398	0.72	0.3422	0.0004
0.80	3	50	-1.11	0.9398	1.00	0.9117	0.0004	0.14	0.9398	1.00	0.9133	0.0004
		100	-0.01	0.9475	1.00	0.9352	0.0002	0.29	0.9475	1.00	0.9354	0.0002
		200	-0.71	0.9514	0.99	0.9467	0.0001	0.54	0.9514	0.99	0.9468	0.0001
		500	0.86	0.9538	0.99	0.9522	0.0001	0.86	0.9538	0.99	0.9522	0.0001
	4	50	-6.46	0.7843	0.99	0.7305	0.0010	0.04	0.7855	0.99	0.7593	0.0006
		100	-0.83	0.8085	0.97	0.7895	0.0004	0.14	0.8085	0.97	0.7918	0.0004
		200	0.17	0.8211	0.95	0.8131	0.0003	0.42	0.8211	0.95	0.8132	0.0003
		500	0.58	0.8289	0.92	0.8259	0.0001	0.58	0.8289	0.92	0.8259	0.0001
	5	50	-35.51	0.5827	0.97	0.3756	0.0034	0.03	0.6599	0.97	0.6498	0.0006
		100	-21.59	0.6516	0.95	0.5819	0.0015	0.08	0.6663	0.95	0.6535	0.0006
		200	-1.59	0.6842	0.94	0.6595	0.0006	0.12	0.6849	0.94	0.6706	0.0005
		500	0.20	0.7047	0.89	0.6968	0.0003	0.30	0.7047	0.89	0.6970	0.0003
	6	50	-30.85	0.1996	0.96	-0.2625	0.0069	0.08	0.5983	0.96	0.5932	0.0006
		100	-17.18	0.3520	0.94	0.1631	0.0034	0.08	0.5516	0.94	0.5497	0.0005
		200	-8.00	0.4226	0.87	0.3589	0.0014	0.11	0.5166	0.87	0.5155	0.0005
		500	-0.93	0.4664	0.83	0.4472	0.0006	0.10	0.4901	0.83	0.4909	0.0004

Although the defined range lies between 0 and 1, there are negative minimum and maximum values of DD. In that case, AODD should be used instead of ODD. According to represented results in Table 2, the estimates of ODD and AODD differ for each scenario. The results in Table 2 show that when the correlation between raters increases, the values of ODD and AODD also increase. When the sample size increases, the value of ODD and AODD decrease for the tables with low correlation ($\rho = 0.20$). The tables with medium or high correlation, with $R \leq 4$ categories, and with $n > 50$ sample size, are not affected by the sample sizes. When the sample size increases, ODD and AODD decrease for the tables with medium or high correlation and with $R \geq 5$ categories. When the number of categories increases, ODD and AODD decrease for the tables with medium or high correlation. The tables with three categories have the highest distinguishability.

3.3 Modeling agreement coefficients and AODD

As Darroch and McCloud (1986) remark in their study, one of the components of rater agreement is the category distinguishability. If raters cannot distinguish the categories, this may cause wrong decisions and the wrong decisions may affect the level of agreement. At this part of the study, we proposed to investigate rater agreement and category distinguishability together.

The polynomial regression models of linearly weighted kappa coefficient and AC2 are discussed on the following equations. A pilot study is performed first with the different polynomial degrees and the fits of the models are compared. There are no statistically significant differences between the models with 5 and 10 degrees, but there are difference between the models with 3 and 5 degrees. Thus, the maximum degree of polynomial terms is accepted as five. In Equation (3.8), weighted kappa coefficient is explained by a function of AODD. In Equation (3.9), AC2 is explained by a function of AODD.

$$\hat{\kappa}_w = \beta_0 + \beta_1\delta^a + \beta_2(\delta^a)^2 + \beta_3(\delta^a)^3 + \beta_4(\delta^a)^4 + \beta_5(\delta^a)^5 \tag{3.8}$$

$$\widehat{AC2} = \beta_0 + \beta_1\delta^a + \beta_2(\delta^a)^2 + \beta_3(\delta^a)^3 + \beta_4(\delta^a)^4 + \beta_5(\delta^a)^5 \tag{3.9}$$

Table 3. The estimates and standard errors (SE) of the polynomial regression models for $\rho = 0.20$.

			β_0		β_1		β_2		β_3		β_4		β_5		
R	n	Model	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	p-value
3	50	κ_w	0.087	0.020	-1.062	0.157	4.653	0.442	-7.089	0.529	3.924	0.227			< 0.001
		AC2	0.615	0.005	0.501	0.016			-1.155	0.043	0.973	0.034			< 0.001
	100	κ_w	0.043	0.009	-0.560	0.079	2.873	0.252	-4.643	0.335	2.782	0.158			< 0.001
		AC2	0.709	0.007	0.162	0.059	0.429	0.186	-1.342	0.247	0.974	0.117			< 0.001
	200	κ_w	0.018	0.004	-0.227	0.039	1.694	0.145	-2.944	0.219	1.912	0.116			< 0.001
		AC2	0.764	0.001	0.360	0.020	-0.749	0.051	0.525	0.035					< 0.001
	500	κ_w	0.017	0.002	-0.214	0.041	2.340	0.257	-6.551	0.720	8.483	0.929	-3.904	0.449	< 0.001
		AC2	0.780	0.001			0.130	0.014	-0.235	0.042	0.207	0.033			< 0.001
4	50	κ_w	0.120	0.020	-0.209	0.157	1.006	0.442	-1.599	0.529	1.010	0.227			< 0.001
		AC2	0.356	0.006	0.127	0.043	-0.236	0.092	0.250	0.063					< 0.001
	100	κ_w	0.089	0.004	0.091	0.031	-0.209	0.076	0.313	0.058					< 0.001
		AC2	0.360	0.002			0.187	0.050	-0.345	0.142	0.316	0.109			< 0.001
	200	κ_w	0.089	0.001	0.267	0.046	-0.491	0.152	0.491	0.134					< 0.001
		AC2	0.372	0.006	-0.293	0.103	2.103	0.679	-5.809	2.050	7.509	2.867	-3.495	1.502	< 0.001
	500	κ_w	0.089	0.001			0.279	0.013	-0.111	0.026					< 0.001
		AC2	0.358	0.001			0.199	0.012	-0.071	0.024					< 0.001
5	50	κ_w	0.090	0.008	0.143	0.051	-0.254	0.103	0.202	0.066			0.090	0.008	< 0.001
		AC2	0.436	0.004	-0.052	0.016	0.076	0.015							< 0.001
	100	κ_w	0.102	0.001	0.021	0.002							0.102	0.001	< 0.001
		AC2	0.412	0.001			0.066	0.014	-0.046	0.017					< 0.001
	200	κ_w	0.103	0.001					0.073	0.019	-0.082	0.028	0.103	0.001	< 0.001
		AC2	0.421	0.004	-0.066	0.028	0.200	0.069	-0.148	0.054					< 0.001
	500	κ_w	0.100	0.001			0.054	0.002					0.100	0.001	< 0.001
		AC2	0.411	0.001			0.047	0.002							< 0.001
6	50	κ_w													0.142
		AC2													0.162
	100	κ_w	0.127	0.001					-0.023	0.003					< 0.001
		AC2	0.139	0.001					-0.032	0.003					< 0.001
	200	κ_w	0.122	0.001			-0.021	0.002							< 0.001
		AC2	0.470	0.003	-0.391	0.045	1.142	0.180							< 0.001
	500	κ_w	0.116	0.001			0.005	0.002							0.002
		AC2													0.062

According to the simulation study results in Tables 3–5, different models have been found for each scenario. All the models are found statistically significant ($p < 0.05$), except for $\rho = 0.20$ where $R = 6$ of $n = 50$ and $n = 500$. The results in Tables 3 and 4 show that AODD explains κ_w better than AC2 for the tables with low or medium correlations. For the tables with high correlation, even though the fit of κ_w model is better than the fit of AC2 model where $R = 3$, AC2 results are better where $R > 3$. In other words, AODD explains κ_w better than AC2 where $R = 3$, and explains AC2 better than κ_w where $R > 3$.

