Scientific Research Publishing

Open Journal of Statistics, 2024, 14, 481-491
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Interrater Reliability Estimation via Maximum Likelihood for Gwet's Chance Agreement Model

Alek M. Westover[1], Tara M. Westover[2], M. Brandon Westover[2*]

[1]Massachusetts Institute of Technology, Boston, MA, USA
[2]Harvard Medical School, Beth Israel Deaconess Medical Center, Boston, MA, USA
Email: *bwestove@bidmc.harvard.edu

How to cite this paper: Westover, A.M., Westover, T.M. and Westover, M.B. (2024) Interrater Reliability Estimation via Maximum Likelihood for Gwet's Chance Agreement Model. *Open Journal of Statistics*, **14**, 481-491.
https://doi.org/10.4236/ojs.2024.145021

Received: August 29, 2024
Accepted: October 25, 2024
Published: October 28, 2024

Abstract

Interrater reliability (IRR) statistics, like Cohen's kappa, measure agreement between raters beyond what is expected by chance when classifying items into categories. While Cohen's kappa has been widely used, it has several limitations, prompting development of Gwet's agreement statistic, an alternative "kappa"statistic which models chance agreement via an "occasional guessing" model. However, we show that Gwet's formula for estimating the proportion of agreement due to chance is itself biased for intermediate levels of agreement, despite overcoming limitations of Cohen's kappa at high and low agreement levels. We derive a maximum likelihood estimator for the occasional guessing model that yields an unbiased estimator of the IRR, which we call the maximum likelihood kappa (κ_{ML}). The key result is that the chance agreement probability under the occasional guessing model is simply equal to the observed rate of disagreement between raters. The κ_{ML} statistic provides a theoretically principled approach to quantifying IRR that addresses limitations of previous κ coefficients. Given the widespread use of IRR measures, having an unbiased estimator is important for reliable inference across domains where rater judgments are analyzed.

Keywords

Interrater Reliability, Agreement, Reliability, Kappa

1. Introduction

Interrater reliability (IRR) (also known as "kappa" (κ)) statistics, are used to measure agreement between two raters or coders classifying items into mutually exclusive categories. κ statistics are widely used in fields such as psychology and medicine to evaluate the reliability or consistency of expert judgments [1].

Simply calculating the percentage of cases where raters agree does not account for the possibility that some agreement occurs by chance. κ is designed to measure the the degree of agreement between raters *beyond* what is expected by chance. Assume two raters independently classify N cases into categories + and −, and denote by N_a the number of cases on which they agree. Assume N_c agreements occur by chance, and the rest N_k are due to knowledge (not due to chance), so that $N_a = N_c + N_k$. The number of cases remaining after subtracting chance agreements is $N - N_c$. Thus the percentage of the observed agreement N_a in excess of chance agreement is:

$$\kappa = \frac{N_a - N_c}{N - N_c} = \frac{N_k}{N - N_c} = \frac{P_a - P_c}{1 - P_c},$$

where $P_a = N_a / N$ denotes the observed percent agreement, and $P_c = N_c / N$ is the percent agreement due to chance. P_a is observed, whereas P_c must be estimated.

Several approaches have been proposed to estimate the probability of chance agreement. The approach used most commonly in the past (Cohen's κ) has recently fallen under criticism [2] [3], leading to a new approach (Gwet's κ) which has been gained popularity over the past several years [1] [4]-[7]. However, we show that the new approach is biased. We demonstrate an unbiased approach to estimating κ based on maximum likelihood estimation.

2. Cohen's Kappa and Its Limitations

Historically, the most commonly used κ statistic has been Cohen's κ [8] [9], which quantifies interrater reliability for two raters applying binary ratings. Other approaches are discussed at length in [10]-[12].

