MethodsX 10 (2023) 102212

Contents lists available at ScienceDirect

MethodsX

journal homepage: www.elsevier.com/locate/mex

Gwet's AC1 is not a substitute for Cohen's kappa – A comparison of basic properties

Werner Vach[a,b,*], Oke Gerke[c,d]

[a] Basel Academy for Quality and Research in Medicine, Steinenring 6, CH-4031Basel, Switzerland
[b] Department of Environmental Sciences, University of Basel, Spalenring 145, CH-4055 Basel, Switzerland
[c] Department of Nuclear Medicine, Odense University Hospital, J.B. Winsløws Vej 4, DK-5000 Odense C, Denmark
[d] Department of Clinical Research, University of Southern Denmark, J.B. Winsløws Vej 19.3, DK-5000 Odense C, Denmark

REVIEW HIGHLIGHTS

- Both Gwet's AC1 and Cohen's kappa compare the observed agreement rate with a comparator, but they use conceptually completely different comparators.
- Gwet's AC1 and Cohens's kappa behave completely different with respect to the influence of the prevalence of positive ratings and with respect to possible values in the case of independence between the raters.
- These fundamental differences suggest that Gwet's AC1 is not a substitute for Cohen's kappa.

ARTICLE INFO

Method name:
Gwet's AC1

Keywords:
Agreement
Method comparison
Observer
Rater
Reliability
Repeatability
Reproducibility

ABSTRACT

Gwet's AC1 has been proposed as an alternative to Cohen's kappa in evaluating the agreement between two binary ratings. This approach is becoming increasingly popular, and researchers have been criticized for still using Cohen's kappa. However, a rigorous discussion of properties of Gwet's AC1 is still missing. In this paper several basic properties of Gwet's AC1 are investigated and compared with those of Cohen's kappa, in particular the dependence on the prevalence of positive ratings for a given agreement rate and the behaviour in case of no association or maximal disagreement. Both approaches compare the observed agreement rate with a comparative number. Cohen's kappa uses an expected agreement rate as comparator, whereas Gwet's AC1 uses an expected disagreement rate. Consequently, for a fixed agreement rate, Gwet's AC1 increases with increasing difference of the prevalence of positive ratings from 0.5. In contrast, Cohen's kappa decreases. Gwet's AC1 can take positive and negative values in the case of no association between the two raters, whereas Cohen's kappa is 0. Due to these fundamental differences, Gwet's AC1 should not be seen as a substitute for Cohen's kappa. In particular, the verbal classification of kappa values by Landis & Koch should not be applied to Gwet's AC1.

* Corresponding author at: Department of Environmental Sciences, Univetsity of Basel, Spalenring 145, CH-4055 Basel, Switzerland.
E-mail address: werner.vach@basel-academy.ch (W. Vach).

https://doi.org/10.1016/j.mex.2023.102212
Received 15 February 2023; Accepted 7 May 2023
Available online 10 May 2023

Specifications Table

Subject area:	Mathematics and Statistics
More specific subject area:	Measurement of agreement
Name of the reviewed methodology:	Gwet's AC1
Keywords:	Agreement; method comparison; observer; rater; reliability; repeatability; reproducibility.
Resource availability:	Not applicable
Review question:	General: Can Gwet's AC1 be seen as a substitute of Cohens's kappa allowing to apply the classification of Landis and Koch? Specific: (1) Do the methods take the prevalence of positive ratings into account in a similar manner? (2) Do they behave similar in the case of no association? (3) Are the values of comparable magnitude?

Introduction

In many situations in research, there is an interest in assessing the reliability of binary ratings such as positive vs. negative or absent vs. present. This requires performing some type of experiment in which ratings are repeated under different conditions, by different raters, or over time. If two raters perform such a rating in a sample of well-defined sampling units (e.g., single patients or single scans), there are only two possible outcomes at the level of a single unit: The raters can agree or disagree. Consequently, the agreement rate (or disagreement rate) provides the essential information. However, it was noted very early that the agreement rate does not directly reflect the degree of agreement; the interpretation has to take into account the prevalence of positive ratings. Already in 1960, Cohen proposed the kappa coefficient as a simple measure to tackle this issue [1], and it became the most widely used measure for the analysis of binary agreement studies.

