Psicológica (2011), 32, 385-402.

A simulation study of rater agreement measures with 2x2 contingency tables

Manuel Ato[*], Juan José López and Ana Benavente

Universidad de Murcia (Spain)

A comparison between six rater agreement measures obtained using three different approaches was achieved by means of a simulation study. Rater coefficients suggested by Bennet's σ (1954), Scott's π (1955), Cohen's κ (1960) and Gwet's γ (2008) were selected to represent the classical, descriptive approach, α agreement parameter from Aickin (1990) to represent loglinear and mixture model approaches and Δ measure from Martín and Femia (2004) to represent multiple-choice test. Main results confirm that π and κ descriptive measures present high levels of mean bias in presence of extreme values of prevalence and rater bias but small to null levels with moderate values. The best behavior was observed with Bennet and Martín and Femia agreement measures for all levels of prevalence.

There are a lot of behavioral research applications where is needed to quantify the homogeneity of agreement between responses given by two (or more) observers or between two (or more) measurement devices. With responses in a numerical scale, the classical intraclass correlation coefficient (Shrout & Fleiss, 1972; McGraw & Gow, 1996) and more recent concordance coefficient (Lin *et al.*, 2002) are the most frequently used alternatives, and it has been demonstrated that Lin's coefficient can be estimated by means of an special intraclass coefficient (Carrasco & Jover, 2003). With measures in a categorical scale there are a greater variety of options (Shroukri, 2004, von Eye and Mun, 2005). In both cases the methodological scenario is matched pairs where two o more raters classify N targets on a categorical scale with M categories producing a $M \times M$ contingency table also known as agreement table (Agresti, 2002).

[*] Funds for this research were provided by a grant from the Spanish Ministry of Science and Culture, reference number SEJ2006/09025. Correspondence address: Manuel Ato García. Facultad de Psicología. Dpto. Psicología Básica y Metodología. Universidad de Murcia. Apartado 4021. 30080 Murcia (Spain). E-mail: matogar@um.es

Three general approaches to rater agreement measurement for categorical data leads to three different forms of agreement coefficients (Ato, Benavente y Lopez, 2006). A first, descriptive approach started from a suggestion of Scott (1955) to correct from observed agreement the proportion of cases in which agreement occurred only at random. Several descriptive measures have been defined within this approach. In this paper the interest will be focused on the π-coefficient (Scott, 1955), on a proposal firstly suggested by Bennet *et al.* (1954), reused afterwards with different names and relabeled as σ-coefficient (see Zwick, 1988; Hsu & Smith, 2003), the classical κ -coefficient (Cohen, 1960) and a more recent proposal of Gwett's γ -coefficient (2008). Differences between these measures are relative to the particular definition of chance correction.

Loglinear and mixture modelling is a second approach which is used when the focus is the detailed examination of agreement and disagreement patterns (Tanner & Young, 1986, Agresti, 1992). Mixture modelling is a generalization of loglinear approach with an unobserved (categorical) latent variable. The set of targets to be classified is assumed to be drawn from a population that is a mixture of two subpopulations (latent classes), one related to objects easy to classify by both raters (systematic agreement) and the other to objects hard to classify (random agreement and disagreement). Within this approach it is also possible to reproduce all the descriptive rater measures (Guggenmoos-Holtzman & Vonk, 1998; Schuster & von Eye, 2001, Schuster & Smith, 2002; Ato, Benavente y López, 2006) and also to define new rater agreement measures as α -coefficient (Aickin, 1990).

A third alternative approach is inspired in the tradition of multiple-choice test and developed in order to overcome the limitations shown for many descriptive measures. Within this tradition, Martín & Femia (2004, 2008) defined a new measure of agreement, the Δ-coefficient, as 'the proportion of agreements that are not due to chance'.

In this paper we use a simulation study to compare the behavior of these rater agreement measures for 2 (rater) $\times$ 2 (categories) agreement tables. The rest of this paper is organized as follows. The second section comments the main notation and formulas used for descriptive, loglinear and mixture rater agreement measures. The third section describes with detail the simulation process of this study. The final section shows the main results obtained and some implications for research practice.

Rater agreement measures

Notation

Let A and B denote 2 raters classifying a set of targets into M categories, with responses $i = 1,..., M$ for observer A and $j = 1,\ldots, M$ for observer B. In this work we confine our interest to the 2 (raters) × 2 (categories) agreement table case. Let $\{\pi_{ij}\}$ the joint probabilities of responses in *i* row and *j* column given for both raters, and $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$ the row and column marginal distributions resulting of summing the joint probabilities where $\sum_i \pi_{i+} = \sum_j \pi_{+j} = \sum_i \sum_j \pi_{ij} = 1.$

Given a sample of N objects to be classified, Table 1 summarizes the notation used in this paper characterizing four cells inside the agreement table: $p_{11} = n_{11} / N$ refers to the proportion of responses of both raters for first (or positive) category, $p_{22} = n_{22} / N$ to the proportion of agreement responses for second (or negative) category, and $p_{12} = n_{12} / N$ and $p_{21} = n_{21} / N$ are proportions of disagreement responses between raters. Similarly, $p_{1+} = n_{1+} / N$ and $p_{2+} = n_{2+} / N$ are marginal proportions for both categories corresponding to rater A and $p_{+1} = n_{+1} / N$ and $p_{+2} = n_{+2} / N$ are marginal proportions for rater B.

Table 1. Joint and marginal proportions for the 2 x 2 agreement table.

	Rater B		
Rater A	*1*	*2*	***Marginal A***
1	p_{11}	p_{12}	p_{1+}
2	p_{21}	p_{22}	p_{2+}
Marginal B	p_{+1}	p_{+2}	p_{++}

Descriptive measures

A simple formula to measure agreement is the observed proportion of agreement which is the sum of diagonal cells of Table 1. For the 2×2 agreement table, $p_o = p_{11} + p_{22}$. This formula is a common component in all measures of agreement considered in this paper but it fails to point us how much of observed agreement is due to chance. Thus the concept of "random corrected agreement" (RCA) is the basic notion that pervades the practical utilization of descriptive measures on agreement.