The observed and predicted values of κ_w and AC2 under the models in Tables 3–5 are summarized in Figures 2 and 3, respectively. According to the results, observed and predicted values of κ_w are found very similar. While the sample size increases, the difference between these values decreases and becomes very similar. The medians of the predicted values of

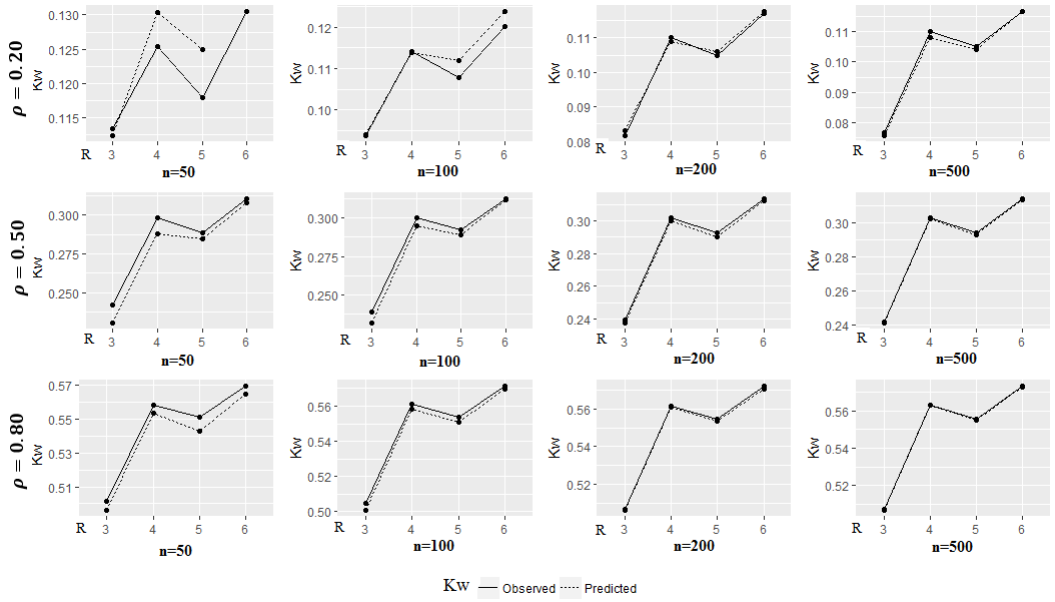

Figure 2. The of observed and predicted weighted kappa results by median.

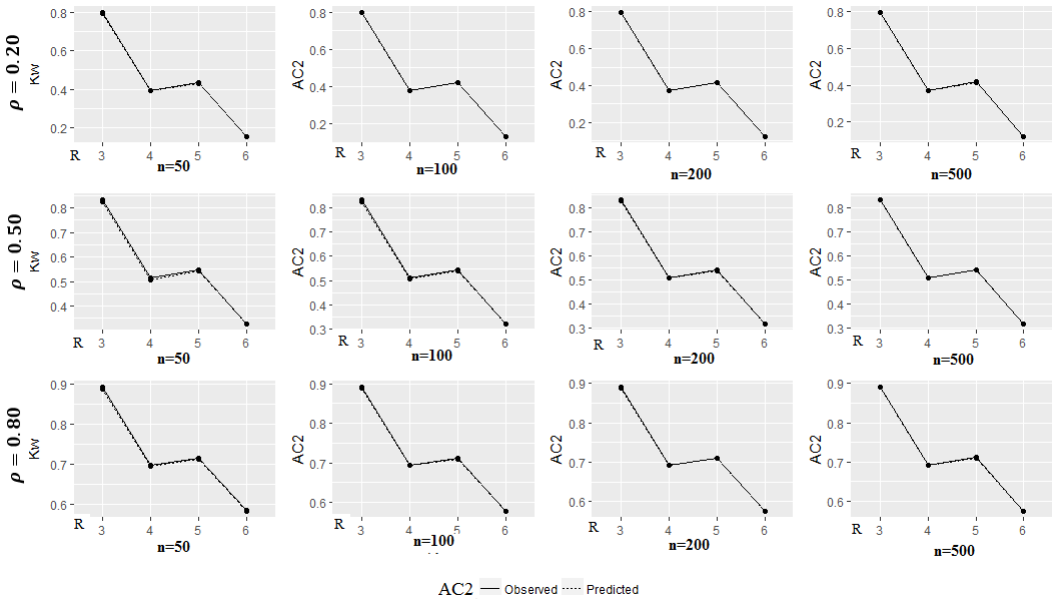

Figure 3. The observed and predicted AC2 results by median.

AC2 are also found very similar to observed ones. The results show that the value of AODD affects the value of agreement coefficients and there is a non-linear correlation between them.

Table 4. The estimates and standard errors (SE) of the polynomial regression models for $\rho = 0.50$.