However, Cohen's kappa has also been criticized [2–4] and there have been various attempts to define alternative measures. One of these measures is Gwet's AC1, which was first proposed at the start of this millennium [5,6]. Gwet presented a more rigorous development and justification in 2008 [7]. In the following years, the work became more and more popular (cf. Table 1), in particular after publication of a case study in 2013 [8]. In this case study, Gwet's AC1 was compared with Cohen's kappa analysing the agreement between raters on several measures of personality disorders. In addition, the continuous publication of a handbook [5,9,10] probably contributed to the popularity. Gwet's AC1 has been implemented in major statistical packages (SAS [11], R [12], Stata [13]), and mathematicians have started to work on sophisticated inference procedures and generalizations [14–16]. Especially, Gwet's AC1 became popular in the evaluation of assessment tools used in systematic reviews [17–23]. Recently, authors of a paper published in the *Journal of Clinical Epidemiology* on the reliability of a risk of bias tool were criticized in a letter to the editor for using Cohen's kappa instead of Gwet's AC1 [24]. However, some authors required a more systematic investigations of properties of Gwet's AC1 in order to come to recommendations about its use [25,26].

Already [8] applied the classification suggested by Landis and Koch to Gwet's AC1, and many other authors followed this example. This classification was originally suggested to assist in the interpretation of Cohen's kappa by defining ranges which can be translated into the terms "Slight", "Fair", "Moderate", "Substantial" or "Almost Perfect". The application of this classification to Gwet's AC1 requires that Gwet's AC1 can be seen as a substitute for Cohen's kappa. It is the aim of this paper to investigate some mathematical properties of Gwet's AC1 and Cohen's kappa to check whether this is indeed the case.

Methods

Notation

If two observers A and B perform a binary rating on a sample of units (i.e., resulting in the values 0 and 1), the results can be presented as a simple cross tabulation as shown in Table 2. Here p_{ij} denotes the relative frequency to observe the rating i by observer A and the rating j by observer B, and π_A and π_B denote the probability of a positive rating by observer A and observer B, respectively. With $a = p_{00} + p_{11}$ we denote the agreement rate, i.e., the relative frequency of agreement.

The need to take the prevalence of positive ratings into account in interpreting an agreement rate

Interpreting the degree of agreement between two raters based on the agreement rate a is a nontrivial task. The crucial point is that even if the raters tend to come to opposite decisions, there will be a certain amount of agreeing decisions depending on the

Table 1
Number of references (according to Web of Science, 02/02/23) to two publications [7,8].

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Gwet [7]	3	16	14	25	42	34	41	53	64	81	115	160
Wongparkan et al. [8]				2	11	15	16	24	44	44	77	107

Table 2
The results of a simple agreement study represented as a cross tabulation.

		B		
		0	1	
A	0	p_{00}	p_{01}	$1-\pi_A$
	1	p_{10}	p_{11}	π_A
		$1-\pi_B$	π_B	

Table 3
Minimal agreement and maximal disagreement between two raters with a prevalence of 90%.

		B		
		0	1	
A	0	0%	10%	10%
	1	10%	80%	90%
		10%	90%	

prevalence values π_A and π_B. For example, if both observers have a prevalence of 15%, they can maximally disagree on 30% of the units, and hence a will be at least 70%. Only if both observers have exactly a prevalence of 50%, it is possible that they can disagree on all units.

Besides such worst-case scenarios of maximal disagreement, the issue can be also illustrated by considering the case that the raters perform the rating completely independent from each other, i.e., there is no statistical association between the decisions made by the raters. If both observers have a prevalence of 50% and rate independently, an agreement rate of 50% is to be expected ($0.5^2+(1-0.5)^2 = 0.5$). If both observers have a prevalence of 70% and rate independently, the probability of a joint positive rating is 49% and of a joint negative rating 9%, i.e., an agreement rate of 58% is to be expected ($0.7^2+(1-0.7)^2 = 0.58$). If both observers have a prevalence of 90%, the probabilities are 81% and 1%, i.e., an agreement rate of 82% is to be expected ($0.9^2+(1-0.9)^2 = 0.82$).

Consequently, it is not sufficient to inspect the agreement rate a in order to judge the degree of agreement. An agreement rate of 80% may be an indication of a relevant degree of agreement if the prevalence is close to 50%, but it is definitely a poor degree of agreement if the prevalence is 90% – it is the least possible agreement rate in this situation as the raters can maximally disagree in 20% of the subjects. This situation is depicted in Table 3.