A general measure of RCA is

$$RCA = \frac{p_o - p_e}{1 - p_e} = \frac{p_o}{1 - p_e} - \frac{p_e}{1 - p_e} \qquad (1)$$

where p_e is chance agreement proportion and $p_o - p_e$ corrects the excess of observed agreement proportion that is assumed to be computed with p_o. The difference must be weighted with $1 - p_e$, the possible maximum of non chance agreement proportion. The values of RCA are constrained to lie inside the interval $[-p_e/(1-p_e); \ +p_o/(1-p_e)]$, where $RCA = -p_e/(1-p_e) > -1$ is the lower limit associated with perfect disagreement, $RCA = 0$ means that observed agreement is equal to chance agreement probability and $RCA = p_o/(1-p_e) \leq 1$ is the higher limit associated with perfect agreement that is only possible when chance agreement proportion is zero.

The four descriptive measures used in this study are based on the general formula (1) and assume independence between raters in the classification process. Differences between them are due to the specific definition of chance-corrected agreement p_e.

A first solution was originally proposed by Bennet *et al*. (1954) by using as chance correction formula a fixed value, the inverse of the number of categories. This solution has been relabeled as σ-coefficient, and chance-corrected proportion as $p_e^{\sigma}=1/M$. More recently, σ-coefficient has become a consolidated agreement measure and reconsidered in works such as Holley & Guilford (1964), Janson & Vegelius (1979), Maxwell (1977) and Brennan & Prediger (1981). See Zwick, 1988 and Hsu & Field, 2003, for a more detailed explanation of how researchers have been using different names for the same procedure. This form of chance correction assumes that observers uniformly classify targets in categories, and then is based on an uniform distribution of targets. For $M = 2$ categories, $p_e^{\sigma} = .5$, and using (1) σ-coefficient can be estimated as

$$\hat{\sigma} = \frac{p_o}{0.5} - \frac{0.5}{0.5} = 2(p_{11} + p_{22}) - 1 \qquad (2)$$

A second solution was proposed by Scott (1955) using the squared mean of row and column marginal proportions for each category as chance-corrected agreement. Defining the mean of marginal probabilities for each category simply as $p_1 = (p_{1+} + p_{+1})/2$ and $p_2 = (p_{2+} + p_{+2})/2$, then

$$p_e^{\pi} = p_1^2 + p_2^2 \qquad (3)$$

This formula assumes that observers classify targets using a common homogeneous distribution. The resulting π-coefficient can be estimated using (1) as

$$\hat{\pi} = \frac{p_o - p_e^{\pi}}{1 - p_e^{\pi}} \qquad (4)$$

A third solution was proposed by Cohen (1960) using as chance-corrected formula the sum of products of row and column marginal probabilities for each category

$$p_e^{\kappa} = p_{1+}p_{+1} + p_{2+}p_{+2} \qquad (5)$$

and so assumes that each observer classify targets using his/her own distribution. The resulting κ-coefficient is also estimated using (1) as

$$\hat{\kappa} = \frac{p_o - p_e^{\kappa}}{1 - p_e^{\kappa}} \qquad (6)$$

Research literature (see Feinstein & Cichetti, 1990; Byrt, Bishop & Carlin, 1993; Agresti, Ghosh & Bini, 1995; Lantz & Nebenzahl, 1996 and Hoehler, 2000, among others) reports that κ- and π-coefficients have two main limitations as a direct consequence of using these chance-corrected agreement measures: prevalence problems refer to the particular distribution of data across the categories and arise in presence of extreme values for one of the two categories and rater bias problems appear particularly with extreme marginal distributions of two raters.

More recently, Gwett (2001, 2008) proposed a RCA formula that seems to be more stable under certain conditions and uses the mean of marginal probabilities for each category simply as chance-corrected agreement

$$p_e^\gamma = 2p_1p_2 \tag{7}$$

and the resulting g -coefficient is estimated using (1) as

$$\hat{\gamma} = \frac{p_o - p_e^\gamma}{1 - p_e^\gamma} \tag{8}$$

Descriptive measures, and κ -coefficient in particular, are very popular between researchers of behavioral and social sciences. The interpretation of these measures has been generally focused as a classical null hypothesis of RCA equal zero (see Fleiss, Levin and Paik, 2003). It has been shown that they can also be understood as a restricted loglinear or mixture model within the QI family (see Ato, Benavente y López, 2006).

Loglinear and mixture model measures

Given a set of N targets to be classified by 2 raters in 2 response categories, loglinear models distinguish between components such as random expected agreement and non random expected agreement and so they can evaluate the fit of the model to the data (von Eye & Mun, 2005). The basic, starting model representing random expected agreement is independence model, $\log(m_{ij}) = \lambda + \lambda_i^A + \lambda_j^B$, where m_{ij} are expected values, λ_i^A and λ_j^B are individual rater effects and the model releases $M^2 - 2M + 1$ residual degrees of freedom with M being the number of categories. Because rater agreement is concentrated on diagonal cells, a more appropriate model is the quasi-independence (QI) model,

$$\log(m_{ij}) = \lambda + \lambda_i^A + \lambda_j^B + \delta_i \tag{9}$$

where δ_i is a diagonal parameter that represents systematic agreement for i-th category. This model releases $M^2 - 3M + 1$ residual degrees of freedom and so cannot be directly estimated with 2×2 agreement tables.

QI model is directly related with the concept of RCA. A general formula to obtain a model-based agreement measure, which allows correcting from observed agreement a component due to chance, can be defined (see Guggenmoos-Holtzmann, 1993; Guggenmoos-Holtzmann and Vonk, 1998; Ato, Benavente and López, 2006) by:

$$RCA(QI) = \sum_i \left[p_{ii} - \frac{p_{ii}}{\exp(\hat{\delta}_i)} \right] \qquad (10)$$

where p_{ii} is the observed agreement proportion in *i*-th diagonal cell and $\exp(\hat{\delta}_i)$ is the estimation of exponential transformation of δ_i diagonal parameter in the QI model (9). As it can be seen, the framework of equation (10) is very similar to *RCA* equation (1) and the range of possible values spreads from $\sum_i p_{ii} \leq 1$, when there is no disagreement, to $-\sum_i p_{ii} / \exp(\delta_i) > -1$, when agreement is null.