			β_0		β_1		β_2		β_3		β_4		β_5		
R	n	Model	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	p-value
3	50	κ_w	0.260	0.035	-2.883	0.249	10.92	0.635	-15.25	0.692	7.589	0.273			< 0.001
		AC2	0.667	0.003	1.748	0.044	-3.455	0.085	1.979	0.045					< 0.001
	100	κ_w	0.089	0.022	-1.321	0.157	6.225	0.413	-9.495	0.463	5.100	0.188			< 0.001
		AC2	0.698	0.004	0.366	0.013	-1.025	0.030	0.886	0.022					< 0.001
	200	κ_w	-0.010	0.024	-0.293	0.171	3.039	0.436	-5.560	0.478	3.396	0.191			< 0.001
		AC2	0.813	0.015	-0.319	0.107	1.492	0.272	-2.404	0.298	1.339	0.119			< 0.001
	500	κ_w	0.297	0.091	-1.811	0.554	5.688	1.251	-7.477	1.236	3.859	0.452			< 0.001
		AC2	0.620	0.018	1.023	0.078	-1.812	0.110	1.078	0.052					< 0.001
4	50	κ_w	0.199	0.007	0.202	0.045	-0.321	0.091	0.458	0.058					< 0.001
		AC2	0.453	0.002	0.051	0.006			0.165	0.006					< 0.001
	100	κ_w	0.211	0.002	0.121	0.005			0.176	0.006					< 0.001
		AC2	0.455	0.002	0.016	0.007	0.175	0.007							< 0.001
	200	κ_w	0.218	0.002	0.091	0.007	0.152	0.008							< 0.001
		AC2	0.453	0.001	0.450	0.030	-0.672	0.082	0.435	0.061					< 0.001
	500	κ_w	0.199	0.004	0.277	0.032	-0.319	0.070	0.343	0.051					< 0.001
		AC2	0.462	0.001	0.024	0.006	0.132	0.006							< 0.001
5	50	κ_w	0.239	0.004			0.546	0.083	-1.190	0.197	0.855	0.128	0.239	0.004	< 0.001
		AC2	0.514	0.003			0.242	0.062	-0.482	0.148	0.365	0.096			< 0.001
	100	κ_w	0.261	0.001			0.114	0.002					0.261	0.001	< 0.001
		AC2	0.518	0.001			0.091	0.002							< 0.001
	200	κ_w	0.265	0.001			0.138	0.002					0.265	0.001	< 0.001
		AC2	0.520	0.001			0.100	0.001							< 0.001
	500	κ_w	0.260	0.001	0.063	0.006	0.063	0.008					0.260	0.001	< 0.001
		AC2	0.518	0.001	0.040	0.005	0.044	0.006							< 0.001
6	50	κ_w	0.274	0.014	0.192	0.080	-0.373	0.152	0.247	0.093					< 0.001
		AC2	0.288	0.013	0.225	0.074	-0.438	0.140	0.283	0.086					< 0.001
	100	κ_w	0.304	0.002			0.053	0.017	-0.043	0.020					< 0.001
		AC2	0.315	0.001			0.049	0.015	-0.046	0.019					< 0.001
	200	κ_w	0.293	0.003			0.740	0.193	-3.090	0.863	4.910	1.395	-2.684	0.774	< 0.001
		AC2	0.337	0.018	-0.505	0.245	3.217	1.278	-8.814	3.170	11.15	3.746	-5.288	1.695	< 0.001
	500	κ_w	0.300	0.001	0.027	0.008	0.037	0.011							< 0.001
		AC2	0.303	0.001	0.026	0.007	0.033	0.010							< 0.001

Table 5. The estimates and standard errors (SE) of the polynomial regression models for $\rho = 0.80$.

			β_0		β_1		β_2		β_3		β_4		β_5		
R	n	Model	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	p-value
3	50	κ_w	-16.52	0.860	133.8	6.480	-422.0	19.09	654.7	27.50	-499.7	19.41	150.5	5.380	< 0.001
		AC2	-3.761	0.312	36.22	2.352	-113.5	6.931	176.1	9.988	-134.9	7.050	40.78	1.953	< 0.001
	100	κ_w	-52.94	2.45	383.6	16.93	-1091.1	45.96	1533.1	61.51	-1066.2	40.62	294.4	10.60	< 0.001
		AC2	-13.72	0.957	105.4	6.594	-300.6	17.90	423.6	23.96	-295.6	15.82	81.87	4.128	< 0.001
	200	κ_w	48.26	2.360	-254.6	11.63	503.2	21.29	-439.6	17.16	143.5	5.140			< 0.001
		AC2	-46.40	7.130	314.3	44.99	-828.4	112.3	1083.7	138.9	-704.8	85.11	182.5	20.71	< 0.001
	500	κ_w	-148.3	8.070	492.6	25.74	-545.9	27.36	202.4	9.690					< 0.001
		AC2	-42.18	3.720	143.0	11.87	-158.6	12.61	58.76	4.470					< 0.001
4	50	κ_w	0.502	0.025	-1.621	0.187	6.054	0.486	-8.000	0.539	3.850	0.217			< 0.001
		AC2	0.673	0.016	-1.245	0.113	4.583	0.295	-6.027	0.328	2.873	0.132			< 0.001
	100	κ_w	0.631	0.054	-2.461	0.354	8.061	0.839	-10.10	0.863	4.649	0.325			< 0.001
		AC2	0.654	0.033	-0.962	0.215	3.624	0.508	-4.896	0.523	2.422	0.197			< 0.001
	200	κ_w	-0.410	0.075	3.726	0.299	-5.556	0.395	2.995	0.173					< 0.001
		AC2	0.069	0.045	2.443	0.177	-3.737	0.234	2.054	0.102					< 0.001
	500	κ_w	-0.659	0.305	4.603	1.138	-6.606	1.415	3.416	0.586					< 0.001
		AC2	1.095	0.015	-1.464	0.038	1.179	0.023							< 0.001
5	50	κ_w	0.530	0.022	-0.771	0.173	2.868	0.591	-3.784	0.491	1.903	0.256	0.530	0.022	< 0.001
		AC2	0.697	0.014	-0.435	0.110	1.607	0.312	-2.077	0.375	1.045	0.163			< 0.001
	100	κ_w	0.515	0.021	-0.507	0.162	2.002	0.453	-2.615	0.541	1.314	0.234	0.515	0.021	< 0.001
		AC2	0.671	0.014	-0.242	0.105	1.178	0.293	-1.696	0.350	<0.001	0.152			< 0.001
	200	κ_w	0.390	0.011	0.522	0.056	-0.816	0.094	0.587	0.051			0.390	0.011	< 0.001
		AC2	0.712	0.017	-0.472	0.124	1.718	0.337	-2.315	0.395	1.178	0.169			< 0.001
	500	κ_w	0.360	0.029	0.690	0.133	-1.115	0.206	0.752	0.105			0.360	0.029	< 0.001
		AC2	0.581	0.019	0.494	0.088	-0.843	0.136	0.570	0.069					< 0.001
6	50	κ_w	0.526	0.005			0.233	0.082	-0.371	0.185	0.276	0.114			< 0.001
		AC2	0.459	0.040	0.994	0.440	-4.113	1.837	8.592	3.662	-8.542	3.501	3.318	1.289	< 0.001
	100	κ_w	0.532	0.002					0.812	0.087	-1.578	0.220	0.935	0.145	< 0.001
		AC2	0.533	0.003			0.363	0.052	-0.614	0.124	0.392	0.081			< 0.001
	200	κ_w	0.522	0.001	0.092	0.003							0.053	0.006	< 0.001
		AC2	0.530	0.001	0.081	0.003					0.047	0.005			< 0.001
	500	κ_w	0.509	0.007	0.224	0.047	-0.308	0.098	0.237	0.067					< 0.001
		AC2	0.536	0.002	0.061	0.008	0.035	0.008							< 0.001

3.4 The benchmark scale of AODD

Yilmaz and Saracbasi (2019) suggest the interpretation levels for ADD in 2×2 tables. For ordinal tables, even if there are some general ones, there is not any scale to interpret the level of category distinguishability.

To develop mixed tables of the linearly weighted kappa coefficient and AODD, we generate 50000 tables for each scenario. For each simulation, we use a random sample size between 40 and 1000; and a random correlation between 0.10 and 0.90. Then, we classify the values of linearly weighted kappa into nine categories and class interval is accepted as 0.10. According

to the defined intervals, descriptive statistics of AODD values are summarized in Table 6.

