CHAPTER 3

Agreement Coefficients for Nominal Ratings: A Review

OBJECTIVE

This chapter presents a critical review of several agreement coefficients proposed in the literature in the past few decades for analyzing nominal ratings. Among other coefficients, I discuss the Kappa coefficient of Cohen (1960), its meaning and its limitations. The different components of Kappa are teased apart and their influence on the agreement coefficient discussed. I explore the case of two raters and two response categories first before expanding to the more general situation of multiple raters and multiple-item response scales. This chapter also treats the important problem of missing ratings often overlooked in the literature. Figure 3.1 is a flowchart that shows the different agreement coefficients reviewed, the conditions under which they can be used and their equation numbers that provide a convenient way to locate them in this chapter.

Contents

"When you can measure what you are speaking about and express it in numbers, you know something about it. But when you cannot – your knowledge is of meager and unsatisfactory kind. —" .
- Lord Kelvin (1824-1907) -

3.1 The Problem

The objective of this chapter is to present a number of agreement coefficients that have been proposed in the literature for quantifying the extent of agreement among raters when the ratings are data of the nominal type. Such ratings are independent categories, which cannot be ranked neither by order of importance, severity nor any other attribute. Table 3.1 for example shows the distribution of 223 psychiatric patients by diagnosis category and method used to obtain the diagnosis. The first method named "Clinical Diagnosis" (also known as "Facility Diagnosis") is used in a service facility (e.g. public hospital, or a community unit) and does not rely on a rigorous application of research criteria. The second method known as "Research Diagnosis" is based on a strict application of research criteria. Fenning et al. (1994) conducted this study to investigate the extent of agreement between clinical and Research Diagnoses, using the following 4 diagnostic categories:

- Schizophrenia
- Bipolar Disorder
- Depression
- Other

This inter-rater reliability experiment involves two raters and four possible categories into which the patients may be classified. The two raters are the diagnosis methods "Clinical Diagnosis" and "Research Diagnosis." The rating scale is considered nominal because the four categories cannot be ranked, although it is more accurate to state that this study does not consider the ranking of these categories to be of any interest. The basic problem is to quantify the extent to which the two methods agree about the diagnoses they produce.

The most fundamental and intuitive approach to this problem is to consider the percent agreement as an agreement coefficient. The percent agreement is calculated by summing all four diagonal numbers of Table 3.1 and dividing the sum by the total number of patients. That is,

$$\text{Percent Agreement} = (40 + 25 + 21 + 45)/223 = 131/223 = 58.7\%.$$

Several authors have attempted to identify who first proposed this coefficient. There is still a confusion regarding this issue. While some authors refer to the percent agreement as the Osgood's coefficient, others refer to it as the Holsti's coefficient due to Osgood (1959) and Holsti (1969) recommending its use at a given point in

time. There is ample evidence that the percent agreement has been used numerous times well before these works were released. Since the percent agreement is a crude application of the notion of empirical probability that did not require much investigation, my recommendation would be to put this debate to rest so we can move on to other things.

Researchers observed early in the history of inter-rater reliability estimation that two raters may agree for cause following a clear deterministic rating procedure, or they may agree by pure chance. The problem of chance agreement is best seen in a two-category inter-rater reliability experiment, where two raters must assign subjects to a positive and a negative categories. If two raters are unclear about the categorization of a subject and independently decide to make a subjective choice, they still have a chance to agree that is considerably high given the limited number of options they have to chose from. Because this type of agreement is unpredictable and difficult to justify, it is clearly not the way any researcher will want the raters to agree. Therefore, agreement by chance is undesirable since it cannot be seen as evidence that the raters master the rating process. Unfortunately the percent agreement accounts for both types of agreement and can be expected to overstate the "true" extent of agreement among raters. This is the problem that led several authors to propose what is known today as chance-corrected agreement coefficients. The important notion of chance agreement is further discussed in section 3.2.

Psychiatric diagnoses for example, are difficult to make due to the fuzzy boundaries that define various psychiatric disorders. A high degree of consistency between different methods permits each method to validate the other and eventually be used with confidence and interchangeably on a routine basis. We saw in an example earlier that the clinical and research methods yield the same diagnosis on approximately 58.7% of patients. One can assume that some of these agreements did occur by pure chance. An agreement by chance is not a false agreement. It represents a form of gift or bonus that inflates the relative number of subjects in agreement without resulting from the diagnostic methods' inherent properties. Therefore a patient associated with an agreement by chance does not carry useful information regarding the degree of consistency that can be expected from the methods' intrinsic properties. Consequently, the figure 58.7% overestimates the extent of agreement between the two methods.