The definition of Cohen's kappa and Gwet's AC1

Cohen's kappa uses the case of statistical independence between the two raters as benchmarking. Consequently, in a first step, the expected agreement rate is defined as

$$e = \pi_A\pi_B + (1-\pi_A)(1-\pi_B).$$

This expected agreement rate is the probability of agreement between the two raters if they rate independently of each other, but keep the observed prevalence values π_A and π_B .

In a second step, the expected agreement rate is related to the observed agreement rate a by taking the difference, and this is related to the case of full agreement, i.e., $a = 1$:

$$\kappa = \frac{a-e}{1-e}$$

Gwet's AC1 has the same structure

$$g = \frac{a-\gamma}{1-\gamma}$$

but e is replaced by the quantity

$$\gamma = 2\pi(1-\pi)$$

with

$$\pi = \frac{1}{2}(\pi_A + \pi_B).$$

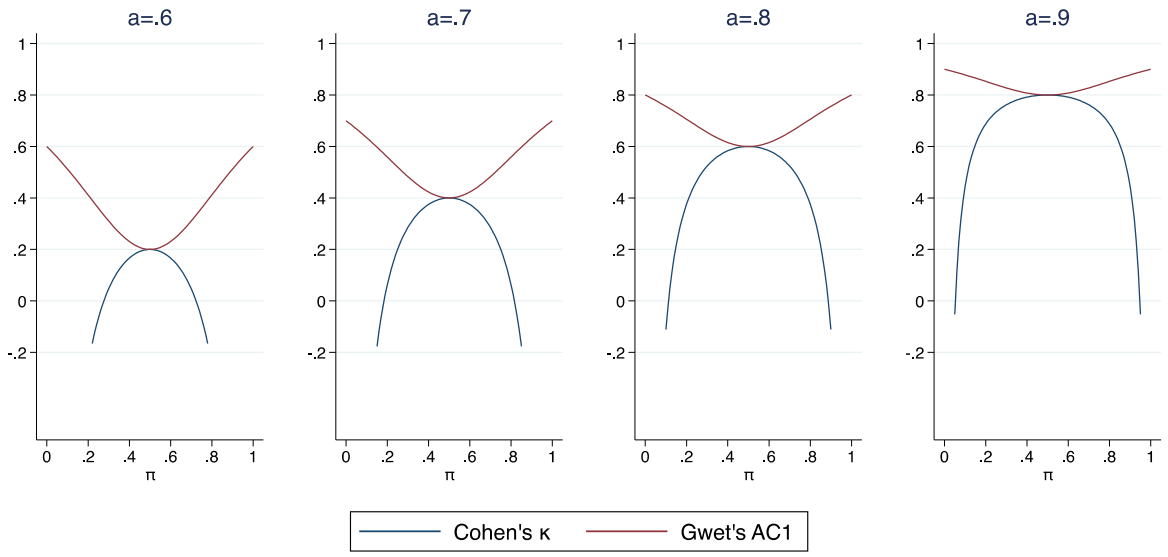

Fig. 1. Cohen's kappa and Gwet's AC1 as a function of the prevalence π for the special case $\pi = \pi_A = \pi_B$.

In the derivation of the formula for γ, Gwet assumes that each rater performs a random rating by flipping a fair coin for some sample units, whereas the decision is certain for other sample units [7]. Defining the two events G={The two raters A and B agree} and R = {A, B or both perform a random rating}, the probability of chance agreement is then defined as $p_e = P(G \text{ and } R)$, i.e. the probability that both raters agree and at least one has performed a random rating. Flipping a coin implies that $P(G|R) = 0.5$, and hence $p_e = P(G|R)\ P(R) = 0.5\ P(R)$. Gwet pointed out that $P(R)$ can only be approximated and that a balanced distribution of positive and negative ratings may be a hint to random rating. Finally, he suggested to approximate $P(R)$ by the ratio of the variance of the prevalence to the maximum possible variance, i.e., by $\pi\ (1-\pi)\ /\ (0.5\ (1-0.5)) = 4\pi\ (1-\pi)$, which then implies $p_e = 2\ \pi\ (1-\pi)$. This coincides with the value of γ given above.