Table 6. The descriptive statistics of the predicted AODD values.

R	κ_w interval	Min	Median	Max	R	κ_w interval	Min	Median	Max
3	0.00-0.10	0.01	0.42	0.90	5	0.00-0.10	0.02	0.30	0.88
	0.11-0.20	0.33	0.67	0.91		0.11-0.20	0.04	0.31	0.87
	0.21-0.30	0.46	0.81	0.94		0.21-0.30	0.03	0.36	0.85
	0.31-0.40	0.48	0.89	0.97		0.31-0.40	0.04	0.45	0.87
	0.41-0.50	0.50	0.94	0.98		0.41-0.50	0.16	0.57	0.88
	0.51-0.60	0.78	0.96	0.99		0.51-0.60	0.27	0.70	0.90
	0.61-0.70	0.86	0.98	0.99		0.61-0.70	0.41	0.81	0.93
	0.71-0.80	0.97	0.99	1.00		0.71-0.80	0.68	0.89	0.96
	0.81-1.00	0.99	0.99	1.00		0.81-1.00	0.95	0.98	0.99
4	0.00-0.10	0.02	0.24	0.78	6	0.00-0.10	0.04	0.32	0.86
	0.11-0.20	0.01	0.31	0.80		0.11-0.20	0.05	0.32	0.86
	0.21-0.30	0.05	0.44	0.87		0.21-0.30	0.04	0.33	0.86
	0.31-0.40	0.17	0.58	0.87		0.31-0.40	0.07	0.36	0.93
	0.41-0.50	0.29	0.65	0.90		0.41-0.50	0.07	0.41	0.81
	0.51-0.60	0.46	0.80	0.94		0.51-0.60	0.11	0.61	0.85
	0.61-0.70	0.71	0.90	0.97		0.61-0.70	0.30	0.78	0.89
	0.71-0.80	0.85	0.94	0.97					
	0.81-1.00	0.95	0.96	0.96					

We suggest the benchmark scales in Table 7 for AODD given in Equation (2.7) by considering medians in Table 6. We use the midpoints of medians to decide the thresholds. To give an example, the first threshold of 3×3 tables is the midpoint of $3rd$ and $4th$ classes $((0.81 + 0.89)/2 = 0.85)$. The third threshold is the midpoint of $5th$ and $6th$ classes $((0.94 + 0.95)/2 = 0.95)$.

Table 7. The benchmark scales of AODD.

R	κ_w	δ^a	**Strength of** δ^a
3	$\geq$0.51	$\geq$0.95	Good
	0.31-0.50	0.85-0.94	Moderate
	$\leq$0.30	$\leq$0.84	Fair
4	$\geq$0.71	$\geq$0.92	Good
	0.51-0.70	0.72-0.91	Moderate
	$\leq$0.50	$\leq$0.71	Fair
5	$\geq$0.81	$\geq$0.94	Good
	0.61-0.80	0.76-0.93	Moderate
	$\leq$0.60	$\leq$0.75	Fair
6	$\leq$0.60	$\leq$0.70	Good
	$\leq$0.60	$\leq$0.70	Fair

In order to test the validity of the defined intervals, a simulation study is performed with 50,000 replications for each scenario. Linearly weighted kappa coefficient and AODD are calculated for each replication. The correct classification rates are calculated for each scenario and given in Table 8. The correct classification rates in Table 8 change between 0.63 and 1.00.

Table 8. The correct classification rates.

		ρ		
R	n	0.20	0.50	0.80
3	50	0.95	0.81	0.71
	100	0.98	0.84	0.73
	200	0.98	0.87	0.74
	500	1.00	0.94	0.63
4	50	0.93	0.86	0.72
	100	0.98	0.92	0.80
	200	0.99	0.97	0.89
	500	1.00	1.00	0.98
5	50	0.94	0.92	0.72
	100	0.98	0.98	0.77
	200	0.99	0.99	0.78
	500	1.00	1.00	0.80
6	50	0.82	0.83	0.69
	100	0.92	0.94	0.71
	200	0.99	0.99	0.80
	500	1.00	1.00	0.90

3.5 Illustrative Examples

Example 1: In the study of Walsh et al. (2014), two raters classify 159 children in terms of immediate "gestalt" impression of overall clinical appearance. After examining the children, they are classified again in terms of their clinical impression. Intra-rater and inter-rater reliabilities are considered and summarize in Tables 9 and 10.

Table 9. The initial gestalt classifications (classifications following examination) of two raters.

	First Rater			
Second Rater	Not ill appearing	Unsure	Ill appearing	Total
Not ill appearing	94 (103)	11 (6)	13 (14)	118 (123)
Unsure	12 (8)	0 (0)	2 (1)	14 (9)
Ill appearing	14 (14)	5 (2)	8 (11)	27 (27)
Total	120 (125)	16 (8)	23 (26)	159

Table 10. The first (second) rater's initial-after classifications.

	After Examining Xhild			
Gestalt Impression	Not ill appearing	Unsure	Ill appearing	Total
Not ill appearing	113 (113)	3 (3)	2 (4)	118 (120)
Unsure	8 (9)	4 (5)	2 (2)	14 (16)
Ill appearing	2(3)	2 (0)	23 (20)	27 (23)
Total	123 (125)	9 (8)	27 (26)	159

The calculated overall categories are summarized in Table 11. R codes for calculation of ODD and AODD are provided in Appendix A. Firstly, the existence of zero cells are detected

and if there are any, adding 0.5 to each cell. Then, the DD and ADD values are calculated for the adjacent categories and represented as a matrix. The means of the DD and ADD values give the ODD and AODD, respectively.

The linearly weighted kappa coefficient results show that the agreements between the classifications are found as "slight", "fair", "substantial", and "substantial", respectively. The linearly AC2 coefficient results indicate "moderate" agreements for all the tables. Because of the negative values of DD, the ODD of initial classifications are also found as negative. Thus, AODD is preferred over ODD. The distinguishabilities are found as "fair", "fair", "good", and "good", respectively. Because of the low levels of agreement between the raters, the indistinguishable categories are investigated by using the pairwise ADD.

Table 11. The calculated overall coefficients.

Tables	κ_w	AC2	δ	δ^a
Initial	0.177	0.477	-1.174	0.681
After	0.261	0.518	0.368	0.206
Rater 1	0.777	0.574	0.967	0.952
Rater 2	0.714	0.551	0.976	0.968

The calculated DD and ADD values for each pair of categories are summarized in Table 12. The results show that DD of initial classifications between "Not ill appearing"–"Unsure" and "Unsure"–"Ill appearing" and DD of after classifications between "Not ill appearing"–"Unsure" are calculated as negative. In this case, ADD can be used instead of DD. The initial classifications indicate the fair and homogeneous distinguishabilities and after classifications indicate the poor distinguishabilities, except for 'Not ill appearing"–"Ill appearing".

Table 12. The calculated DD and ADD for gestalt data.

