An Evaluation of the Impact of Design on the Analysis of Nominal-Scale Inter-Rater Reliability Studies

Kilem L. Gwet, PhD
Advanced Analytics, LLC., PO Box 2696, Gaithersburg, MD USA

February 10, 2018

Correspondence should be sent to Kilem L. Gwet, Advanced Analytics, LLC., PO Box 2696, Gaithersburg, MD USA. E-Mail: gwet@agreestat.com

Abstract

Inter-rater reliability studies have become an integral part of most data quality control programs used in various fields of research. The ultimate goal of these studies to quantify the extent of agreement among raters is often achieved without incorporating the specific way they were designed into the statistical procedures. As a result, the interpretation of agreement coefficients is often incomplete, and sometimes misleading. Many agreement coefficients in the literature implicitly assume a fully-crossed design where each rater is expected to rate all subjects even though fractional designs with different groups of raters assigned to different subjects are common. The purpose of this paper is to propose a new probabilistic methodology for interpreting agreement coefficients based on their design-based standard errors. Simulation results demonstrating the validity of the design-based standard errors are presented. We also provide guidelines for determining the minimum sample sizes required to obtain optimal study designs.
keywords: inter-rater reliability; Cohen's kappa coefficient; Gwet's AC2; Fleiss' kappa.

1 Introduction

Two raters with a high inter-rater reliability coefficient can rate subjects interchangeably with minimal rater effect. Any variation in the ratings is then seen as an attribute of the subjects, and not that of the raters nor the subject-rater interaction. Only under this ideal scenario may a researcher use the ratings to qualify subjects with respect to the phenomenon under investigation.

Several methods aimed at quantifying the extent of agreement between two raters were proposed in the literature, most of which being expressed in the form $\kappa = (p_a - p_e)/(1 - p_e)$, where p_a is the percent agreement, and p_e the percent chance agreement. While many known agreement coefficients share the same percent agreement, they tend to differ in the way they quantify the percent chance agreement. Guttman (1945), Bennett et *al.* (1954), Holley and Guilford (1964), Maxwell (1977), as well as Janson and Vegelius (1979) independently developed the same agreement coefficient for 2 raters and 2 categories, evaluating the percent chance agreement as the inverse of the number of categories used in the experiment. Brennan and Prediger (1981) later extended it to an arbitrarily large number of categories. Still in the context of 2 raters, Scott (1955), and Cohen (1960) both proposed two agreement coefficients that compute the percent chance agreement based on the observed raters' classification probabilities. Both coefficients are known in the literature as Scott's π and Cohen's κ. Cohen's kappa gained in popularity among researchers since it was published, despite the fact that Scott's Pi was published 5 years earlier, and that both suffer from serious and well-documented deficiencies (see Cicchetti and Feinstein, 1990, or Feinstein and Cicchetti, 1990). Alternatives to Scott's Pi and Cohen's Kappa include Brennan-Prediger coefficient (see Brennan and Prediger, 1981), or Gwet's AC$_1$ and AC$_2$ (see Gwet, 2008 and Gwet, 2014), two coefficients that are more resistant to high prevalence and lack of uniformity in raters' classification probabilities.

All two-rater agreement coefficients are based on 3 fundamental and often overlooked assumptions that define their statistical framework:

- Study findings will not be extended beyond the 2 participating raters.
- Each rater is expected to rate all subjects recruited for the experiment.
- The subjects recruited for the experiment constitute a representative

sample of a larger population of subjects they were selected from.

Even if all these 3 conditions are violated, one can still use Cohen's equation and claim to have calculated kappa. For example two-rater inter-rater reliability experiments where the 2 raters change from subject to subject are common in practice (see Hallgren, 2012). Although this framework is different from that of Cohen (1960, 1968), kappa is still often used. Here is a typical situation where the study design is ignored in the calculation and interpretation of an agreement coefficient. We will show in this paper that using the correct standard error associated with the agreement coefficient is pivotal for incorporating the design into the analysis of inter-rater reliability studies. Fleiss, Cohen, and Everitt (1968) proposed valid standard errors of kappa and weighted kappa. Unfortunately, researchers are not using these expressions very often. This is mainly due to a lack of clear guidelines for using them when interpreting the magnitude of agreement coefficients.

Fleiss (1971) proposed the first extention of Cohen's kappa coefficient to multiple raters. Conger (1980), Light (1971), and Gwet (2008*a*) recommended alternative approaches. All these generalizations kept the original two-rater format of agreement coefficients as the ratio $(p_a - p_e)/(1 - p_e)$ of the propensity of agreement beyond chance to the propensity of non chance agreement.

Almost all multiple-rater agreement coefficients in the inter-rater reliability literature are formulated under the assumption that each rater rates all subjects. That is n subjects and r raters participating in a study will produce a total of rn ratings. Let us refer to this as a *fully-crossed design*. For scoring processes that are demanding, the researcher often reduces the workload of raters by assigning different groups of 2, 3, or 4 raters to different subjects. Although Fleiss' generalized kappa coefficient may still be calculated under this fractional design, its interpretation however will be different.