It should be noted that this derivation is based on a different notation of "chance agreement" than Cohen's kappa. In both cases "chance agreement" refers to (stochastic) independence between the two raters. However, for Cohen's kappa, this is assumed in all sample units with the raters keeping their prevalence, whereas in Gwet's AC1, this is assumed only in a subset of sample units in which the raters perform a random rating with a prevalence of 0.5.

Results

Gwet's AC1 compares the observed agreement with an expected disagreement

The quantity γ can be also interpreted as a probability under the assumption of independence between the two raters given that both raters have the same prevalence π. Under this assumption γ is just the probability of disagreement between the two raters. Disagreement happens when the raters score differently: the first rater gives the rating 1 and the second the rating 0, or the first rater gives the rating 0 and the second the rating 1. The probability of the first event is $\pi\ (1-\pi)$, the probability of the second is $(1-\pi)\pi$. Consequently, $\gamma = 2\ \pi\ (1-\pi)$ is the probability of disagreement under this assumption.

This indicates a fundamental difference to the labelling used by Gwet [7] and subsequent authors, who call γ a "chance-agreement probability".

Gwet's AC1 reverses the relation to the prevalence

Both Cohen's kappa and Gwet's AC1 are functions of the agreement rate a and the prevalence values π_A and π_B. Hence, we can easily study how the values of these two statistics depend on the prevalence values, when the agreement rate a is fixed.

Fig. 1 depicts this dependence for the special case of $\pi = \pi_A = \pi_B$. For Cohen's kappa we observe that it decreases with increasing distance from $\pi = 0.5$. This is exactly what is desired following our considerations above. Gwet's AC1 behaves exactly in the opposite manner: It increases with increasing distance from $\pi = 0.5$. The same pattern can be also observed if the prevalence differs between the two observers. This is illustrated for four special cases in Fig. 2.

The opposite behaviour of Gwet's AC1 is a simple consequence of using the expected disagreement instead of the expected agreement in the computation. The expected agreement tends to increase with increasing distance of the prevalence from 0.5. The expected disagreement tends to decrease.

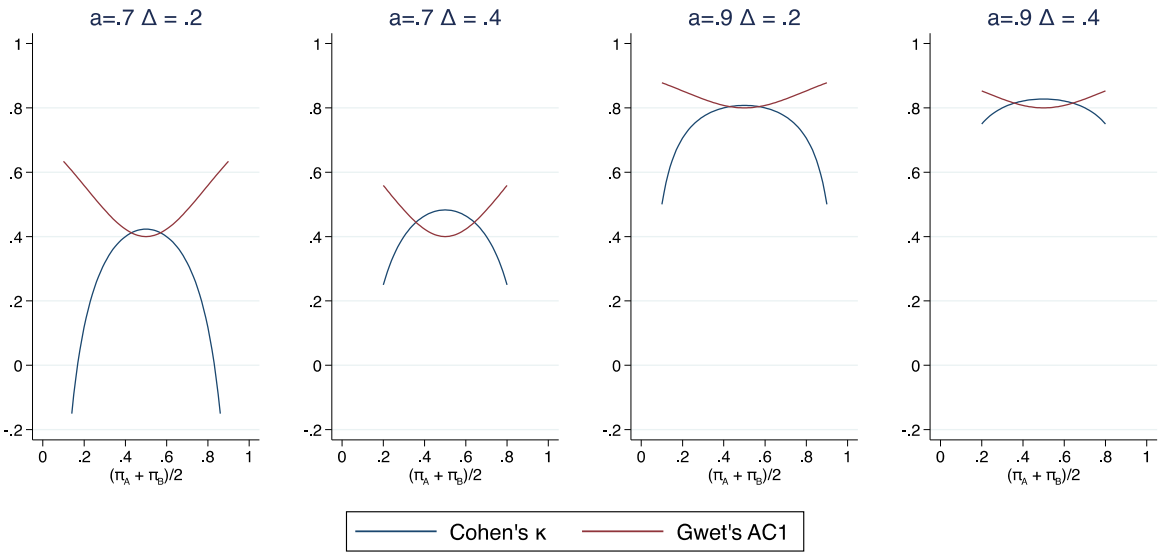

Fig. 2. Cohen's kappa and Gwet's AC1 as a function of the average prevalence $\frac{1}{2}(\pi_A + \pi_B)$ for four choices of the agreement rate a and the difference $\Delta = \pi_B - \pi_A$.