QI model is the most general of a family of models whose members can be defined applying some restrictions on the basic model. A basic kind of restriction is the constant quasi-independence (QIC) model (Ato, Benavente and López, 2006),

$$\log(m_{ij}) = \lambda + \lambda_i^A + \lambda_j^B + \delta \qquad (11)$$

where δ represents systematic agreement and is assumed constant for all categories. This model, which was firstly proposed by Tanner & Young (1996), releases $M^2 - 2M$ residual degrees of freedom being saturated for 2×2 tables and can be defined using a similar formulation to (10) by

$$RCA(QIC) = \sum_i \left[p_{ii} - \frac{p_{ii}}{\exp(\hat{\delta})} \right] \qquad (12)$$

A generalization of loglinear models including a latent variable leads to latent-class mixture models where targets to be classified are assumed to come from a population representing a mixture of two finite subpopulations. Each subpopulation identifies a cluster of homogeneous items, one related with easy to classify items where both raters give the same response (systematic agreement), and the other related with items of difficult classification where both raters give a random response (random agreement and disagreement). This generalization change the status of $M \times M$ agreement table into a $2 \times M \times M$ table, but it hardly affect to the loglinear model required to estimate its parameters. Assuming a QI loglinear model, the representation of a QI mixture model is very similar to (9),

$$\log(m_{ijk}) = \lambda + \lambda_i^A + \lambda_j^B + \xi_i \qquad (13)$$

where mixture ξ_i parameters are related with loglinear δ_i parameters (Guggenmoos-Holtzman, 1993, Ato, Benavente & López, 2006) by

$$\exp(\xi_i) = \exp(\delta_i) - 1 \qquad (14)$$

and the new subscript k indicates the re-dimensionality of agreement table due to the latent variable.

A popular agreement measure derived from a mixture model is α (Aickin, 1990) and its equivalent loglinear model is QIC (11). The representation of QIC mixture model is

$$\log(m_{ijk}) = \lambda + \lambda_i^A + \lambda_j^B + \xi \qquad (15)$$

where again the correspondence (14) allows to estimate a QIC mixture model by means of its equivalent loglinear model using (12). For the case of a 2 × 2 agreement table, Guggenmoos-Holtzman (1993: 2194) also showed that constant parameter exp(δ)could also be estimated using the odds ratio

$$\exp(\hat{\delta}) = \sqrt{(p_{11}p_{22})/(p_{21}p_{12})} \qquad (16)$$

and the agreement measure can be obtained using (12).

Multiple-choice test measures

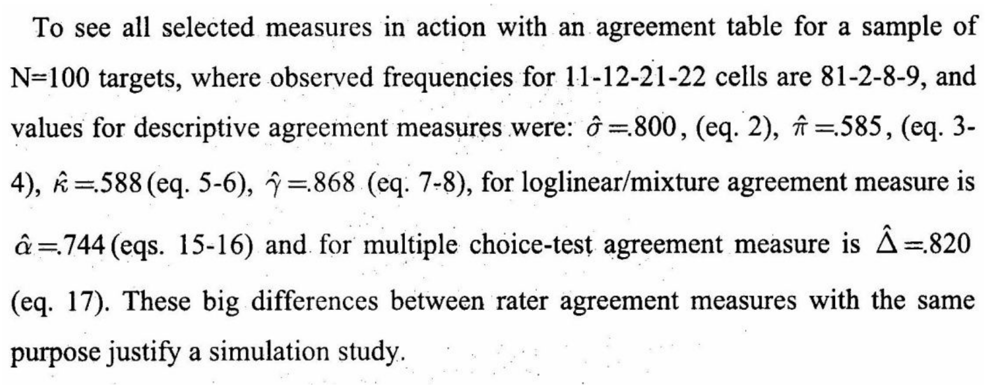
Martín & Femia (2004) proposed a new rater agreement measure, called Delta (Δ), which was developed in the context of multiple-choice test where a student has to choose among one of M possible responses for each of N targets known to the evaluator. If a student knows a fraction (say Δ =.4) of responses and fill out all the test, then it is assumed that the 40% of responses will be accurately recognized and the other 60% will be classified at random. The response model postulated for this situation (see Martín and Luna, 1989) is that the student will give a correct reply if the response is known and will pick a response at random if the response is unknown. In this case Δ is really a measure of the conformity of student with evaluator.

Δ-coefficient, as will be used in this paper, is the generalization of the situation of multiple choice-test where there is a sole object, to the situation of agreement where there are several objects and the intensity of recognitions for each object need not be necessarily the same (Martín and Femia, 2004).

Although inspired in a different tradition, in practice Δ-coefficient seems to reproduce exactly the same results than that obtained from *RCA(QI)* model (10) for the general *M* x *M* case. But for the 2 x 2 agreement table, as was pointed before, Δ and *RCA(QI)* cannot be estimated. Nevertheless Martín & Femia (2004: 9) suggested an simple asymptotic approximation of Δ -coefficient formulated as

$$\Delta = (p_{11} + p_{22}) - 2\sqrt{p_{12}p_{21}} \tag{17}$$

which can be used as a consistent rater agreement measure in this context. A more recent asymptotic re-formulation of Δ-coefficient (Martin & Femia, 2008: 766), which is specially useful for contingency tables with zero cells, is a simple modification of equation (16) which can be obtained in practice after adding the constant 1 to all the observed frequency cells of a contingency table, but it will not be used in this paper.