Cohen proposed calculating the probability of chance agreement P_c based on an 'always guess' model. Suppose two raters A and B independently assign N items to two categories, + and −. Let the numbers of items assigned to each category be N_A^+, N_A^-, N_B^+, N_B^-, and the number of items on which they agree be N_a. Now consider what percentage of cases raters A and B would be expected to agree on if they assigned the same numbers of items to each category as they do in the observed data, but made the assignments at random ("guessing"). Under this model, A and B classify items as + with probabilities $p_A^+ = N_A^+ / N$, $p_B^+ = N_B^+ / N$, and as − with probabilities $p_A^- = N_A^- / N$, $p_B^- = N_B^- / N$. Any agreements under this model occur by chance, with probability

$$P_c = p_A^+ p_B^+ + p_A^- p_B^-.$$

Critiques of Cohen's Model

Two main criticisms have been raised against Cohen's κ. First, Cohen's κ produces "paradoxical" results under certain circumstances [2] [10] [11] [13]: high levels of observed agreement can accompany a low κ value. This happens because Cohen's κ depends only on the rates of ratings in the data. Thus, if raters

A and B score most cases as class +, it may be because they correctly recognize that most cases are +, yet Cohen's κ cannot give credit for agreement due to expertise. This problem is most pronounced when the proportion of classes in the data deviates from 50% [12].

Second, some authors [12] [14] dispute the idea that κ "takes into account" chance agreement. Truly doing this requires a realistic model of how chance affects rater decisions; Cohen's 'always guess' model is unrealistic as a model of how raters behave. For this reason κ can be misleading in situations such as the diagnosis of rare diseases. In these scenarios, κ tends to underestimate agreement on the rare category [15]. κ is thus considered an overly conservative measure of agreement [16].

3. Gwet's Kappa: An Improved Model of Chance Agreement

Gwet proposed an alternative to Cohen's κ, which we call Gwet's κ (also known as AC1 (Agreement Coefficient 1)) that addresses the limitations discussed above [12]. Gwet's key contribution was a more realistic model of chance agreement, P_c, which we call the "occasional guessing" model. Because this model addresses the limitations of Cohen's κ, Gwet's κ has been increasingly adopted in studies of IRR [1] [4]-[7]. However, as we show below, Gwet's κ also has important limitations. Specifically, the formula Gwet proposed for estimating κ is biased.

3.1. The "Occasional Guessing" Model for Chance Agreement

Gwet suggested that a more realistic model for how chance agreement occurs is:

1) Cases are easy or hard. Raters always classify easy cases correctly, and for hard cases, they guess with equal probability. Thus, for hard cases, the probability of agreement is 1/2.

2) The fraction of hard cases is r.

3.2. Theoretical Value of κ under the Occasional Guessing Model

Using this model, we can calculate the theoretical true value of κ, denoted κ^*. For any case evaluated by two raters consider the following events: A = {Raters agree}, and R = {the case is hard: raters guess randomly}. Then the probability of agreement due to chance (arising out of guessing) for any case is

$$P_c = P(A,R) = P(R)P(A \mid R) = r/2$$

The overall probability of agreement is

$$\begin{aligned} P_a &= P(A) \\ &= P(A,R) + P(A,\bar{R}) \\ &= P(R)P(A \mid R) + P(\bar{R})P(A \mid \bar{R}) \\ &= r/2 + (1-r) \\ &= 1 - r/2. \end{aligned}$$

Thus, the expected proportion of beyond-chance agreement is

$$\kappa^* = \frac{P_a - P_c}{1 - P_c} = \frac{1 - r}{1 - r/2}.$$

We note that r can also be expressed in terms of κ, as

$$r = \frac{1 - \kappa^*}{1 - \kappa^*/2}.$$

It is easy to check that $0 < \kappa < P_a$. Also, noting that $\kappa = \kappa(r)$, we observe that for high and low values of r, we get $\kappa(0) = 1$, $\kappa(1) = 0$.

Any estimate of κ whose expected value deviates from the theoretical value κ^* is said to be *biased*. We next consider Gwet's proposal for estimating κ, and will show that it is biased in some important settings.