If we are able to identify all patients that are susceptible to chance agreement, then we could remove them from our pool of study participants before evaluating the percent agreement. But the sole existence of these special patients does not make them identifiable. A patient is associated with an agreement by chance if the processes that led to a particular diagnosis are not an integral part of the methods. However, Table 3.1, which constitutes the basis for our analysis, contains no information regarding the processes behind the diagnoses. Moreover, some of these processes

may even be cognitive and difficult to capture with precision. Still, an inter-rater reliability coefficient will yield a useful measure of the extent to which two methods are concurrent, only if it is corrected for chance agreement. How one defines chance agreement will determine the form a particular inter-rater reliability coefficient will take.

The oldest chance-corrected agreement coefficient mentioned in the literature is likely from Benini (1901). Other early efforts to solve the chance agreement issue include authors such as Guttman (1945), Bennett et al. (1954), Holley and Guilford (1964), Maxwell (1977), Janson and Vegelius (1979), and Brennan and Prediger (1981) who independently developed the same coefficient giving it different names. This simple coefficient, often referred to in the literature as the Brennan-Prediger coefficient is given by the ratio $(p_a - 1/q)/(1 - 1/q)$, where p_a is the percent agreement and q the number of nominal categories in the rating scale.

Brennan and Prediger (1981) recommended their coefficient in the case of two raters and an arbitrary number q of categories, while most authors before recommended it in the simpler case of two raters and two categories only. It can further be extended to the more general case of three raters or more as will be seen in subsequent sections. Holley and Guilford (1964) were the first to formally study this coefficient as a way to compute inter-rater reliability, even though others mentioned it before in various contexts. They named this coefficient the G-Index. These agreement coefficients and many others will be discussed in greater details in subsequent sections.

Some authors have criticized this coefficient under the ground that a practitioner may artificially increase the number of categories. The benefit of this operation would be a smaller chance-agreement probability (i.e. $1/q$), which in turn would increase the magnitude of the agreement coefficient. I believe that this criticism is unfounded. A practitioner who adds dummy categories for the sole purpose of jacking up the agreement coefficient engages in malpractice to obtain an undeserving reward. This is a behavioral problem. Time spent looking for a statistical fix to a problem that stems from a rigged experimental design is not time well spent.

While several inter-rater reliability coefficients have been proposed in the literature since the late forties and early fifties, the Kappa statistic proposed by Cohen (1960) became overtime the most widely-used agreement index of its genre. Despite its popularity, Kappa has many well-documented weaknesses that researchers have been slow to take into consideration when selecting an agreement coefficient. In the next few sections, I will discuss various properties of this coefficient and will highlight some of its shortcomings.

Table 3.1: Distribution of 223 Psychiatric Patients by Type of Psychiatric Disorder and Diagnosis Method.

Clinical Diagnosis	Research Diagnosis				Total
	Schizo	Bipolar	Depress	Other	
Schizo	**40**	6	4	15	65
Bipolar	4	**25**	1	5	35
Depress	4	2	**21**	9	36
Other	17	13	12	**45**	87
Total	65	46	38	74	**223**

3.2 Agreement for two Raters and two Categories

A simple inter-rater reliability study consists of evaluating the extent of agreement between two raters who have each classified for example the same 100 individuals into one of two non-overlapping response categories. To be concrete, I will refer to the two raters as A and B and to the two categories as 1 and 2. Ratings obtained from such a study are often organized in a contingency table such as Table 3.2, which contains fictitious data. This table will be used later in this chapter for illustration purposes. Table 3.3 on the other hand, contains similar agreement data in their abstract form. I will appeal to the abstract agreement table throughout this chapter to describe the computational methods in their general form.