To see all selected measures in action with an agreement table for a sample of N=100 targets, where observed frequencies for 11-12-21-22 cells are 81-2-8-9, and values for descriptive agreement measures were: $\hat{\sigma}$=.800, (eq. 2), $\hat{\pi}$=.585, (eq. 3-4), $\hat{\kappa}$=.588 (eq. 5-6), $\hat{\gamma}$=.868 (eq. 7-8), for loglinear/mixture agreement measure is $\hat{\alpha}$=.744 (eqs. 15-16) and for multiple choice-test agreement measure is $\hat{\Delta}$=.820 (eq. 17). These big differences between rater agreement measures with the same purpose justify a simulation study.

A SIMULATION STUDY

A simulation study with these descriptive, loglinear/mixture and multiple-choice test rater measures for 2×2 agreement tables was performed varying parameters of prevalence of first or positive category (PCP, from 0.10 to 0.90 in steps of 0.10), discrimination probability (DP, from 0.50 to 0.90 in steps of 0.10) and number of targets (N of 30, 100 and 300). DP is a key variable that is used to capture the random responses of raters. It refers to the ability of raters to discriminate between easy and difficult to classify targets and so allows distinguishing between reliable (with high DP) and unreliable (with low DP) raters. So a total of 3 (number of targets, N) x 9 (levels of PCP) x 5 (levels of DP) combinations were generated for this simulation study.

The model of simulation process borrowed the essential aspects from the latent mixture and multiple-choice test models, but it also has some peculiarities added in order to refine the random behavior of raters in the choice of categories.

The process is initiated fixing the number N of targets, a prevalence value for the first or positive category value PCP and a discrimination probability DP. 1000 empty 2×2 agreement tables were generated for each combination of N, PCP and DP. In all cases, we assumed that discrimination probability was the same in both observers, a reasonable assumption if raters receive the same training as is usual with many rater agreement studies. Given both values of prevalence and discrimination, the average proportion of p_{11} would be PCP*DP+(1-DP)/4, which can be partitioned in a portion due to systematic agreement, (PCP*DP), and another due to random agreement, (1-DP)/4, whereas the average proportion of p_{22} would be (1-PCP)*DP+(1-DP)/4, also partitioned in a portion of systematic agreement, (1-PCP)*DP, and another of random agreement, (1-DP)/4.

For each case of an empty agreement table, the element was assigned to the positive or negative category (depending of the fixed PCP), and a random number *R* between 0 and 1 was generated and evaluated. If R £ DP, the case was included inside of discrimination range being easy to classify and so was considered as correctly classified. If R > DP, the case was considered as difficult to classify and so it was randomly assigned to any one of the four cells. In some cases where was detected the presence of zero values for one or more cells in an agreement table, the table was deleted and proceeding with the following. A few cases occurred in particular with extreme values of DP and number of targets of N=30 (see the total number of tables used for all combinations of PCP and DP in Table 2). A flow diagram of the simulation process is represented on Figure 1.

Since the raw observed agreement (ROA) is the sum of proportions for responses to diagonal cells of each agreement table, the simulation process distinguished between two components of ROA: systematic agreement (SA), as the non-random proportion of easy to classify targets and so correctly classified in the diagonal cells, and random agreement (RA), as the proportion of difficult to classify targets that were randomly classified in the diagonal cells.

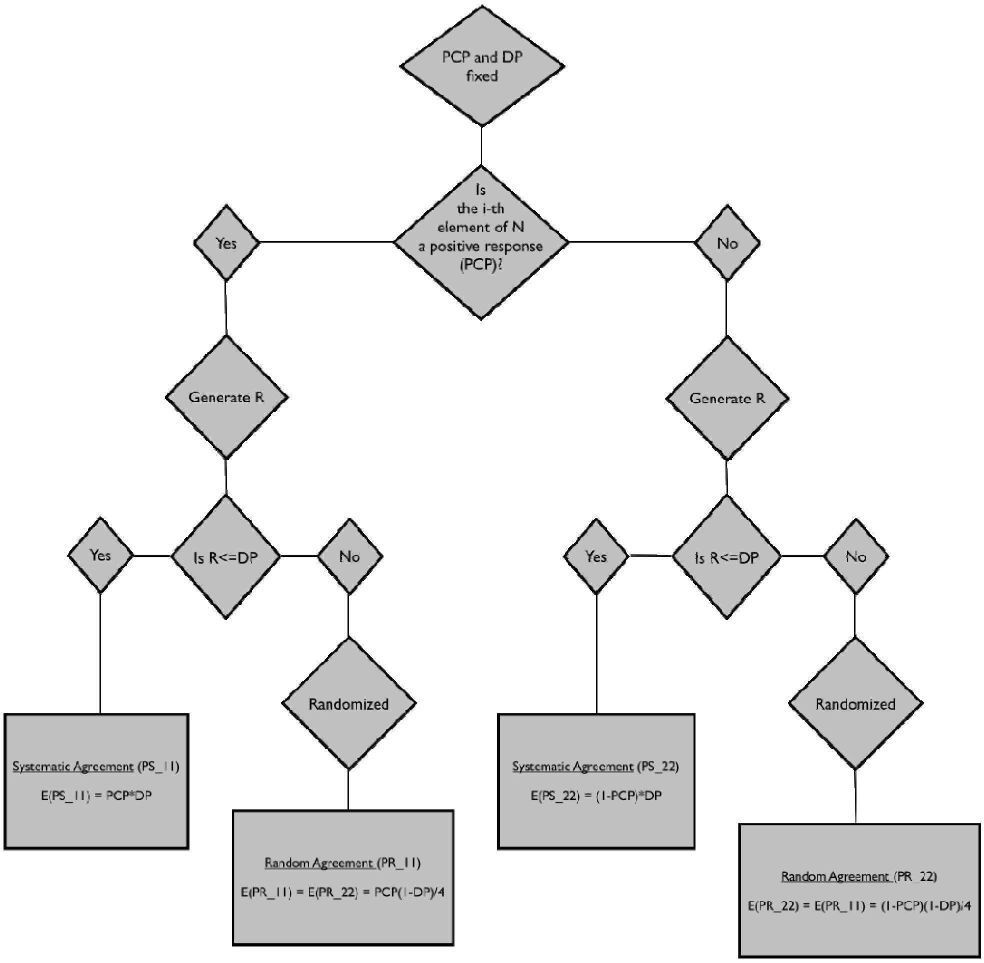

PCP and DP fixed
Is the i-th element of N a positive response (PCP)?
Yes
No
Generate R
Generate R
Yes
Is R<=DP
No
Yes
Is R<=DP
No
Randomized
Randomized
Systematic Agreement (PS_11)
E(PS_11) = PCP*DP
Random Agreement (PR_11)
E(PR_11) = E(PR_22) = PCP(1-DP)/4
Systematic Agreement (PS_22)
E(PS_22) = (1-PCP)*DP
Random Agreement (PR_22)
E(PR_22) = E(PR_11) = (1-PCP)(1-DP)/4

Figure 1. Flow diagram of the simulation process.