	Initial		After	
Categories	δ_{ij}	δ^a_{ij}	δ_{ij}	δ^a_{ij}
Not ill appearing-Unsure	-2.042	0.671	-0.068	0.063
Unsure-Ill appearing	-2.235	0.691	0.348	0.348
Not ill appearing-Ill appearing	0.756	–	0.823	–

Even if high levels of intra-rater agreements and distinguishabilities are found, the inter-rater ones are found at low levels. The reason is the inability of the raters to distinguish the categories, even if they are consistent in themselves. Because there are poor distinguishabilities of categories, we reclassify the categories (Table 13). After the reclassification of initial classifications, the weighted kappa, AC2, and DD are increased to 0.194, 0.528, and 0.696, respectively. For the reclassification following examination, the weighted kappa, AC2, and DD are increased to 0.298, 0.427, and 0.814, respectively.

Table 13. The reclassified initial (following examination) results of two raters.

	First Rater		
Second Rater	Not ill appearing+Unsure	Ill appearing	Total
Not ill appearing+Unsure	117 (117)	15 (15)	132 (132)
Ill appearing	19 (16)	8 (11)	27 (27)
Total	136 (133)	23 (26)	159

Example 2: The data is taken from Oh (2009) and given in Table 14. The radiographs of 60 patients are shown to two groups of doctors (two trauma surgeons and two radiologists).

Table 14. The ratings given by trauma surgeons and radiologists.

Trauma	Radiologists				
Surgeons	0	1	2	3	Total
0	3	15	1	2	21
1	1	11	13	1	26
2	1	5	4	2	12
3	0	0	1	0	1
Total	5	31	9	5	60

For Table 14, linearly weighted kappa coefficient is calculated as 0.11. The correlation between doctors is $\hat{\rho} = 0.29$. The calculated values for each pair of categories are summarized in Table 15.

Table 15. The calculated DD and ADD for radiographs data.

	0-1	0-2	0-3	1-2	1-3	2-3
δ_{ij}	0.42	0.86	0.29	-0.43	0.87	-0.67
δ^a_{ij}	0.42	–	–	0.30	–	0.40

DD's between 2–3 and 3–4 are calculated as negative. The category distinguishability between 0–2 and 1–3 are in the good level. Even if $\delta = 0.22$, it decreases to -0.67 when calculating only for the adjacent categories. It is expected that while the distance between categories increases, the level of category distinguishability also increases. Thus, the calculation of ODD over all the category pairs has an increasing effect on its level.

Because the negative values affect its value, AODD is preferred over ODD. AODD for radiographs data is calculated as $\delta^a = 0.38$. When we interpret the level of ODD by the suggested benchmarking scales in Table 7, the calculated values suggest a "fair" agreement between the doctors's decisions and the categories are indistinguishable. In that case, it is possible to define the indistinguishable categories by using the pairwise ADD.

From the functional equations for $n = 50$ and $\rho = 0.20$ in Tables 3, the weighted kappa coefficient is predicted as 0.12 which is similar to the observed one. This example indicates the accuracy of the regression models in Equation (3.8) and (3.9).

4. Conclusions, Limitations, and Future Research

In recent studies, inter-rater agreement analysis has grown extensively. There are different ideas between researchers when the subject is agreement. In the agreement studies, it is necessary to discuss the term of category distinguishability. Yilmaz and Saracbasi (2019) discuss that the value of degree of distinguishability may be calculated outside the defined range as negative and they suggest adjusted degree of distinguishability as an alternative measure to degree of distinguishability. In this study, in order to detect the conditions where degree of distinguishability is negative, we conduct a detailed investigation of the frequency distribution of sub-tables.

For ordered agreement tables, overall degree of distinguishability can be used to detect the general distinguishability of the table. We discuss that the negative values of degree

of distinguishability affect the value of overall degree of distinguishability because it is an average of the all the degree of distinguishability's. Simulation study results in Section 3 show that overall degree of distinguishability can also be found negative which is which is outside the defined range and the results prove the necessity of a correction on the coefficient. We also discuss that degree of distinguishability only for the adjacent categories instead of all the pairs should be used to calculate the overall degree of distinguishability. In this paper, we propose adjusted overall degree of distinguishability as an alternative to overall degree of distinguishability. The new coefficient overcomes the problems of overall degree of distinguishability: (1) solve the problem of its calculation outside the defined range and (2) solve the problem of the inflation of its value by using only the adjacent categories. The adjusted degree of distinguishability and adjusted overall degree of distinguishability can be used for all the ordinal classifications of two raters, one device-one rater, or same rater-two different times in dermatological (see Valet et al. (2007, 2009), psychiatric, and pathological researches, reading ultrasound pictures (see Bagheban et al. (2008)), word-sense distinguishabilities (see Bruce and Wiebe (1998)), etc.

Darroch and McCloud (1986) discuss that weighted kappa coefficient is a measure of observer agreement and it is unsatisfactory as a measure of overall category distinguishability. In other words, even if weighted kappa coefficient explains how well two raters agree with each other, it does not describe how well any rater can distinguish the categories from each other. In this study, we focus on the assessment of weighted kappa, AC2 coefficients, and overall degree of distinguishability; and we discuss their interpretation. Weighted kappa and AC2 coefficients are agreement indexes while degree of distinguishability is a measurement of category distinguishability. Even though they are different coefficients, we focused on how these coefficients associated with each other. The polynomial regression model study results show that category distinguishability has an important influence on the value of agreement coefficients. Adjusted overall degree of distinguishability explains κ_w better than AC2 for the tables with poor and medium correlations, and explains AC2 better than κ_w for the tables with high correlation $R > 3$.

In general, the values of weighted kappa coefficient and adjusted overall degree of distinguishability are affected by the number of categories. When the number of categories increases (especially $R \geq 5$), then the ability of the raters to distinguish the categories becomes weaker. Low distinguishability may affect the agreement, as well. In the rater agreement studies, it is proposed to use weighted kappa coefficient and adjusted overall degree of distinguishability simultaneously. In order to get more distinguishable tables, it is proposed to avoid the tables with more than five categories.

We also propose a benchmarking scale shown in Table 7. It is possible to interpret adjusted overall degree of distinguishability by using these intervals. If the category distinguishability is at poor level, then it means that the raters have some difficulties to distinguish the categories. The reason may be about the unclearly defined categories or non-expert raters. If the problem is about unclearly defined categories, it is suggested to combine the categories by considering the category distinguishability between adjacent categories.

The tables with poor distinguishability indicate poor agreement. Although indistinguishable categories point towards a poor agreement, the good or substantial distinguishability does not always point towards a good agreement. To get a good agreement level, the categories should be distinguishable as well. Besides, marginal homogeneity is also important.

The distributions of adjusted overall degree of distinguishability associated with the linearly weighted kappa coefficient for different number of categories are summarized in the Appendix B. These tables can be used in two ways: (1) If we calculate linearly weighted kappa coefficient, we can find adjusted degree of distinguishability and (2) If we calculate adjusted degree of distinguishability, we can find the linearly weighted kappa coefficient. Let us illustrate how these tables work using a hypothetical inter-rater experiment featur-

ing two raters who have classified 100 patients into four categories. Suppose the weighted kappa coefficient is calculated as 0.44. Then, the value of adjusted degree of distinguishability can be found using the tables similar to using a z-table. The intersection point of 0.4 from the second column and 0.04 from the first row is found. The value is found as 0.68 which is at "fair" level. On the other hand, for the same conditions, if we have a "good" distinguishability ($\delta^a = 0.95$), then the weighted kappa coefficient is expected to be around 0.75 which is at "substantial" level.