Table 3.2 shows that raters A and B both classified 35 of the 100 subjects into category 1 and 40 of the 100 subjects into category 2. Therefore, both raters agreed about the classification of 75 subjects for a percent agreement of 75%. However, they disagreed about the classification of 25 subjects, classifying 5 into categories 2 and 1 and 20 into categories 1 and 2 respectively. Likewise, using the abstract Table 3.3, I would say that raters A and B agreed on the classification of $n_{11} + n_{22}$ subjects out of a total of n subjects for a percent agreement $(n_{11} + n_{22})/n$. If p_a denotes the percent agreement then its value based on Table 3.2 data is given by:

$$p_a = (35 + 40)/100 = 0.75,$$

and its formula given by:

$$p_a = (n_{11} + n_{22})/n\cdot \tag{3.2.1}$$

It would seem natural to consider 0.75 as a reasonably high extent of agreement between raters A and B. In reality, this number may overstate what one expects the

Table 3.2: Distribution of 100 Subjects by Rater and Category.

Rater A	Rater B		Total
	1	2	
1	35	20	55
2	5	40	45
Total	40	60	100

Table 3.3: Distribution of n Subjects by Rater and Category.

Rater A	Rater B		Total
	1	2	
1	n_{11}	n_{12}	n_{1+}
2	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

inter-rater reliability between A and B to be, due to possible chance agreement as discussed in section 3.1. In this section, I will show how Cohen (1960) adjusted p_a for chance agreement to obtain the Kappa coefficient.

CHANCE-AGREEMENT CORRECTION

The idea of adjusting the percent agreement p_a for chance agreement is often controversial and the definition of what constitutes chance agreement is part of the problem. Rater A for example, ignoring a particular subject's specific characteristics may decide to categorize it randomly[1]. With the number of response categories as small as 2, rater A could still categorize that subject into the exact same group as rater B, creating a lucky agreement that reflects neither the intrinsic properties of the classification system, not rater A's proficiency to use it.

3.2.1 *Cohen's Kappa Definition*

What researchers need, is an approach for measuring agreement beyond chance. To address this problem, Cohen (1960) estimated the expected percent chance agreement (denoted by p_e) and used it to adjust the percent agreement p_a to obtain the Kappa coefficient shown in equation 3.2.3. The percent chance agreement p_e is calculated by summing what Cohen considers to be the two probabilities for the two response categories 1 and 2. Note that the probabilities that raters A and B classify a subject into category 1 are respectively 0.55 and 0.40. These numbers correspond to the raw and column marginal percentages. According to Cohen the two raters are expected to reach agreement on category 1 with probability $0.55 \times 0.40 = 0.22$. Likewise they are expected to reach agreement on category 2 with probability $0.45 \times 0.60 = 0.27$. Consequently, Cohen's percent chance agreement is

[1] I consider a subject categorization to be random if it is not based on any known and predetermined process

given by:

$$p_e = \frac{55}{100} \times \frac{40}{100} + \frac{45}{100} \times \frac{60}{100} = \frac{49}{100} = 0.49.$$

The formula for calculating the percent chance agreement is given by:

$$\begin{aligned} p_e &= p_{1+}p_{+1} + p_{2+}p_{+2} = \frac{n_{1+}}{n} \times \frac{n_{+1}}{n} + \frac{n_{2+}}{n} \times \frac{n_{+2}}{n}, \\ &= p_{1+}p_{+1} + (1 - p_{1+})(1 - p_{+1}), \end{aligned} \tag{3.2.2}$$

where $n_{1+} = n_{11} + n_{12}$ and $n_{+1} = n_{11} + n_{21}$ are the marginal counts and p_{1+} and p_{+1} the associated marginal probabilities. Cohen (1960) defined the Kappa coefficient as follows:

$$\boxed{\widehat{\kappa}_{\text{C}} = \frac{p_a - p_e}{1 - p_e}.} \tag{3.2.3}$$

Although Cohen's original notation for the Kappa was κ (the Greek character "kappa"), I am using a different notation $\widehat{\kappa}_{\text{C}}$ (read "kappa hat-C"). In this new notation, the subscript C is a specific label that identifies Cohen's version of Kappa among other versions to be studied later, κ_{C} (without the hat) represents the true and error-free value of Kappa also known as the estimand and the hat ($\widehat{\ }$) indicates an approximation of the estimand based on observed ratings. The notion of "true" value or estimand reminds us that calculated numbers are always one concrete representation of an abstract (and elusive) reality (some authors will refer to it as a construct) that constitutes our primary interest. These subtleties appeal to more sophisticated statistical concepts in the field of statistical inference, to be discussed in chapter 6.