3.3. Gwet's Formula for the Probability of Chance Agreement

Gwet proposed a formula for $r = P(R)$ based on the following heuristic argument. Consider the random variable

$$X_+ = \begin{cases} 1 & \text{if a rater classifies a given case as +} \\ 0 & \text{otherwise} \end{cases}$$

The variance of X_+ is $\mathrm{Var}(X_+) = \pi_+(1 - \pi_+)$, where π_+ is the average rate at which raters assign cases to the "+" category. The maximum possible variance for classification is reached when rating is done completely at random, with each category assigned with probability 1/2, in which case the variance is $\mathrm{Var}_{\max} = 1/2(1 - 1/2) = 1/4$. Gwet suggested that a reasonable measure of the randomness with which raters choose the + category is the ratio of the observed choice variance to the maximal possible variance, *i.e.* $P(R) \approx \mathrm{Var}(X_+)/V_{\max}$, thus:

$$r = P(R) = \frac{\pi_+(1 - \pi_+)}{1/2(1 - 1/2)} = 4\pi_+(1 - \pi_+),$$

This leads to chance agreement probability of

$$P_c = r/2 = 2\pi_+(1 - \pi_+),$$

which can be substituted into $\kappa = (P_a - P_c)/(1 - P_c)$.

3.4. Gwet's κ Is Biased

Gwet showed that, when considered from the point of view of the 'occasional guessing' model of chance agreement, Cohen's κ and several other well-known κ and κ-like statistics for interrater agreement are biased, particularly at high levels of agreement [1] [12]. By contrast, Gwet's formula is accurate (nearly unbiased, *i.e.* $\kappa \approx \kappa^*$) when agreement between raters P_a is high or low, overcoming a key limitation of Cohen's κ [1] [12]. This is easy to show: When agreement is high, $P_a \approx 1$, we have $\kappa \approx (1 - P_c)/(1 - P_c) = 1$, regardless of P_c. When agreement is low (both raters guessing all the time, $r = 1$), agreement occurs in

approximately half the cases, $P_a \approx 1/2$, approximately half of the ratings are positive, $\pi_+ \approx 1/2$, and $P_c = 2(1/2)(1-1/2) = 1/2$, and $\kappa = (1/2-1/2)/(1-1/2) = 0$.

However, for intermediate levels of agreement, Gwet's formula is biased. We show this by expressing π_+ in terms of *r*, substituting into Gwet's formula for P_c, then comparing this with the true value $P_c = r/2$. The proportion of + ratings is the sum of the proportions of + ratings on hard cases, *r*/2, and easy cases, $(1-r)q$, where $q \in [0,1]$ is the proportion of easy cases whose true rating is +. Thus $\pi_+ = r/2 + (1-r)q$, and Gwet's formula gives $P_c = 2\left(r/2 + (1-r)q\right)\left(1-r/2-(1-r)\right)q$. The deviation of Gwet's formula for P_c from the true value *r*/2 is

$$\Delta P_c = 2\left(r/2 + (1-r)q\right)\left(1-r/2-(1-r)q\right) - r/2 = r/2 - r^2/2.$$

Note that this bias does not depend on *q*. Figure 1(A) & Figure 1(B) illustrate the bias and 95% confidence intervals for 2 raters scoring $N = 100$ cases, where $q = 0.2$, over the entire range of possible true values κ^* of the underlying IRR.