RESULTS

For each agreement table, the response of interest was focused on absolute bias, defined as the absolute difference between systematic agreement proportion (SA) and estimates for each one of six rater agreement measures. Absolute bias was firstly evaluated for all levels of discrimination (DP), prevalence of positive category (PCP) and number of targets (N), but due to the high similarity of results for different number of targets, all data from three options were merged. Table 2 shows the absolute mean bias (with standard errors in brackets) of six rater measures for a selected 20 of the 45 combinations of PCP and DP and the number of 2 x 2 contingency tables used for each combination with all merged data after discarding tables containing zero values on any of the four cells.

Three different sets of rater agreement measures emerged. All of them were easily detected at all levels of discrimination probability. Ordered from greater to smaller mean bias, the first set was integrated with Cohen's κ and Scott's π coefficients, the second one included Gwet's g and Aickin's α, and the third with Bennet's σ and Martin's Δ coefficients.

First and second sets scored similar absolute mean bias results with moderate prevalence levels ranged between .30 and .70, but diverged from these stable levels with more extreme prevalence values. We also noted that all rater measures presented a moderate deviation from a stable performance inside the range 30-70, but deviation was more marked with κ and π set, than with α and g set whereas σ and Δ set were near of unbiased performance.

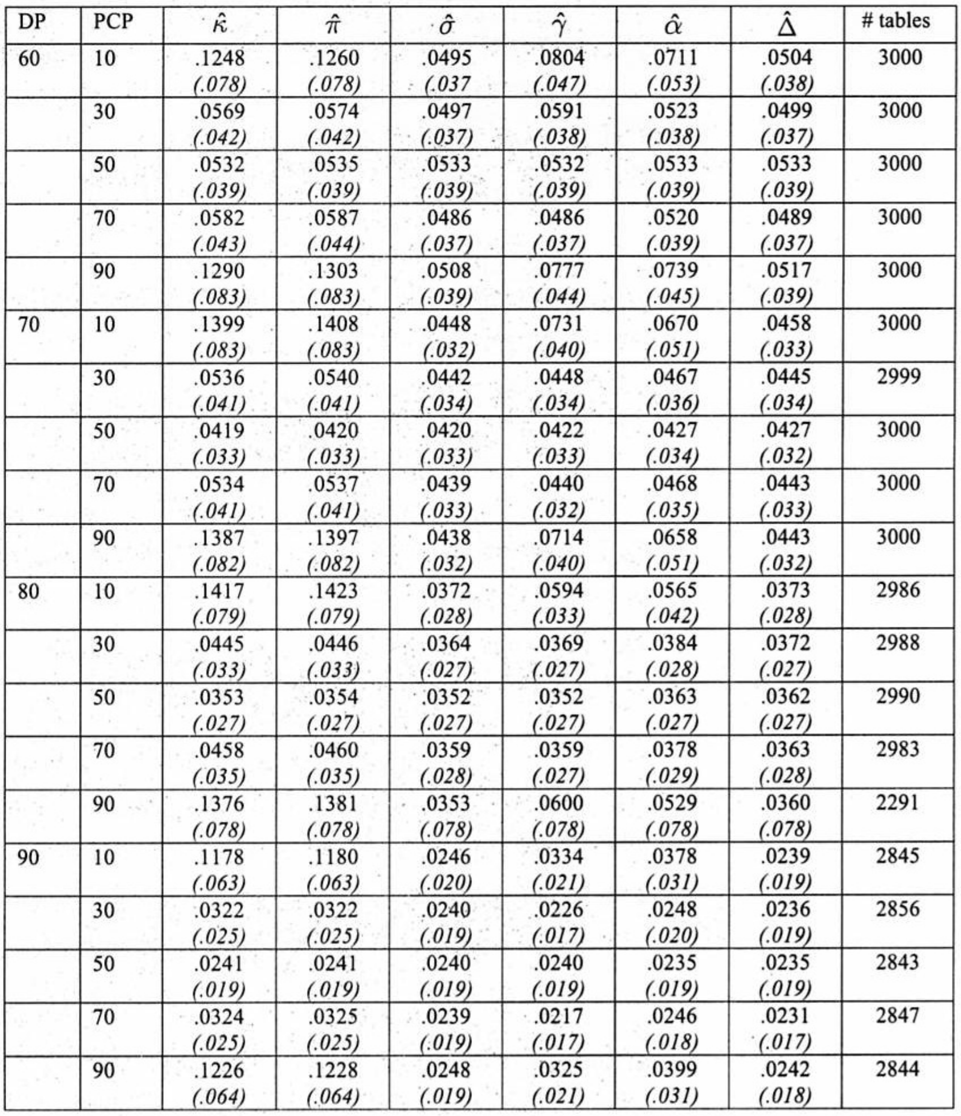
The behavior of the first set was the most biased of all rater agreement measures evaluated, particularly with low and high prevalence levels (see Figures 1 to 4). This result confirms a solid finding of research literature concerning the high sensibility of κ (and π) coefficients to trait prevalence in the population and reinforces the recommendation to avoid using any of both coefficients as reliable rater measures with extreme prevalence levels (Byrt, Bishop and Carlin, 1993; Feinstein and Cichetti, 1990; Lantz and Nebenzahl, 1996).