To understand the meaning of the proposed notation $\widehat{\kappa}_{\text{C}}$, which is further discussed in chapter 6, the reader should remember that the Kappa value is calculated using one specific sample[2] of subjects. Consequently, a different sample of subjects selected by another researcher is expected to lead to a different value of Kappa. One may then wonder whether there exists a "true", fixed and unique value for Kappa. The answer is yes, there is a unique "true" Kappa specific to a predefined universe or population of subjects. The subject population of interest is made up of subjects that participated in the reliability study, as well as all those subjects that could potentially be rated in the future and to whom the researcher wants to extend the findings of the inter-rater reliability experiment. Defining this subject population at the beginning of any reliability study is an important task often overlooked by researchers, but which is essential for calculating the precision of our statistics.

[2] A sample of subjects in this context does not represent a single unit as is often the case in some medical fields(e.g. a blood sample). Instead, it represents the entire pool of subjects that participated in the reliability study.

Kappa's denominator represents the percent of subjects for which one would not expect any agreement by chance, while its numerator according to Cohen (1960) represents *" ... the percent of units in which beyond-chance agreement occurred ..."* Cohen (1960) sees Kappa as a measure of *"... the proportion of agreement after chance agreement is removed from consideration ..."* I will show in chapter 4 that this fundamental goal set by Cohen for Kappa can be achieved with alternative and more efficient methods.

It follows from Table 3.2 data and from the values of p_a and p_e obtained earlier in this section that the inter-rater reliability between raters A and B as measured by Kappa is given by:

$$\widehat{\kappa}_{\text{C}} = \frac{0.75 - 0.49}{1 - 0.49} \cong 0.51.$$

That is, the Kappa-based extent of agreement between raters A and B is approximately equal to 0.51. This represents a "Moderate" agreement level between two raters according to the Landis-Koch benchmark scale (Landis and G., 1977). Although widely-used by researchers, this benchmark scale is not without flaws and is further discussed in chapter 7.

3.2.2 *What is Chance Agreement?*

While the idea of correcting agreement coefficients for chance agreement is justified, the very notion of chance agreement introduced in the previous section is loosely defined. When we claim that two raters A and B have agreed by chance, what do we really mean? Does p_e (Cohen's percent chance agreement) measure what it is supposed to measure? These are two important questions that need to be addressed.

- By claiming that raters A and B have agreed by chance about the classification of a subject, do we mean that one of the two raters not knowing in which category the subject belongs, took a chance by randomly classifying it (perhaps with an equal probability of 0.5 (i.e. the 50:50 rule)) into one of the two possible categories? This view ties the notion of chance agreement to that of random rating.

- Rather than using the 50:50 rule when randomly categorizing a subject, one may consider the marginal classification probabilities p_{1+} and p_{+1} as being the raters' propensity for classifying a subject into category 1. Even if the rating is random, raters A and B would choose category 1 with probabilities p_{1+} and p_{+1} respectively. They will then agree by chance if one of them performs a random classification according to the observed marginal probabilities. The classification can be seen as having been carried out either independently of the subject's specific characteristics, or following an unknown judgmental process with no apparent logic connecting the subject to the rating.

In both situations described above one of the raters must perform a random classification for concurrence to be considered chance agreement. Based on the second scenario, Cohen (1960) evaluated the chance-agreement probability as shown in equation 3.2.2. This equation could be problematic for the following reason:

> *The expression* $p_{1+}p_{+1} + (1 - p_{1+})(1 - p_{+1})$ *represents a probability of agreement between raters A and B only if the ratings are known to be independent*[3]*. In case of independence, the percent agreement* p_a *and the percent chance agreement* p_e *will be very close. If the ratings are not independent then the expression* $p_{1+}p_{+1}+(1-p_{1+})(1-p_{+1})$ *does not have any particular meaning and does not represent a measure of agreement. Using it in the Kappa equation may yield unpredictable results.*

Krippendorff (2011) argues that Cohen's percent chance agreement is based on the concept of statistical independence, which in his opinion *"... is only marginally related to how units are coded and data are made and does not yield valid coefficients for assessing the reliability of coding processes ..."*. One of the few instances in statistical science where both expressions p_a ("observed proportion" of agreement) and p_e ("expected proportion" of agreement due to chance) are part of the same equation, occurs when testing the statistical hypothesis of independence between two events with the Chi-Square test. In this case, the two expressions are used to define the test statistic, which does not represent any particular metric. Instead, the role of the test statistic is to determine whether the difference between observed and expected values under the hypothesis of independence, is sufficiently large to exclude the possibility that it may have been caused by sampling variation alone. I will further discuss the limitations of Kappa in section 3.5.