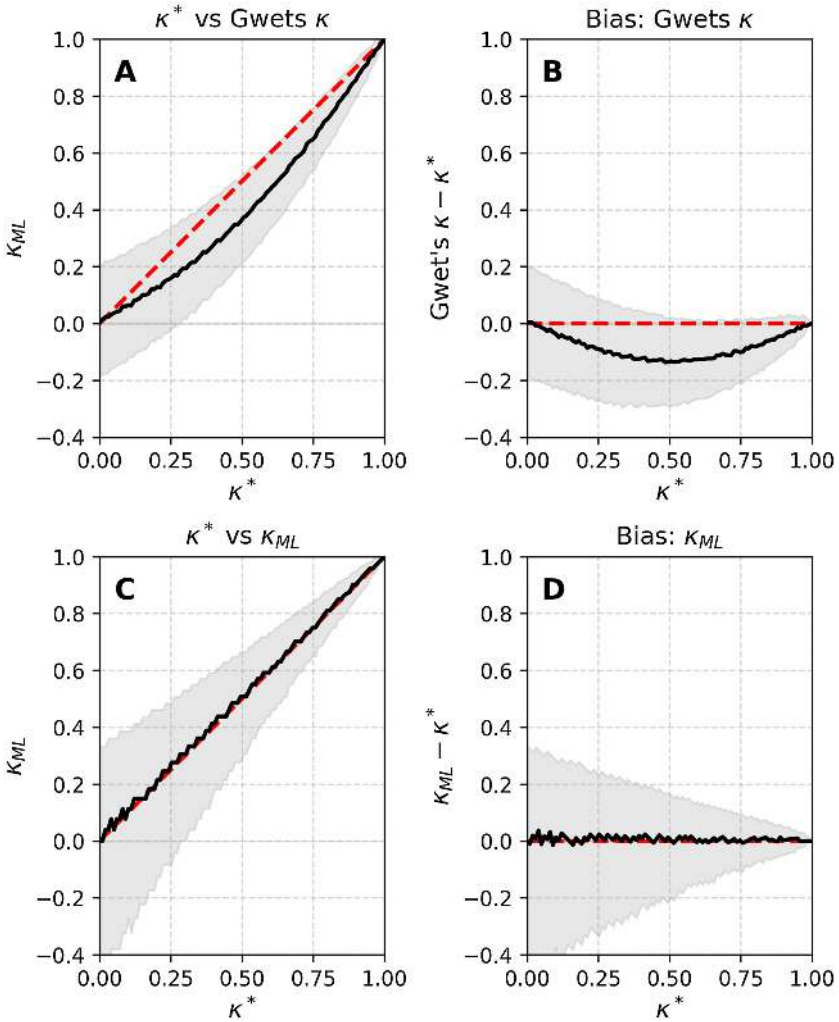

Figure 1. (A) True $\kappa = \kappa^*$ vs Gwet's κ. (B) Bias (Gwet's $\kappa - \kappa^*$). (C) κ^* vs κ_{ML}. (D) Bias ($\kappa_{ML} - \kappa^*$).

4. Maximum Likelihood Estimation of $P(R)$

Here we present a direct approach to estimating $P(R)=r$ in Gwet's occasional guessing model. Unlike Gwet's κ, the ML κ is not based on a heuristic approximation. Rather, we derive κ_{ML} by writing down the likelihood of the observed data under the occasional guessing model and then solving for the r that maximizes that likelihood.

Let $X=[X_1,X_2,\cdots,X_N]$ represent the agreement and disagreements for the N cases, where $X_i=0$ indicates disagreement and $X_i=1$ indicates agreement. When event R occurs (random guessing), we have $P(X_i=0\,|\,R)=P(X_i=1\,|\,R)=1/2$. For easy cases, raters are not guessing (*i.e.* $\bar{R}$ occurs), and we have $P(X_i=0\,|\,\bar{R})=0$, $P(X_i=1\,|\,\bar{R})=1$. The probability that raters guess is $P(R)=r$. The probabilities for X_i conditional on r are

$$P(X_i=0\,|\,r)=P(R)P(X_i=0\,|\,R)+P(\bar{R})P(X_i=0\,|\,\bar{R})=r/2$$

$$P(X_i=1\,|\,r)=P(R)P(X_i=1\,|\,R)+P(\bar{R})P(X_i=1\,|\,\bar{R})=1-r/2$$

The likelihood function for the data is: $P(X\,|\,r)=\prod_{i=1}^{N}P(X_i\,|\,r)$, so the log-likelihood is $L(X\,|\,r)=\sum_{i=1}^{N}\log P(X_i\,|\,r)$. Splitting the sum into N_d terms in which they disagree ($X_i=0$) and N_a terms in which they agree ($X_i=1$), we get

$$\begin{aligned}L(X\,|\,r)&=N_d\log P(X=0\,|\,r)+N_a\log P(X=1\,|\,r)\\&=N_d\log\frac{r}{2}+N_a\log(1-r/2)\end{aligned}$$

Taking the derivative of $L(X\,|\,r)$ with respect to r, setting it equal to zero, and solving, we get:

$$\frac{\partial}{\partial r}L(X,r)=N_d/r_{ML}-\frac{1}{2}N_a/(1-r_{ML}/2)=0$$

$$\Rightarrow r_{ML}=\frac{2N_d}{N},$$

where $N=N_d+N_a$. Note that $N_d/N=P_d$ the probability of disagreement.