This study is limited with the linearly and quadratically weighted kappa and AC2 coefficients. The recent studies show the effect of weighting schemes and coefficients on the agreement Tran et al. (2020); Yilmaz and Altas (2018). In the future studies, different coefficients with different weighting schemes can also be considered. Besides, the adjusted degree of distinguishability and adjusted overall degree of distinguishability can be extended for the tables with multiple raters.

Author contributions Conceptualization, A.E.Y.; methodology, A.E.Y.; software, A.E.Y.; validation, A.E.Y.; formal analysis, A.E.Y.; investigation, A.E.Y.; data curation, A.E.Y.; writing-original draft preparation, A.E.Y.; writing-review and editing, A.E.Y.; visualization, A.E.Y.; supervision, A.E.Y. The author has read and agreed the published version of the manuscript.

Acknowledgements The author thanks to the editors, associate editor, and the anonymous referees for their valuable comments to a improve this manuscript.

Funding The author received no financial support for the research, authorship, and/or publication of this article.

Conflicts of interest The authors declare no conflict of interest.

References

Agresti, A., 1988. A model for agreement between ratings on an ordinal scale. Biometrics, 44, 539-548.

Agresti, A., 2002. Categorical Data Analysis. John Wiley & Sons, New York, pp. 397-398.

Bagheban, A.A., Zayeri, F., Anaraki, F.B., and Elahipanah, Z., 2008. The reliability and distinguishability of ultrasound diagnosis of ovarian masses. Indian Journal of Medical Sciences, 62, 217-221.

Becker, M.P., 1989. Using association models to analyse agreement data: Two examples. Statistics in Medicine, 8, 1199-1207.

Bruce, R. and Wiebe, J., 1998. Word-sense distinguishability and inter-coder agreement. In Proceedings of the Third Conference on Empirical Methods for Natural Language Processing, 53-60.

Cicchetti, D.V. and Allison, T., 1971. A new procedure for assessing reliability of scoring EEG sleep recordings. American Journal EEG Technology, 11, 101-109.

Cohen, J., 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 70, 213-220.

Darroch, J.N. and McCloud, P.I., 1986. Category distinguishability and observer agreement. Australian Journal of Statistics, 28, 371-388.

Goktas, A., and Isci, O., 2011. A comparison of the most commonly used measures of association for doubly ordered square contingency tables via simulation. Metodoloski Zvezki, 8, 17-37.

Gwet, K.L., 2012. Handbook of Inter-Rater Reliability, The definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC, Maryland.

Gwet, K.L., 2019. irrICC: Intraclass Correlations for Quantifying Inter-Rater Reliability. R package version 1.0. https://CRAN.R-project.org/package=irrICC.

Oh, M., 2009. Inference on measurements of agreement using marginal association. Journal of the Korean Statistical Society, 38, 41-46.

Perkins, S.M. and Becker, M.P., 2002. Assessing rater agreement using marginal association models. Statistics in Medicine, 21, 1743-1760.

Shoukri, M.M., 2004. Measures of Interrater Agreement. Chapman & Hall/CRC Press LLC, Florida.

Tran, T., Dolgun, A., and Demirhan, D., 2020. Weighted inter-rater agreement measures for ordinal outcomes. Communications in Statistics - Simulation and Computation, 49, 989-1003.

Valet, F., Guinot, C., and Mary, J.Y., 2007. Log-linear non-uniform association models for agreement between two ratings on an ordinal scale. Statistics in Medicine, 26, 647-662.

Valet, F., Ezzedine, K., Malvy, D., Mary, J.Y., and Guinot, C., 2009. Assessing the reliability of four severity scales depicting skin ageing features. British Journal of Dermatology, 161, 153-158.

Valet, F. and Mary, J-Y., 2011. Power estimation of tests in log-linear nonuniform association models for ordinal agreement. BMC Medical Research Methodology, 11, 70-80.

Vélez, J.I. and Marmolejo-Ramos, F., 2017. Extension of a graphical diagnostic test for contingency tables. Chilean Journal of Statistics, 8, 53-65.

Venables, W.N. and Ripley, B.D., 2002. Modern Applied Statistics with S, Fourth edition. Springer, New York. ISBN 0-387-95457-0, http://www.stats.ox.ac.uk/pub/MASS4/.

Walsh, P., Thornton, J., Asato, J., Walker, N., McCoy, G., Baal, J., Baal, J., Mendoza, N., and Banimahd, F., 2014. Approaches to describing inter-rater reliability of the overall clinical appearance of febrile infants and toddlers in the emergency department. PeerJ, 2, 1-19.

Yilmaz, A.E. and Aktas, S., 2018. Ridit and exponential type scores for estimating the kappa statistic. Kuwait Journal of Science, 45, 89-99.

Yilmaz, A.E. and Saracbasi, T., 2019. Agreement and adjusted degree of distinguishability for square contingency tables. Hacettepe Journal of Mathematics and Statistics, 48, 592-604.

Appendix A. R Codes

```
R <- 3   #number of categories
ratings <- matrix(c(94, 12, 14, 11, 0, 5, 13, 2, 8), nrow=R) #data matrix
if (any(ratings=='0')){ratings <- ratings + 0.5}
comb <- factorial(R)/factorial(R-2)/factorial(2)
DD <- matrix(NA, R, R)
ADD <- matrix(NA, R, R)

for (i in 1:(R-1)){
        for (j in (i+1):R){
                DD[i,j]  <- 1-1/(ratings[i,i]*ratings[j,j]/ratings[i,j]/ratings[j,i])
}
}
for (i in 1:(R-1)){
        ADD[i,i+1]  <- 1-1/(ratings[i,i]*ratings[i+1,i+1]/ratings[i,i+1]/ratings[i+1,i])
        if ( ADD[i,i+1]<0){
        ADD[i,i+1]  <- 1-ratings[i,i]*ratings[i+1,i+1]/ratings[i,i+1]/ratings[i+1,i]}
}
ODD <- sum(DD,na.rm = TRUE)/comb  #overall degree of distinguisability
AODD <- sum(ADD,na.rm = TRUE)/(R-1) #adjusted overall degree of distinguisability
```

Appendix B. Tables

Table A1. The distribution of AODD associated with the weighted kappa coefficient for 3×3 tables.