Although this section focuses on simple reliability experiments where two raters classify subjects into two distinct categories, many experiments in practice use more categories. This generalization is discussed in section 3.3. Moreover, the treatment of missing ratings is done in the more general framework of 3 raters or more. The results presented in that context apply to the case of 2 raters as well (see section 3.4).

3.2.3 *Scott's Pi Coefficient*

About 5 years before Cohen's Kappa was published, Scott (1955) recommended the use of an agreement coefficient named Pi (Scott used the Greek character

[3]Note that two ratings from two raters A and B are independent if the knowledge of one rating makes the other neither more probable nor less probable. This may be the case for a small percent of subjects only. If two raters have high agreement, then for the majority of subjects, the knowledge of one rating provides a strong indication of what the other rating is (they are likely to be the same).

π to designate his coefficient). Scott's coefficient too is based on the same percent agreement of equation 3.2.1 and on a new percent chance agreement that is calculated with Table 3.2 data as follows:

$$p_e^2 = \left(\frac{(55/100) + (40/100)}{2}\right)^2 + \left(\frac{(45/100) + (60/100)}{2}\right)^2,$$
$$= (0.95/2)^2 + (1.05/2)^2 = 0.5013\cdot$$

The abstract equation based on Table 3.3 is given by $p_e = \widehat{\pi}_1^2 + (1 - \widehat{\pi}_1)^2$, where $\widehat{\pi}_1 = (p_{1+} + p_{+1})/2$ and is defined by Scott (1955) as the frequency with which category 1 is used by coders[4]. This formulation of the percent chance agreement was later criticized by Cohen (1960) as one that assumes a unique propensity for classification into a particular category for all coders. Scott did not explicitly make such an assumption. Instead he was interested in the frequency of use of each category by either rater.

Scott's agreement coefficient is formally defined as follows:

$$\widehat{\kappa}_S = \frac{p_a - p_e}{1 - p_e}, \text{ where } p_e = \widehat{\pi}_1^2 + (1 - \widehat{\pi}_1)^2. \tag{3.2.4}$$

Using Table 3.2 data, this agreement coefficient is obtained as, $\widehat{\kappa}_S = (0.75 - 0.5013)/(1 - 0.5013) = 0.4988$. Scott's Pi coefficient has limitations that have been abundantly documented in the literature and are known to be similar to those of Cohen's kappa. These limitations are discussed in section 3.5 with a particular focus on the Kappa coefficient. The main problem with Scott's Pi coefficient revolves around the calculation of the percent chance agreement p_e, which appears to be disconnected from experimental facts. Does p_e describe a phenomenon that occurred during the rating process and which must be subtracted from the percent agreement p_a? It is well conceivable that some agreements did happen by pure chance, but certainly not all of them. This is one of the issues that led some authors to raise doubts about the quality of Scott's coefficient.

3.2.4 *Krippendorff's Alpha Coefficient*

Krippendorff (1970, 2012) proposed an agreement coefficient named α (read "alpha"), which is often used by researchers in the field of communication. The procedure for computing Krippendorff's alpha is often described in terms of coincidence tables and difference functions (see Hayes and Krippendorff, 2007) . I will stay away

[4]The frequency is denoted as $\widehat{\pi}_1$ (read "Pi Hat One"), which is the Greek character π with a hat. The hat indicates that this quantity is an estimation from a sample and is subject to sampling errors. π_1 would be the "true" and unknown frequency.

from these two concepts and replace them with the more common notations and concepts along the lines of Cohen (1960, 1968).

It is essential to realize that Krippendorff's alpha is solely based on subjects that are rated by two raters or more. All subjects rated by a single rater must be eliminated upfront before the calculations begin. For simple data such as described in Tables 3.2 and 3.3, this coefficient is fairly simple to describe. Let $\varepsilon_n = 1/(2n)$, where n is the number of subjects rated by both raters. Krippendorff's coefficient is calculated as follows:

$$\alpha_{\text{K}} = (p'_a - p_e)/(1 - p_e), \text{ where } \begin{cases} p_e &= \widehat{\pi}_1^2 + (1 - \widehat{\pi}_1)^2, \\ p'_a &= (1 - \varepsilon_n)p_a + \varepsilon_n, \end{cases} \tag{3.2.5}$$

with p_a being the percent agreement of equation 3.2.1. Note that the expression defining ε_n will be different when 3 raters or more are involved. Moreover, one may get the false impression that the percent chance agreement associated with Krippendorff's alpha is identical to that of Scott's Pi coefficient. This is true only if your dataset does not contain missing ratings. With missing ratings, the two quantities will be different. Krippendorff's percent chance agreement uses only subjects that are rated by both raters, while Scott's percent chance agreement uses all subjects including those rated by a single rater.