This result makes sense: Given that the probability of agreement when raters guess is 1/2, the best estimate from the data of the number of times at least one rater was in fact guessing is twice the number of observed disagreements.

From the above calculation it follows that the estimated probability of agreement due to chance is

$$P_c=P(R)P(A\,|\,R)=r_{ML}/2=N_d/N.$$

κ_{ML} Is Unbiased

We now show that the expected value of the ML estimator for κ is equal to the theoretical value, hence κ_{ML} is an unbiased estimator of κ^*.

Recall that $r_{ML}=(2N_d)/N$ is the probability of chance agreement used in

calculating κ_{ML}, where N_d is the number of disagreements observed between the two raters performing binary assignments. We can rewrite this as $N_d = \mathrm{E}\left[\sum_{i=1}^{N} \overline{X_i}\right]$, since $X_i = 0$ denotes disagreements, and $X_i = 1$ in cases of agreement. Thus,

$$\begin{aligned}\mathrm{E}[r_{ML}] &= \mathrm{E}[2N_d/N] \\ &= \frac{2}{N}\mathrm{E}\left[\sum_{i=1}^{N} \overline{X_i}\right] \\ &= \frac{2}{N} N \cdot (r/2) \\ &= r\end{aligned}$$

Consequently,

$$\mathrm{E}[\kappa_{ML}] = \frac{1-\mathrm{E}[r_{ML}]}{1-\mathrm{E}[r_{ML}]/2} = \frac{1-r}{1-r/2} = \kappa^*$$

Figure 1(C) & Figure 1(D) illustrate the estimation of κ_{ML} in a case with $N = 100$ cases scored by 2 raters, including bootstrap estimates of the 95% confidence intervals.

5. Variance of κ_{ML}

We now compute the variance of our estimate of r_{ML}. The key computation is computing the second moment of N_d.

$$\begin{aligned}\mathrm{E}\left[N_d^2\right] &= \mathrm{E}\left[\left(\sum_i \overline{X_i}\right)^2\right] \\ &= \sum_{i \neq j} P\left(\overline{X_i X_j}\right) + \sum_i P\left(\overline{X_i}\right) \\ &= \left(N^2 - N\right) r^2/4 + Nr/2.\end{aligned}$$

Thus,

$$\begin{aligned}\mathrm{Var}[r_{ML}] &= \frac{4}{N^2}\left(\mathrm{E}\left[N_d^2\right] - \mathrm{E}[N_d]^2\right) \\ &= \frac{4}{N^2}\left(\left(N^2 - N\right) r^2/4 + Nr/2 - (Nr/2)^2\right) \\ &= \frac{r(2-r)}{N}.\end{aligned}$$

Let $f(r) = \frac{1-r}{1-r/2}$. The maximum derivative of f over $r \in [0,1]$ is 2. Thus, we have for all $\epsilon > 0$, $r \in [\epsilon, 1]$:

$$\left|f(r) - f(r-\epsilon)\right| \leq 2\epsilon.$$

In other words, a confidence interval for $r \in [r_0 - \delta, r_0 + \delta]$ translates into a confidence interval for κ_{ML} which is $\kappa_{ML} \in \left[f(r_0) - 2\delta, f(r_0) + 2\delta\right]$. Confidence intervals can also be calculated numerically using bootstrapping, as shown in Figure 1.