The behavior of the second set was less biased than the first set, particularly with higher levels of prevalence but the bias was the same with intermediate levels of prevalence for all DP levels. We were disappointed in particular with the behavior of Aickin's α-coefficient, but as was pointed before, the assumption of homogeneity of x parameter is a restrictive assumption for 2 x 2 agreement tables.

The behavior of the third set was excellent, scoring next of null mean bias for all levels of prevalence of positive category and for all levels of discrimination probability used. Both σ and Δ coefficients may be considered as unbiased rater agreement coefficients with 2×2 agreement tables. Although differences between measures of third set were negligible, the best behavior was related in all cases with the asymptotic approximation of Martín and Femia's Δ. σ coefficient had an excellent behavior with 2 x 2 agreement tables, but due to the uniform distribution of targets we suspect that it cannot be extrapolated to agreement tables of higher dimensionality where uniformity feature could be severely penalized.

Comparing the behavior of rater agreement measures between levels of discrimination probability, a noticeable result was that mean bias was higher in the range DP=.70-.80 than for DP ≤ .69 or DP ≥ .81, particularly with extreme levels of prevalence, and with first and second sets of rater measures, a strange result that deserve to be further investigated.

Table 2. Absolute mean bias (standard error in brackets) and number of tables used for selected combinations of discrimination (DP) and prevalence (PCP) for all agreement measures.



DP	PCP	$\hat{\kappa}$	$\hat{\pi}$	$\hat{\sigma}$	$\hat{\gamma}$	$\hat{\alpha}$	$\hat{\Delta}$	# tables
60	10	.1248 *(.078)*	.1260 *(.078)*	.0495 *(.037*	.0804 *(.047)*	.0711 *(.053)*	.0504 *(.038)*	3000
	30	.0569 *(.042)*	.0574 *(.042)*	.0497 *(.037)*	.0591 *(.038)*	.0523 *(.038)*	.0499 *(.037)*	3000
	50	.0532 *(.039)*	.0535 *(.039)*	.0533 *(.039)*	.0532 *(.039)*	.0533 *(.039)*	.0533 *(.039)*	3000
	70	.0582 *(.043)*	.0587 *(.044)*	.0486 *(.037)*	.0486 *(.037)*	.0520 *(.039)*	.0489 *(.037)*	3000
	90	.1290 *(.083)*	.1303 *(.083)*	.0508 *(.039)*	.0777 *(.044)*	.0739 *(.045)*	.0517 *(.039)*	3000
70	10	.1399 *(.083)*	.1408 *(.083)*	.0448 *(.032)*	.0731 *(.040)*	.0670 *(.051)*	.0458 *(.033)*	3000
	30	.0536 *(.041)*	.0540 *(.041)*	.0442 *(.034)*	.0448 *(.034)*	.0467 *(.036)*	.0445 *(.034)*	2999
	50	.0419 *(.033)*	.0420 *(.033)*	.0420 *(.033)*	.0422 *(.033)*	.0427 *(.034)*	.0427 *(.032)*	3000
	70	.0534 *(.041)*	.0537 *(.041)*	.0439 *(.033)*	.0440 *(.032)*	.0468 *(.035)*	.0443 *(.033)*	3000
	90	.1387 *(.082)*	.1397 *(.082)*	.0438 *(.032)*	.0714 *(.040)*	.0658 *(.051)*	.0443 *(.032)*	3000
80	10	.1417 *(.079)*	.1423 *(.079)*	.0372 *(.028)*	.0594 *(.033)*	.0565 *(.042)*	.0373 *(.028)*	2986
	30	.0445 *(.033)*	.0446 *(.033)*	.0364 *(.027)*	.0369 *(.027)*	.0384 *(.028)*	.0372 *(.027)*	2988
	50	.0353 *(.027)*	.0354 *(.027)*	.0352 *(.027)*	.0352 *(.027)*	.0363 *(.027)*	.0362 *(.027)*	2990
	70	.0458 *(.035)*	.0460 *(.035)*	.0359 *(.028)*	.0359 *(.027)*	.0378 *(.029)*	.0363 *(.028)*	2983
	90	.1376 *(.078)*	.1381 *(.078)*	.0353 *(.078)*	.0600 *(.078)*	.0529 *(.078)*	.0360 *(.078)*	2291
90	10	.1178 *(.063)*	.1180 *(.063)*	.0246 *(.020)*	.0334 *(.021)*	.0378 *(.031)*	.0239 *(.019)*	2845
	30	.0322 *(.025)*	.0322 *(.025)*	.0240 *(.019)*	.0226 *(.017)*	.0248 *(.020)*	.0236 *(.019)*	2856
	50	.0241 *(.019)*	.0241 *(.019)*	.0240 *(.019)*	.0240 *(.019)*	.0235 *(.019)*	.0235 *(.019)*	2843
	70	.0324 *(.025)*	.0325 *(.025)*	.0239 *(.019)*	.0217 *(.017)*	.0246 *(.018)*	.0231 *(.017)*	2847
	90	.1226 *(.064)*	.1228 *(.064)*	.0248 *(.019)*	.0325 *(.021)*	.0399 *(.031)*	.0242 *(.018)*	2844