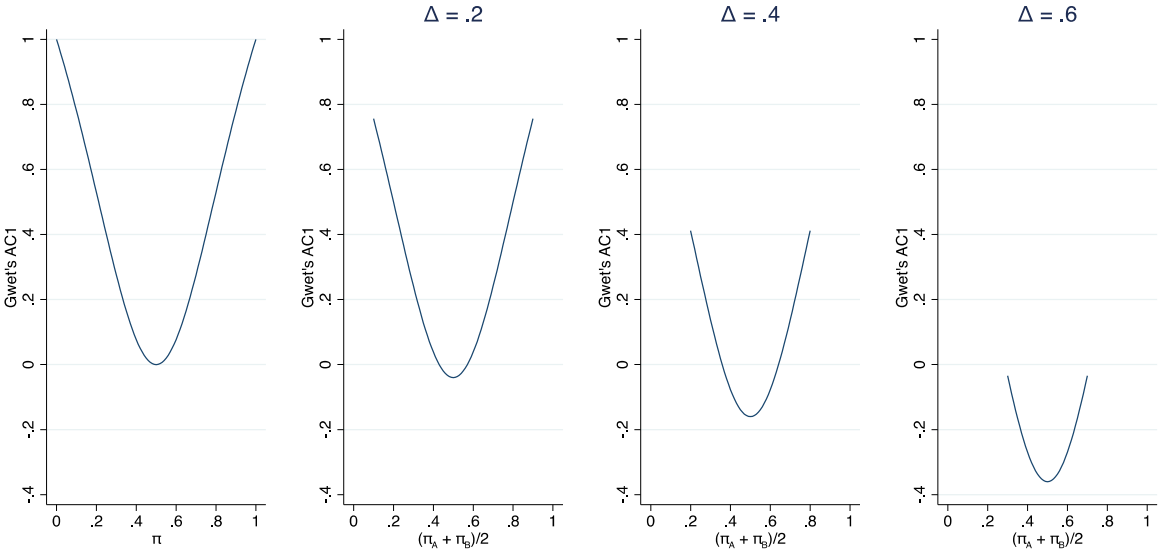

Fig. 3. Gwet's AC1 as a function of the prevalence in the case of independent rating by the two raters, i.e., $a = e$. The leftmost panel considers the case $\pi = \pi_A = \pi_B$. The other three panels consider a specific choice of $\Delta = \pi_B - \pi_A$.

Gwet's AC1 can be high in the case of no association

Cohen's kappa is by definition 0, if there is no statistical association between the decisions of the raters, i.e., if the decision of one rater does not influence the probability of a positive decision by the other rater. This does not hold for Gwet's AC1. Actually, Gwet's AC1 can be close to 1 in this situation.

If there is no statistical association between the decisions of the raters, we have $a = e$, and we can easily study the behaviour of Gwet's AC1 in dependence on the prevalence. The leftmost panel in Fig. 3 considers the case of equal prevalence values. It can be observed that Gwet's AC1 is always nonnegative, but it can become close to 1 if the prevalence is close to 0 or 1. This is a consequence of the fact that the expected agreement rate tends to 1 and the expected disagreement rate to 0. If the prevalence values are unequal, Gwet's AC1 may become also negative, but it never shows a tendency to get close to 0, i.e., to the value of Cohen's kappa.

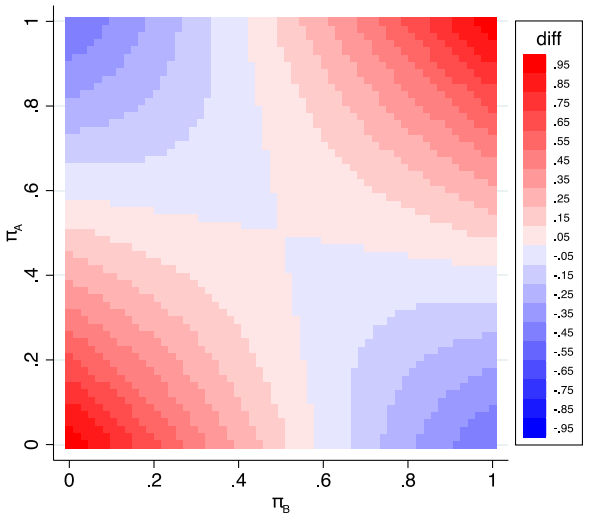

Fig. 4. The value of $e - \gamma$ as a function of π_A and π_B.