6. Multiple Categories

The preceding sections have dealt with the case of classifying into 2 categories. We can analogously derive κ_{ML} and r in the case where there is instead an arbitrary number, n, of classes. To do this, we generalize the "occasional guessing" model so that, for hard cases, raters guess all n classes with equal probability. Under this model, the probability of agreement by guessing is

$$P_c = P(A,R) = P(R)P(A \mid R) = r/n,$$

and the overall probability of agreement is

$$\begin{aligned} P_a &= P(A,R) + P(A,\bar{R}) \\ &= P(R)P(A \mid R) + P(\bar{R})P(A \mid \bar{R}) \\ &= r/n + (1-r)\cdot 1 \\ &= 1 + \frac{r}{n}(1-n). \end{aligned}$$

Now, to find the theoretical κ^* in terms of r,

$$\begin{aligned} \kappa^* &= \frac{P_a - P_c}{1-P_c} \\ &= \frac{1 + \frac{r}{n}(1-n) - r/n}{1 - r/n} \\ &= \frac{1-r}{1-r/n}. \end{aligned}$$

Next we derive the ML estimator of r. Let $X = [X_1, X_2, \cdots, X_N]$ represent the agreement and disagreements for the N cases, where $X_i = 0$ indicates disagreement and $X_i = 1$ indicates agreement. When event R occurs (random guessing), we have $P(X_i = 0 \mid R, r) = P(X_i = 1 \mid R, r) = 1/2$. When neither rater guesses (*i.e.* event $\bar{R}$ occurs), we have $P(X_i = 0 \mid \bar{R}, r) = 0$, $P(X_i = 1 \mid \bar{R}, r) = 1$. The probability that raters guess randomly is $P(R) = r$. The probabilities for X_i conditional on r are

$$P(X_i = 0 \mid r) = 1 - P_a = \frac{r}{n}(n-1)$$

$$P(X_i = 1 \mid r) = P_a = 1 + \frac{r}{n}(1-n)$$

Now, to find κ_{ML}, we maximize the likelihood function for the data $P(X \mid r) = \prod_{i=1}^{N} P(X_i \mid r)$, or the log-likelihood $L(X \mid r) = \sum_{i=1}^{N} \log P(X_i \mid r)$. Splitting the sum into N_d terms with $X_i = 0$ and N_a terms with $X_i = 1$, we get

$$\begin{aligned} L(X \mid r) &= N_d \log P(X = 0 \mid r) + N_a \log P(X = 1 \mid r) \\ &= N_d \log\left(\frac{r}{n}(n-1)\right) + N_a \log\left(1 + \frac{r}{n}(1-n)\right). \end{aligned}$$

Taking the derivative with respect to r, setting it equal to zero, and solving, we

get:

$$\frac{\partial}{\partial r}L(X,r) = N_d \cdot \frac{n}{r(n-1)} \cdot \frac{n-1}{n} + N_a \frac{1}{1+\frac{r}{n}(1-n)} \cdot \frac{n}{1-n}$$

$$= \frac{N_d}{r} + \frac{N_a}{\frac{n}{1-n}+r} = 0$$

$$\Rightarrow r_{ML} = \frac{N_d}{N}\frac{n}{n-1}$$

where $N = N_d + N_a$.

7. Conclusions

We have presented a maximum likelihood approach to estimating the chance agreement probability P_c in Gwet's "occasional guessing" model of interrater agreement. Our estimator, κ_{ML}, is derived directly from the likelihood function of the data under this model, rather than relying on heuristic approximations as in Gwet's κ.

We have shown that the maximum likelihood estimator r_{ML} for the probability of guessing r is simply twice the observed disagreement rate between raters. Consequently, the chance agreement probability estimate P_c used in κ_{ML} is the observed disagreement rate. We have also generalized this result to the case of raters scoring cases that can belong to multiple classes.

A key advantage of κ_{ML} is that it is an unbiased estimator of the true value of κ predicted by the occasional guessing model. In contrast, we have demonstrated that Gwet's formula for P_c, while overcoming certain limitations of Cohen's κ, is itself biased for intermediate levels of agreement.

We have also provided the variance of the κ_{ML} estimator, which can be used to construct confidence intervals. The variance depends on both the true value of r and the sample size N, decreasing as N increases as expected for a consistent estimator.

In summary, κ_{ML} provides a principled approach to estimating chance agreement in the occasional guessing model, addressing limitations of previous κ statistics. As the use of interrater reliability measures continues to grow across fields, having an unbiased estimator is important for obtaining reliable inferences from data.