	κ_w	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
$n = 50$	0.0	0.48	0.50	0.52	0.53	0.55	0.56	0.57	0.59	0.60	0.61
	0.1	0.63	0.64	0.65	0.66	0.68	0.69	0.70	0.71	0.72	0.73
	0.2	0.74	0.75	0.76	0.77	0.78	0.79	0.79	0.80	0.81	0.82
	0.3	0.83	0.83	0.84	0.85	0.85	0.86	0.87	0.87	0.88	0.88
	0.4	0.89	0.89	0.90	0.90	0.91	0.91	0.92	0.92	0.92	0.93
	0.5	0.93	0.93	0.94	0.94	0.94	0.94	0.95	0.95	0.95	0.95
	0.6	0.95	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96
	0.7	0.98	0.98	0.98	0.98	0.99	0.99	0.99	0.99	0.99	0.99
	0.8	0.99	0.99	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99
	0.9	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	κ_w	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
$n = 100$	0.0	0.39	0.45	0.47	0.49	0.50	0.52	0.54	0.56	0.57	0.59
	0.1	0.60	0.62	0.63	0.65	0.66	0.68	0.69	0.70	0.72	0.73
	0.2	0.74	0.75	0.77	0.78	0.79	0.80	0.81	0.82	0.83	0.84
	0.3	0.84	0.85	0.86	0.87	0.87	0.88	0.88	0.89	0.89	0.90
	0.4	0.90	0.90	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.94
	0.5	0.94	0.94	0.94	0.94	0.94	0.95	0.95	0.95	0.95	0.95
	0.6	0.95	0.95	0.95	0.94	0.97	0.98	0.98	0.98	0.98	0.98
	0.7	0.98	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	0.8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	0.9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	κ_w	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
$n = 200$	0.0	0.33	0.35	0.36	0.38	0.40	0.43	0.45	0.47	0.50	0.52
	0.1	0.55	0.57	0.59	0.61	0.64	0.66	0.68	0.72	0.73	0.74
	0.2	0.76	0.77	0.78	0.79	0.80	0.81	0.82	0.83	0.84	0.84
	0.3	0.85	0.86	0.87	0.88	0.89	0.90	0.92	0.93	0.90	0.91
	0.4	0.91	0.92	0.92	0.93	0.93	0.93	0.94	0.94	0.94	0.94
	0.5	0.95	0.95	0.95	0.96	0.96	0.96	0.96	0.96	0.97	0.97
	0.6	0.97	0.97	0.97	0.98	0.98	0.98	0.98	0.98	0.98	0.98
	0.7	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00
	0.8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	0.9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	κ_w	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
$n = 500$	0.0	0.16	0.23	0.26	0.29	0.32	0.36	0.40	0.43	0.47	0.50
	0.1	0.53	0.56	0.59	0.61	0.64	0.66	0.68	0.70	0.71	0.73
	0.2	0.74	0.76	0.77	0.78	0.78	0.81	0.82	0.83	0.83	0.84
	0.3	0.86	0.87	0.87	0.88	0.89	0.89	0.90	0.90	0.91	0.91
	0.4	0.92	0.92	0.93	0.94	0.94	0.94	0.94	0.94	0.95	0.95
	0.5	0.95	0.95	0.96	0.96	0.96	0.96	0.96	0.97	0.97	0.97
	0.6	0.97	0.97	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.99
	0.7	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	0.8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	0.9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table A2. The distribution of AODD associated with the weighted kappa coefficient for 4×4 tables.

	κ_w	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
$n = 50$	0.0		0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47	0.47
	0.1	0.47	0.48	0.47	0.47	0.47	0.46	0.47	0.47	0.47	0.47
	0.2	0.47	0.48	0.48	0.49	0.50	0.50	0.51	0.51	0.51	0.52
	0.3	0.53	0.54	0.56	0.57	0.57	0.58	0.59	0.59	0.61	0.62
	0.4	0.63	0.64	0.66	0.67	0.68	0.69	0.71	0.71	0.72	0.73
	0.5	0.72	0.73	0.75	0.77	0.78	0.79	0.80	0.80	0.81	0.83
	0.6	0.84	0.85	0.87	0.88	0.89	0.90	0.91	0.91	0.92	0.90
	0.7	0.91	0.91	0.92	0.93	0.93	0.93	0.94	0.94	0.94	0.94
	0.8	0.94	0.94	-	-	-	-	-	-	-	-
	0.9	-	-	-	-	-	-	-	-	-	-
	κ_w	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
$n = 100$	0.0	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40
	0.1	0.40	0.41	0.40	0.40	0.40	0.41	0.41	0.41	0.42	0.42
	0.2	0.43	0.44	0.46	0.47	0.48	0.49	0.50	0.50	0.45	0.47
	0.3	0.49	0.50	0.54	0.55	0.57	0.58	0.60	0.60	0.62	0.63
	0.4	0.60	0.62	0.65	0.66	0.68	0.69	0.71	0.71	0.72	0.73
	0.5	0.74	0.76	0.78	0.79	0.80	0.81	0.81	0.81	0.82	0.83
	0.6	0.84	0.85	0.87	0.88	0.89	0.89	0.90	0.90	0.91	0.91
	0.7	0.92	0.93	0.94	0.94	0.95	0.95	0.95	0.95	0.96	0.96
	0.8	-	-	-	-	-	-	-	-	-	-
	0.9	-	-	-	-	-	-	-	-	-	-
	κ_w	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
$n = 200$	0.0	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32
	0.1	0.33	0.34	0.32	0.33	0.34	0.34	0.35	0.35	0.36	0.38
	0.2	0.39	0.40	0.43	0.44	0.46	0.47	0.49	0.49	0.50	0.51
	0.3	0.53	0.50	0.54	0.55	0.57	0.59	0.61	0.61	0.62	0.60
	0.4	0.62	0.64	0.67	0.68	0.69	0.71	0.72	0.72	0.73	0.74
	0.5	0.74	0.76	0.78	0.79	0.80	0.82	0.83	0.83	0.84	0.85
	0.6	0.86	0.86	0.88	0.89	0.89	0.90	0.91	0.91	0.91	0.92
	0.7	0.92	0.93	0.94	-	-	-	-	-	-	-
	0.8	-	-	-	-	-	-	-	-	-	-
	0.9	-	-	-	-	-	-	-	-	-	-
	κ_w	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
$n = 500$	0.0					0.22	0.22	0.22	0.22	0.23	0.24
	0.1	0.25	0.26	0.29	0.31	0.33	0.35	0.37	0.37	0.30	0.32
	0.2	0.35	0.37	0.41	0.42	0.38	0.41	0.43	0.43	0.45	0.47
	0.3	0.49	0.51	0.55	0.57	0.59	0.61	0.63	0.63	0.64	0.66
	0.4	0.67	0.69	0.72	0.70	0.69	0.70	0.72	0.72	0.73	0.74
	0.5	0.76	0.77	0.79	0.80	0.81	0.82	0.83	0.83	0.84	0.85
	0.6	0.86	0.87	0.89	0.89	0.90	0.91	-	-	-	-
	0.7	-	-	-	-	-	-	-	-	-	-
	0.8	-	-	-	-	-	-	-	-	-	-
	0.9	-	-	-	-	-	-	-	-	-	-

Table A3. The distribution of AODD associated with the weighted kappa coefficient for 5×5 tables.