Gwet's AC1 can be high in the case of the minimal possible agreement

Even in the case of a negative association between the two ratings, Gwet´s AC1 can be high. This can be illustrated using the example shown in Table 3, depicting the situation of minimal positive agreement between two raters both characterized by a prevalence of 90%. In this case we have $\gamma = 2 \times 0.9 \times 0.1 = 0.18$ and $g = (0.8 - 0.18)/(1 - 0.18) = 0.76$. Cohen's kappa is, by the way, −0.11 here.

Gwet's AC1 is often larger than Cohen's kappa

Figs. 1 and 2 have already indicated that Gwet's AC1 is nearly always larger than Cohen's kappa. In the case of equal prevalence values, this is exactly true, as here $e \geq 0.5$ and $\gamma \leq 0.5$ holds, and hence $\gamma \leq e$. The latter implies $g \geq \kappa$ as $(a - x)/(1 - x)$ is a monotonously decreasing function in x.

The general case is considered in Fig. 4 depicting the value of $e - \gamma$ as a function of the two prevalence values. Gwet's AC1 is greater than Cohen's kappa if $e - \gamma$ is positive (indicated in red), whereas Gwet's AC1 is smaller than Cohen's kappa if $e - \gamma$ is negative (indicated in blue). It can be observed that Gwet's AC1 tends to be larger than Cohen's kappa if the two prevalence values are similar, and it only tends to be smaller than Cohen's kappa if the two raters show very different prevalence values. In the first case, the difference $e - \gamma$ becomes in particular large if the prevalence values are close to 0 or 1.

Discussion

Our investigation reveals fundamental differences between Gwet's AC1 and Cohen's kappa. Gwet's AC1 does not compare the observed agreement rate with an expected agreement rate, but with an expected disagreement rate. Consequently, it reverses the relation to the prevalence. For a fixed agreement rate Cohen's kappa decreases with increasing distance of the prevalence from 0.5, which is in accordance with the aim to take into account that high agreement rates get more likely; Cohen's kappa is a measure of agreement beyond what could be expected by chance. In contrast, Gwet's AC1 increases with increasing distance of the prevalence from 0.5. Hence Gwet's AC1 cannot be seen as a substitute for Cohen's kappa.

Moreover, Gwet's AC1 can produce high values even if there is no (statistical) association between the ratings of the two raters. Cohen's kappa results always in a value of 0 in this case. A similar property has been also mentioned with respect to Gwet's AC2 [14], a generalization to ordinal scales. Actually, Gwet's AC1 can also produce high values in the case of a minimally possible agreement rate.

These fundamental differences should forbid to apply the classification of Landis and Koch [27] to Gwet's AC1. This classification was proposed with respect to translating the value of Cohen's kappa into a verbal ranking. It was not intended to translate any measure taking values between 0 and 1. The application of this classification is in particular problematic as Gwet's AC1 tends to produce rather systematically larger (and often much larger) values than Cohen's kappa.

Consequently, it should be stopped to criticize researchers for using Cohen's kappa and to urge them to switch to Gwet's AC and applying the classification of Landis and Koch, as done by [24]. We regard this as an attempt of whitewashing the deficits of binary assessment rules. Similar it is inappropriate and misleading to combine estimates of Gwet's AC1 and Cohen's kappa into one meta-analysis, as done by [28].

Gwet [7] claimed that Gwet's AC1 is less biased than Cohen's kappa, and other authors have used the decreased bias as an argument in favour of using Gwet's AC1 [24]. It should be noted that [7] did not investigate the statistical bias of an estimation procedure in the traditional sense. This would require studying characteristics of the sampling distribution of Gwet's AC1 and Cohen's kappa with respect to the distance to the expected "true" value. This true value differs between Gwet's AC1 and Cohen's kappa. Instead, [7] defines artificially a common true value motivated within the framework used to derive Gwet's AC1. Hence, it is in no way surprising that Gwet´s AC1 is on average closer to this value than Cohen´s kappa.