Author Contributions

Conceptualization, A.M.W., T.M.W. and M.B.W.; methodology, A.M.W., T.M.W. and M.B.W.; writing—original draft preparation, A.M.W., T.M.W. and M.B.W.; writing—review and editing, A.M.W., T.M.W. and M.B.W.; supervision, M.B.W. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The code that supports the findings of this study is available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Gwet, K.L. (2014) Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement among Raters. Advanced Analytics, LLC.

[2] Cicchetti, D.V. and Feinstein, A.R. (1990) High Agreement but Low Kappa: II. Resolving the Paradoxes. *Journal of Clinical Epidemiology*, **43**, 551-558. https://doi.org/10.1016/0895-4356(90)90159-m

[3] Feinstein, A.R. and Cicchetti, D.V. (1990) High Agreement but Low Kappa: I. The Problems of Two Paradoxes. *Journal of Clinical Epidemiology*, **43**, 543-549. https://doi.org/10.1016/0895-4356(90)90158-l

[4] Wongpakaran, N., Wongpakaran, T., Wedding, D. and Gwet, K.L. (2013) A Comparison of Cohen's Kappa and Gwet's AC1 When Calculating Inter-Rater Reliability Coefficients: A Study Conducted with Personality Disorder Samples. *BMC Medical Research Methodology*, **13**, Article No. 61. https://doi.org/10.1186/1471-2288-13-61

[5] Ohyama, T. (2020) Statistical Inference of Gwet's AC1 Coefficient for Multiple Raters and Binary Outcomes. *Communications in Statistics—Theory and Methods*, **50**, 3564-3572. https://doi.org/10.1080/03610926.2019.1708397

[6] Jimenez, A.M. and Zepeda, S.J. (2020) A Comparison of Gwet's AC1 and Kappa When Calculating Inter-Rater Reliability Coefficients in a Teacher Evaluation Context. *Journal of Education Human Resources*, **38**, 290-300. https://doi.org/10.3138/jehr-2019-0001

[7] Gaspard, N., Hirsch, L.J., LaRoche, S.M., Hahn, C.D. and Westover, M.B. (2014) Interrater Agreement for Critical Care EEG Terminology. *Epilepsia*, **55**, 1366-1373. https://doi.org/10.1111/epi.12653

[8] Cohen, J. (1960) A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, **20**, 37-46. https://doi.org/10.1177/001316446002000104

[9] Cohen, J. (1968) Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit. *Psychological Bulletin*, **70**, 213-220. https://doi.org/10.1037/h0026256

[10] Gwet, K. (2002) Kappa Statistic Is Not Satisfactory for Assessing the Extent of Agreement between Raters. *Statistical Methods for Inter-Rater Reliability Assessment*, **1**, 1-6.

[11] Gwet, K. (2002) Inter-Rater Reliability: Dependency on Trait Prevalence and

Marginal Homogeneity. *Statistical Methods for Inter-Rater Reliability Assessment*, **2**, 1-9.

[12] Gwet, K.L. (2008) Computing Inter-Rater Reliability and Its Variance in the Presence of High Agreement. *British Journal of Mathematical and Statistical Psychology*, **61**, 29-48. https://doi.org/10.1348/000711006x126600

[13] Byrt, T., Bishop, J. and Carlin, J.B. (1993) Bias, Prevalence and Kappa. *Journal of Clinical Epidemiology*, **46**, 423-429. https://doi.org/10.1016/0895-4356(93)90018-v

[14] Uebersax, J.S. (1987) Diversity of Decision-Making Models and the Measurement of Interrater Agreement. *Psychological Bulletin*, **101**, 140-146. https://doi.org/10.1037//0033-2909.101.1.140

[15] Viera, A.J. and Garrett, J.M. (2005) Understanding Interobserver Agreement: The Kappa Statistic. *Family Medicine*, **37**, 360-363.

[16] Strijbos, J., Martens, R.L., Prins, F.J. and Jochems, W.M.G. (2006) Content Analysis: What Are They Talking about? *Computers & Education*, **46**, 29-48. https://doi.org/10.1016/j.compedu.2005.04.002

Abbreviations

The following abbreviations are used in this manuscript:

IRR	Interrater reliability
ML	Maximum likelihood
AC1	Agreement coefficient 1