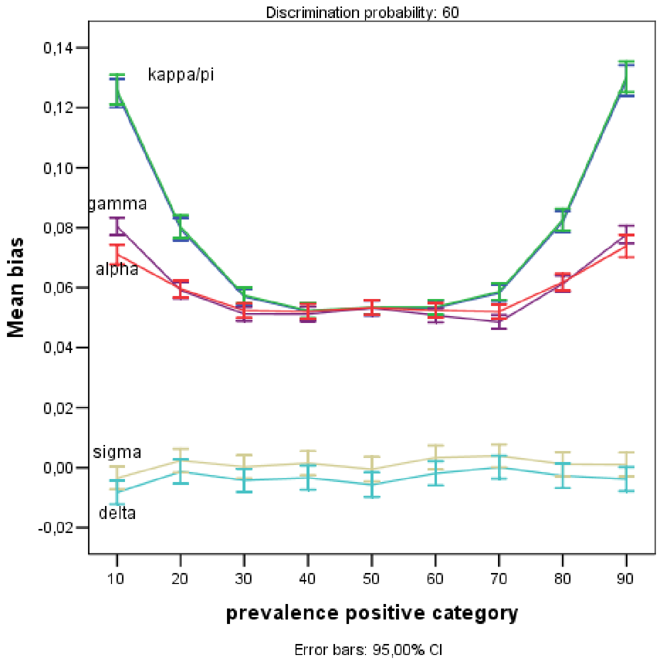


Figure 2. Behaviour of rater measures for DP=.60

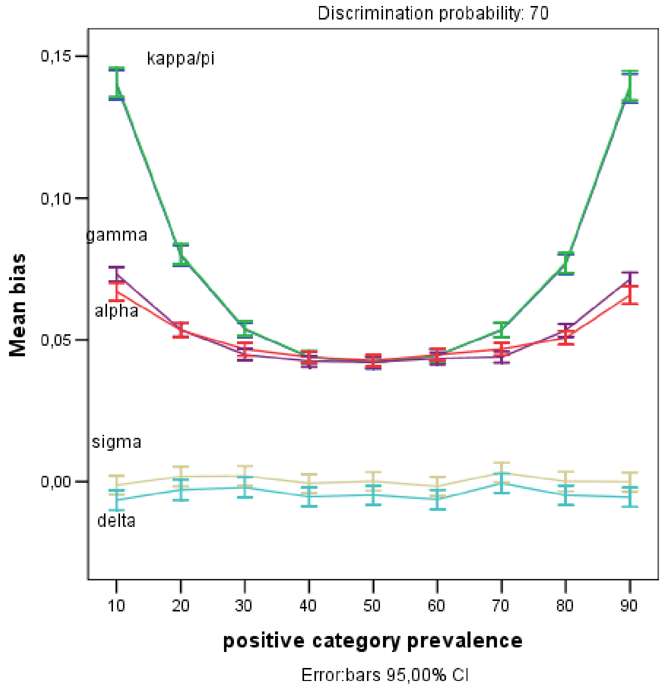


Figure 3. Behaviour of rater measures with DP = .70

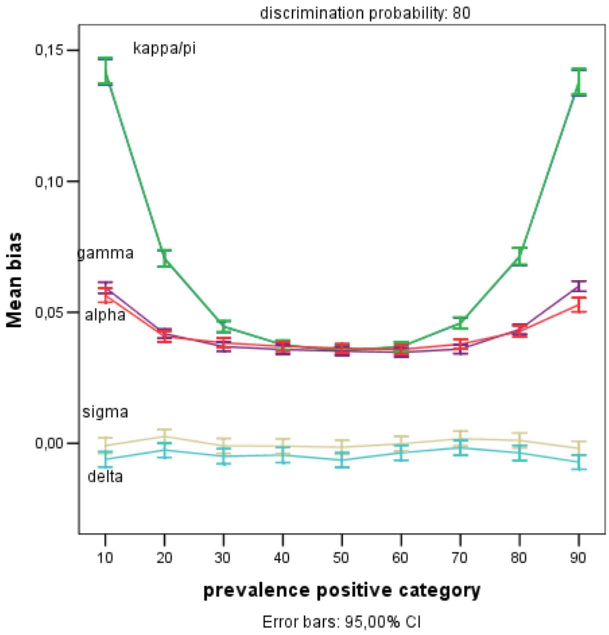


Figure 4. Behaviour of rater measures with DP = .80

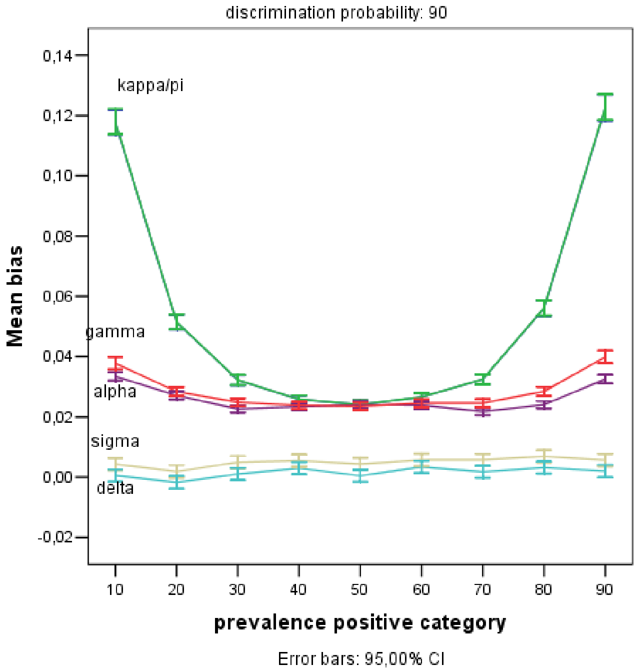


Figure 5. Behaviour of rater measures with DP = .90

RESUMEN

Un estudio de simulación de medidas de acuerdo entre observadores para tablas de contingencia 2x2. Mediante un estudio de simulación se aborda una comparación entre seis medidas obtenidas usando tres enfoques diferentes para la evaluación del acuerdo. Los coeficientes de acuerdo elegidos fueron σ de Bennet (1954), π de Scott (1955), κ de Cohen (1960) y γ de Gwet (2001; 2008) para representar el enfoque clásico descriptivo, el coeficiente α de Aickin (1990), para representar el enfoque de los modelos loglineal y mixtura ("mixture models") y la medida Δ de Martín and Femia (2004) para representar el enfoque de los test de elección multiple. Los resultados obtenidos confirman que los coeficientes π y κ presentan diferencias notables en relación a los restantes coeficientes particularmente en presencia de valores extremos de prevalencia y sesgo entre observadores. El mejor comportamento fue observado con los coeficientes σ de Bennet y Δ de Martín and Femia para todos los valores de prevalencia y sesgo entre observadores.