	κ_w	0.00	0.01	0.20	0.03	0.04	0.05	0.06	0.07	0.08	0.09
$n = 50$	0.0	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53
	0.1	0.53	0.53	0.53	0.53	0.53	0.53	0.54	0.53	0.53	0.53
	0.2	0.53	0.53	0.53	0.53	0.53	0.54	0.54	0.54	0.54	0.54
	0.3	0.55	0.55	0.55	0.56	0.56	0.56	0.57	0.57	0.58	0.58
	0.4	0.58	0.59	0.59	0.60	0.61	0.61	0.62	0.62	0.63	0.64
	0.5	0.64	0.65	0.66	0.67	0.68	0.68	0.68	0.69	0.70	0.71
	0.6	0.72	0.73	0.74	0.75	0.77	0.78	0.79	0.77	0.78	0.79
	0.7	0.81	0.82	0.83	0.85	0.86	0.87	0.89	0.90	0.92	0.93
	0.8	0.95	0.96	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	0.9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	κ_w	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
$n = 100$	0.0	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48
	0.1	0.48	0.48	0.48	0.47	0.47	0.47	0.47	0.47	0.47	0.48
	0.2	0.48	0.48	0.48	0.49	0.49	0.50	0.50	0.51	0.51	0.52
	0.3	0.51	0.51	0.52	0.52	0.53	0.53	0.53	0.54	0.54	0.55
	0.4	0.55	0.53	0.54	0.55	0.56	0.57	0.58	0.59	0.60	0.61
	0.5	0.63	0.64	0.65	0.67	0.68	0.69	0.71	0.72	0.74	0.75
	0.6	0.77	0.73	0.74	0.75	0.77	0.78	0.79	0.80	0.81	0.83
	0.7	0.84	0.85	0.86	0.87	0.87	0.88	0.89	0.90	-	-
	0.8	-	-	-	-	-	-	-	-	-	-
	0.9	-	-	-	-	-	-	-	-	-	-
	κ_w	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
$n = 200$	0.0	-	-	-	-	-	-	-	0.40	0.40	0.40
	0.1	0.40	0.40	0.38	0.38	0.39	0.39	0.39	0.39	0.40	0.40
	0.2	0.40	0.41	0.41	0.42	0.42	0.43	0.44	0.45	0.46	0.48
	0.3	0.43	0.44	0.45	0.46	0.47	0.48	0.49	0.50	0.51	0.47
	0.4	0.49	0.50	0.52	0.53	0.55	0.56	0.58	0.60	0.61	0.63
	0.5	0.65	0.66	0.62	0.64	0.65	0.67	0.68	0.70	0.71	0.73
	0.6	0.74	0.75	0.77	0.78	0.80	0.81	0.82	0.83	0.85	0.86
	0.7	0.87	0.88	0.89	0.90	0.91	0.92	0.93	0.94	0.95	0.96
	0.8	0.96	0.97	0.98	0.98	0.99	0.99	0.99	1.00	1.00	1.00
	0.9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	κ_w	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
$n = 500$	0.0	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29
	0.1	0.30	0.30	0.30	0.31	0.32	0.33	0.29	0.30	0.31	0.32
	0.2	0.33	0.34	0.35	0.36	0.31	0.32	0.33	0.34	0.36	0.37
	0.3	0.39	0.40	0.42	0.43	0.45	0.47	0.49	0.50	0.52	0.55
	0.4	0.57	0.59	0.61	0.47	0.49	0.51	0.53	0.55	0.57	0.59
	0.5	0.61	0.62	0.64	0.66	0.67	0.69	0.71	0.72	0.73	0.75
	0.6	0.76	0.78	0.79	0.80	0.81	0.82	0.83	-	-	-
	0.7	-	-	-	-	-	-	-	-	-	-
	0.8	-	-	-	-	-	-	-	-	-	-
	0.9	-	-	-	-	-	-	-	-	-	-

Table A4. The distribution of AODD associated with the weighted kappa coefficient for 6×6 tables.

	κ_w	0.00	0.01	0.2	0.03	0.04	0.05	0.06	0.07	0.08	0.09
$n = 50$	0.0	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57
	0.1	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57
	0.2	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57
	0.3	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57	0.57
	0.4	0.57	0.58	0.58	0.58	0.58	0.59	0.59	0.59	0.59	0.60
	0.5	0.60	0.61	0.61	0.62	0.62	0.63	0.63	0.64	0.65	0.65
	0.6	0.66	0.67	0.68	0.68	0.69	0.70	0.71	0.72	0.73	0.75
	0.7	0.76	0.77	0.78	0.80	0.81	0.73	0.74	0.75	0.76	0.77
	0.8	0.79	0.80	0.81	0.82	0.84	0.85	0.87	0.88	0.90	0.91
	0.9	0.93	0.95	0.96	0.98	0.99	1.00	1.00	1.00	1.00	1.00
	κ_w	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
$n = 100$	0.0	0.52	0.52	0.52	0.52	0.52	0.52	0.52	0.53	0.53	0.53
	0.1	0.53	0.53	0.53	0.53	0.53	0.54	0.54	0.54	0.54	0.55
	0.2	0.53	0.53	0.53	0.53	0.54	0.54	0.54	0.55	0.55	0.55
	0.3	0.56	0.56	0.57	0.58	0.55	0.55	0.56	0.56	0.57	0.58
	0.4	0.58	0.59	0.60	0.61	0.61	0.62	0.63	0.64	0.65	0.66
	0.5	0.68	0.69	0.54	0.55	0.55	0.56	0.57	0.57	0.58	0.59
	0.6	0.60	0.61	0.62	0.63	0.64	0.65	0.66	0.67	0.68	0.70
	0.7	0.71	0.73	0.74	0.76	0.77	0.79	0.80	0.82	0.84	0.86
	0.8	0.88	0.90	0.92	0.94	0.96	0.98	0.99	1.00	1.00	1.00
	0.9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	κ_w	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
$n = 200$	0.0	0.37	0.37	0.37	0.37	0.37	0.37	0.37	0.37	0.37	0.40
	0.1	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40
	0.2	0.40	0.40	0.41	0.41	0.41	0.41	0.41	0.41	0.41	0.42
	0.3	0.42	0.42	0.42	0.42	0.43	0.43	0.43	0.44	0.44	0.45
	0.4	0.45	0.46	0.46	0.47	0.47	0.48	0.48	0.49	0.50	0.50
	0.5	0.51	0.52	0.52	0.53	0.54	0.54	0.55	0.56	0.57	0.57
	0.6	0.60	0.61	0.62	0.63	0.64	0.65	0.67	0.68	0.70	0.71
	0.7	0.73	0.74	0.76	0.77	0.79	0.81	0.83	0.85	0.87	0.89
	0.8	0.91	0.93	0.95	0.93	0.95	0.96	0.98	0.99	1.00	1.00
	0.9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	κ_w	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
$n = 500$	0.0	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32
	0.1	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32
	0.2	0.32	0.33	0.33	0.33	0.33	0.34	0.34	0.34	0.35	0.35
	0.3	0.36	0.37	0.37	0.38	0.37	0.37	0.38	0.38	0.39	0.40
	0.4	0.41	0.41	0.42	0.43	0.44	0.45	0.47	0.48	0.49	0.50
	0.5	0.52	0.53	0.54	0.56	0.58	0.59	0.61	0.63	0.64	0.66
	0.6	0.68	0.70	0.72	0.74	0.76	0.78	0.81	0.83	0.85	0.87
	0.7	0.90	0.88	0.78	0.80	0.83	0.85	0.88	0.91	0.94	0.97
	0.8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	0.9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00