Our investigation focused on the comparison of Gwet's AC1 with Cohen's kappa to address the question whether Gwet's AC1 can be regarded as a substitute for Cohen's kappa. We provided several arguments why this is not the case. This still leaves open the question whether there is separate role for Gwet's AC1 in the evaluation of agreement. In combination with a classification system tailored to Gwet's AC1 it may supplement the current widespread use of Cohen's kappa or – if conceptual superiority can be demonstrated – it may even replace it. This should be a topic of further discussions. However, in our opinion the comparison of an observed agreement rate with an expected disagreement rate is counterintuitive.

In motivating the use of Gwet's AC1, several authors have referred to problems in interpreting Cohen's kappa. We would like to note that previous research has pointed out that these problems have been exaggerated [29]. They partially just reflect that any reasonable measure has to depend on the agreement rate and the prevalence.

Ethics statements

The Author had followed MethodsX ethical guidelines, this work does not involve human subjects, animal experiments or data collected from social media.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Werner Vach: Conceptualization, Methodology, Investigation, Writing – original draft. **Oke Gerke:** Conceptualization, Writing – review & editing.

Data availability

No data was used for the research described in the article.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1960) 37–46, doi:10.1177/001316446002000104.
[2] A.R. Feinstein, D.V. Cicchetti, High agreement but low kappa: I. the problems of two paradoxes, J. Clin. Epidemiol. 43 (1990) 543–549, doi:10.1016/0895-4356(90)90158-L.
[3] R. Cook, P. Armitage, T. Colton, Kappa and its dependence on marginal rates, Encyclopedia of Biostatistics, Wiley, Chichester, UK, 1998.
[4] D.G. Altman, Practical Statistics For Medical Research, Chapman & Hall/CRC, Boca Raton, FL, 1999.
[5] K.L. Gwet, Handbook of Inter-Rater Reliability, StatAxis Publishing Company, Gaithersburgh, MD, 2001.
[6] K.L. Gwet, Kappa statistic is not satisfactory for assessing the extent of agreement between raters, Stat. Methods Interrater Reliab. Assess. (1) (2002) https://agreestat.com/papers/kappa_statistic_is_not_satisfactory.pdf. accessed May 20, 2022.
[7] K.L. Gwet, Computing inter-rater reliability and its variance in the presence of high agreement, Br. J. Math. Stat. Psychol. 61 (2008) 29–48, doi:10.1348/000711006X126600.
[8] N. Wongpakaran, T. Wongpakaran, D. Wedding, K.L. Gwet, A comparison of Cohen's kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples, BMC Med. Res. Methodol. 13 (2013) 61, doi:10.1186/1471-2288-13-61.
[9] K.L. Gwet, Handbook of Inter-Rater Reliability: the Definitive Guide to Measuring the Extent of Agreement Among Raters, 3rd ed., Advanced Analytics, Gaithersburg, MD, 2012.
[10] K.L. Gwet, Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters, 4th ed., Advances Analytics, LLC, Gaithersburg, Md, 2014.
[11] E. Blood, K.F. Spratt, Disagreement on agreement: two alternative agreement coefficients, SAS global forum 2007: statistics and data analysis. paper 186-2007. (2007). https://support.sas.com/resources/papers/proceedings/proceedings/forum2007/186-2007.pdf (accessed October 11, 2022).
[12] P. Brasil, Gwet's AC1 interrater reliability, R-Sig-Epi. (2012). https://stat.ethz.ch/pipermail/r-sig-epi/2012-May/000273.html (accessed October 4, 2022).
[13] D. Klein, Implementing a general framework for assessing interrater agreement in Stata, Stata J. 18 (2018) 871–901, doi:10.1177/1536867X1801800408.
[14] D. Tran, A. Dolgun, H. Demirhan, Weighted inter-rater agreement measures for ordinal outcomes, Commun. Stat. Simul. Comput. 49 (2020) 989–1003, doi:10.1080/03610918.2018.1490428.
[15] C. Honda, T. Ohyama, Homogeneity score test of AC1 statistics and estimation of common AC1 in multiple or stratified inter-rater agreement studies, BMC Med. Res. Methodol. 20 (2020) 20, doi:10.1186/s12874-019-0887-5.
[16] T. Ohyama, Statistical inference of Gwet's AC_1 coefficient for multiple raters and binary outcomes, Commun. Stat. Theory Methods 50 (2021) 3564–3572, doi:10.1080/03610926.2019.1708397.