REFERENCES

Agresti, A. (1992). Modelling patterns of agreement and disagreement. *Statistical Methods in Medical Research, 1,* 201-218.

Agresti, A.; Ghosh, A. & Bini, M. (1995). Raking kappa: Describing potential impact of marginal distributions on measure of agreement. *Biometrical Journal, 37*, 811-820.

Agresti, A. (2002). *Categorical Data Analysis*. 2nd Edition. New York, NY: Wiley.

Aickin, M. (1970). Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's kappa. *Biometrics, 46,* 293-302.

Ato, M.; Benavente, A. y López, J.J. (2006). Análisis comparativo de tres enfoques para evaluar el acuerdo entre observadores. *Psicothema, 18*, 638-645.

Bennet, E.M.; Alpert, R. & Goldstein, A.C. (1954). Communications through limited response questioning. *Public Opinion Quarterly, 18*, 303-308.

Bloch, D.A. & Kraemer, H.C. (1989). 2×2 kappa coefficients: measures of agreement or association. *Biometrics, 45,* 269-287.

Brennan, R.L. & Prediger, D. (1981). Coefficient kappa: some uses, misuses and alternatives. *Educational and Psychological Measurement, 41,* 687-699.

Byrt, T.; Bishop, J. & Carlin, J.B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology, 46*, 423-429.

Carrasco, J.L. & Jover, Ll. (2003). Estimating the generalized concordance correlation coefficient through variance components. *Biometrics,* 59, 849-858.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Cronbach, L.J.; Gleser, G.C. & Rajaratnam, J. (1972). *The Dependability of Behavioral Measurements*. New York, NY: Wiley.

Darroch, J.M. & McCloud, P.I. (1986). Category distinguishability and observer agreement. *Australian Journal of Statistics, 28.* 371-388.

Dunn, C. (1989). *Design and Analysis of Reliability Studies: the statistical evaluation of measurement errors.* 2nd Edition. London: Arnold.

Feinstein, A. & Cichetti, D. (1990). High agreement but low kappa: I. The problem of two paradoxes. *Journal of Clinical Epidemiology, 43*, 543-549.

Fleiss, J.L.; Cohen, J. & Everitt, B.S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin, 72,* 323-327.

Graham, P. (1995). Modelling covariate effects in observer agreement studies: the case of nominal agreement. *Statistics in Medicine, 14*, 299-310.

Guggenmoos-Holtzmann, I. (1993). How reliable are chance-corrected measures of agreement. *Statistics in Medicine, 12*, 2191-2205.

Guggenmoos-Holtzmann, I. & Vonk, R. (1998). Kappa-like indices of observer agreement viewed from a latent class perspective. *Statistics in Medicine, 17*, 797-812.

Gwet, K. (2001). *Handbook of inter-rater reliability*. Gaithersburg, MA: Stataxis.

Gwet, K. (2008). Computing inter-rater reliability and its variance in presence of high agreement. *British Journal of Mathematical & Statistical Psychology*, *61*, 29-48.

Hoehler, F.K. (2000). Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *Journal of Clinical Epidemiology, 53*, 499-503.

Holley, W. & Guilford, J.P. (1964). A note on the G-index of agreement. *Educational and Psychological Measurement, 24,* 749-753.

Hsu, L.M. & Field, R. (2003). Interrater agreement measures: comments on kappa_n, Cohen's kappa, Scott's π and Aickin's α. *Understanding Statistics, 2*, 205-219.

Janson, S. & Vegelius, J. (1979). On generalizations of the G-index and the phi coefficient to nominal scales. *Multivariate Behavioral Research, 14*, 255-269.

Lantz, C.A. & Nebenzahl, E. (1996). Behavior and interpretation of the κ statistics: resolution of the two paradoxes. *Journal of Clinical Epidemiology, 49*, 431-434.

Lin, L.; Hedayat, A.S.; Sinha, B. & Yang, M. (2002). Statistical methods in assessing agreement: models, issues and tools. *Journal of the American Statistical Association, 97*, 257-270.

Martín, A. & Femia, P. (2004). Delta: a new measure of agreement between two raters. *British Journal of Mathematical and Statistical Psychology, 57*, 1-19.

Martin, A. and Femia, P. (2008). Chance-corrected measures of reliability and validity in 2 x 2 tables. Communications in Statistics – Theory and Methods, 37, 760-772.

Martín, A. & Luna, J.D. (1989). Tests and intervals in multiple choice tests: a modification of the simplest classical model. *British Journal of Mathematical and Statistical Psychology, 42,* 251-263.

McGraw, K.O. & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1,* 30-46

Maxwell, A.E. (1977). Coefficients of agreement between observers and their interpretation. *British Journal of Psychiatry, 116,* 651-655.

Schuster, C. (2002). A mixture model approach to indexing rater agreement. *British Journal of Mathematical and Statistical Psychology, 55*, 289-303.

Schuster, C. & von Eye, A. (2001). Models for ordinal agreement data. *Biometrical Journal, 43*, 795-808.

Schuster, C. & Smith, D.A. (2002). Indexing systematic rater agreement with a latent-class model. *Psychological Methods, 7*, 384-395.

Scott, W.A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly, 19*, 321-325.

Shoukri, M.M. (2004). *Measures of Interobserver Agreement*. Boca Raton, Fl. CRC Press.

Shrout, P.E, & Fleiss, J.L. (1973). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 2,* 420-428.

Tanner, M.A. & Young, M.A. (1985a). Modeling agreement among raters. *Journal of the American Psychological Association, 80,* 175-180.

Tanner, M.A. & Young, M.A. (1985b). Modeling ordinal scale disagreement. *Psychological Bulletin, 98*, 408-415.

Von Eye, A. & Mun, E.Y. (2005). *Analyzing Rater Agreement.* Mahwah, NJ: Lawrence Erlbaum Associates.

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin, 103,* 374-378.

(Manuscript received: 22 December 2009; accepted: 10 September 2010)