Equation 3.2.5 shows that Krippendorff's alpha and Scott's Pi are almost identical, with the only exception being the percent agreement. Krippendorff's version of the percent agreement is a weighted average of the observed percent agreement and its maximum value of 1, which always makes it higher than the observed percent agreement. If the number of subjects is limited to two, then Krippendorff's percent agreement will be $p'_a = 0.75 \times p_a + 0.25$. It is because of this weighting scheme that Krippendorff alpha is often said to apply a small-sample correction. This essentially means that when the number of subjects is small then the percent agreement is adjusted upwards and the magnitude of the adjustment decreases as the number of subjects increases. This adjustment becomes almost insignificant as soon as the number of subjects reaches a modest size as 10. How critical is this adjustment is unclear. The need for such an adjustment and its potential benefits have not been documented. If the number of subjects is small then why do we even need to correct the percent agreement? If we do need such a correction then should the adjustment be done upwards or downwards? All these are unanswered questions?

3.2.5 *Gwet's AC_1 Coefficient*

Gwet (2008a) recommended an agreement coefficient named AC_1, which was developed to overcome many of the limitations associated with Cohen's Kappa.

Kappa's limitations were overcome to a large extent as shown by Wongpakaran et al. (2013). A detailed discussion of these limitations is presented in section 3.5 and a more elaborate description of Gwet's AC_1 can be found in chapter 5. What I like to do here is to provide a very brief review of AC_1 for simple inter-rater reliability experiments where the number of raters and the number of categories are both limited to two.

Gwet's AC_1 is based on the same percent agreement p_a of equation 3.2.1 and on a new percent chance agreement calculated with Table 3.2 data as $p_e = 2 \times 0.475 \times (1 - 0.475) = 0.49875$, where 0.475 is the probability that a subject is classified into category 1 by a random rater.

Gwet's AC_1 coefficient, denoted here by $\widehat{\kappa}_{\text{G}}$ is formally defined as,

$$\widehat{\kappa}_{\text{G}} = \frac{p_a - p_e}{1 - p_e}, \text{ where } p_e = 2\widehat{\pi}_1(1 - \widehat{\pi}_1). \tag{3.2.6}$$

Using Table 3.2 data, this agreement coefficient is obtained as, $\widehat{\kappa}_{\text{G}} = (0.75 - 0.49875)/(1 - 0.49875) = 0.5012$. The percent chance agreement of equation 3.2.6 is actually calculated as $\left[\widehat{\pi}_1(1 - \widehat{\pi}_1) + \widehat{\pi}_2(1 - \widehat{\pi}_2)\right]/(2 - 1)$ and takes into consideration the fact that $\widehat{\pi}_1(1 - \widehat{\pi}_1) = \widehat{\pi}_2(1 - \widehat{\pi}_2)$. The number 2 in the denominator represents the number of categories. Readers interested in learning more about this coefficient, its merits and motivations are invited to read chapter 5.

3.2.6 *G-Index*

The G-index is the simplest chance-corrected agreement coefficient initially introduced by Holley and Guilford (1964), and later generalized to three categories or more by Brennan and Prediger (1981). It is based on the same percent agreement of equation 3.2.1. However, the percent chance agreement is simply 1/2, where 2 represents the number of categories used in the experiment.

The Holley-Guilford G-index denoted here by $\widehat{\kappa}_2$ is formally defined as,

$$\widehat{\kappa}_2 = \frac{p_a - 0.5}{1 - 0.5}. \tag{3.2.7}$$

Based on Table 3.2 data, this agreement coefficient is calculated as $\widehat{\kappa}_2 = (0.75 - 0.5)/(1 - 0.5) = 0.5$. The magnitude of the Holley-Guilford coefficient is generally reasonable compared to that of the percent agreement.