[17] A.V. Margulis, M. Pladevall, N. Riera-Guardia, C. Varas-Lorenzo, L. Hazell, N. Berkman, M. Viswanathan, S. Perez-Gutthann, Quality assessment of observational studies in a drug-safety systematic review, comparison of two tools: the Newcastle-Ottawa Scale and the RTI item bank, CLEP (2014) 359, doi:10.2147/CLEP.S66677.
[18] K.I. Bougioukas, A. Liakos, A. Tsapas, E. Ntzani, A.-.B. Haidich, Preferred reporting items for overviews of systematic reviews including harms checklist: a pilot tool to be used for balanced reporting of benefits and harms, J. Clin. Epidemiol. 93 (2018) 9–24, doi:10.1016/j.jclinepi.2017.10.002.
[19] S. Dosenovic, A. Jelicic Kadic, K. Vucic, N. Markovina, D. Pieper, L. Puljak, Comparison of methodological quality rating of systematic reviews on neuropathic pain using AMSTAR and R-AMSTAR, BMC Med. Res. Methodol. 18 (2018) 37, doi:10.1186/s12874-018-0493-y.
[20] K.I. Bougioukas, E. Bouras, F. Apostolidou-Kiouti, S. Kokkali, M. Arvanitidou, A.B. Haidich, Reporting guidelines on how to write a complete and transparent abstract for overviews of systematic reviews of health care interventions, J. Clin. Epidemiol. 106 (2019) 70–79, doi:10.1016/j.jclinepi.2018.10.005.
[21] N. Black, A.J. Williams, N. Javornik, C. Scott, M. Johnston, M.C. Eisma, S. Michie, J. Hartmann-Boyce, R. West, W. Viechtbauer, M. de Bruin, Enhancing behavior change technique coding methods: identifying behavioral targets and delivery styles in smoking cessation trials, Ann. Behav. Med. 53 (2019) 583–591, doi:10.1093/abm/kay068.
[22] R.C. Lorenz, K. Matthias, D. Pieper, U. Wegewitz, J. Morche, M. Nocon, O. Rissling, J. Schirm, A. Jacobs, A psychometric study found AMSTAR 2 to be a valid and moderately reliable appraisal tool, J. Clin. Epidemiol. 114 (2019) 133–140, doi:10.1016/j.jclinepi.2019.05.028.
[23] Y. Zhang, L. Huang, D. Wang, P. Ren, Q. Hong, D. Kang, The ROBINS-I and the NOS had similar reliability but differed in applicability: a random sampling observational studies of systematic reviews/meta-analysis, J. Evid. Based Med. 14 (2021) 112–122, doi:10.1111/jebm.12427.
[24] M. Loef, H. Walach, S. Schmidt, Interrater reliability of ROB2 – an alternative measure and way of categorization, J. Clin. Epidemiol. 142 (2022) 326–327, doi:10.1016/j.jclinepi.2021.09.003.
[25] S. Kuppens, G. Holden, K. Barker, G. Rosenberg, A kappa-related decision: K, Y, G, or AC1, Soc. Work Res. 35 (2011) 185–189, doi:10.1093/swr/35.3.185.
[26] S. Minozzi, M. Cinquini, S. Gianola, M. Gonzalez-Lorenzo, R. Banzi, Kappa and AC1/2 statistics: beyond the paradox, J. Clin. Epidemiol. 142 (2022) 328–329, doi:10.1016/j.jclinepi.2021.09.004.
[27] J.R. Landis, G.G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1977) 159–174.
[28] J. Ressman, W.J.A. Grooten, E. Rasmussen-Barr, Visual assessment of movement quality: a study on intra- and interrater reliability of a multi-segmental single leg squat test, BMC Sports Sci. Med. Rehabil. 13 (2021) 66, doi:10.1186/s13102-021-00289-x.
[29] W. Vach, The dependence of Cohen's kappa on the prevalence does not matter, J. Clin. Epidemiol. 58 (2005) 655–661, doi:10.1016/j.jclinepi.2004.02.021.