For the remaining part of this paper, we will assume that n subjects were randomly selected from the original subject population comprising N subjects. At times one may assume $n = N$, indicating that all subjects of interest are recruited for the study. Moreover, r raters are randomly selected from an original universe of R raters. Once again, for some studies we could have $r = R$, indicating no sampling of the rater population.

2 Design of Inter-Rater Reliability Studies

There is a long list of factors that can be considered when designing an inter-rater reliability study. This includes the type of rating scale, the number of categories in that scale, the number of subjects, the number of raters and how they are assigned to subjects or the number of subjects. We have decided to focus on the following 3 specific aspects of study design with a direct impact on the statistical properties of agreement coefficients:

- *Who are the subjects and raters targeted by the experiment?*
- *How many subjects and raters should be recruited for the study? How will they be selected?*
- *How would selected raters be assigned to the selected subjects?*

Defining the Target Rater and Subject Populations

The target here represents those raters and subjects to whom the researcher would like to apply the study findings. These target groups will often be too large to be included in an inter-rater reliability experiment in their entirety. In this case, only a sample of subjects and/or raters carefully selected from their respective groups will participate in the experiment. Regardless of who participate in the experiment, we still want to be able to extrapolate our results to the entire target populations of raters and subjects. Let us assume that the target rater universe contains R raters, and the target universe of subjects N subjects, although R and N are often unknown in practice.

Choosing Participating Raters and Subjects

Cost, time, and other practical constraints often lead researchers to use subsets from target subject and rater universes. Let r and n be respectively the number of raters and subjects who will participate in the experiment. Since the selection of samples does not change our focus on the predetermined target populations, we will want the r selected raters, and the n selected subjects to be representative samples of their respective populations. This is often achieved by randomizing the sample selection process, and by ensuring an adequate number of units in the sample. In section 7, we will provide guidelines for determining these sample sizes.

Assigning Raters to Subjects

Once the participating raters and subjects are identified, the researcher must answer the question as to "what rater must rate what subject?" In this paper we will confine ourselves to the following 3 designs:

(i) The fully-crossed design with no sampling of raters (let's refer to this design as "FC_1")
Under this design, each of the r participating raters (with $r = R$) must rate all n participating subjects (see Table 1). Agreement coefficients are subjected to a single source of a variation due to the random selection of subjects. FC_1 stands for "Fully Crossed" with 1 source of variation.

TABLE 1

Table 1: FC_1 Design: Each rater of interest scores all sampled (Only subjects 1, 2, 4, and 6 are selected in the sample) subjects

Subjects	Rater 1	Rater 2	Rater 3
Subject 1	X	X	X
Subject 2	X	X	X
Subject 3			
Subject 4	X	X	X
Subject 5			
Subject 6	X	X	X

(ii) The fully-crossed design with a sample of raters (let's refer to this design as "FC_2")
Under this plan, each of the r participating raters must rate all n participating subjects (see Table 2). The main difference from the FC_1 design is the new source of variation due to the random selection of raters in addition to the random selection of subjects. FC_2 stands for Fully-Crossed with 2 sources of variation.

TABLE 2

Table 2: FC_2 Design: Each sampled (Raters 2, 4, and 5 are selected in the rater sample, while subjects 1, 2, 4, and 6 are selected in the subject sample) rater scores all sampled subjects

Subjects	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Subject 1		X		X	X
Subject 2		X		X	X
Subject 3					
Subject 4		X		X	X
Subject 5					
Subject 6		X		X	X

(iii) The partially-crossed design with a 2-rater sample per subject (let's refer to this design as PC_2 for Partially Crossed with 2 sources of variation) Under this plan, each of the n subjects is assigned 2 raters randomly chosen from the target population of R raters of interest. This design is cost-effective since it minimizes the number of ratings per subject.

TABLE 3

Table 3: PC_2 Design: Each sampled subject is scored by 2 raters randomly chosen from the universe of all raters of interest

Subjects	Rater 1	Rater 2	Rater 3
Subject 1	X		X
Subject 2		X	X
Subject 3			
Subject 4	X	X	
Subject 5			
Subject 6		X	X

3 Using the Standard Error for Interpreting the Magnitude of Agreement Coefficients

Researchers commonly use benchmark scales to interpret the magnitude of agreement coefficients. One such benchmark scale was proposed by Landis and Koch (1977) and described in Table 4. An agreement coefficient that falls between 0.41 and 0.60 for example is considered to be "Moderate." This simplistic interpretation ignores the important fact that an estimated agreement coefficient is calculated with a margin of error, making it impossible to categorize it with certainty. It can only be done with some degree of certainty. Therefore, the correct interpretation of these coefficients must be probabilistic.

If two agreement coefficients have the exact same magnitude with one having a much larger standard error, intuitively one would expect the coefficient with the smaller standard error to be more likely to belong to the high-end intervals of the benchmark scale. Consider two agreement coefficients $\widehat{\kappa}_1 = 0.67$ with standard error $se = 0.15$, and $\widehat{\kappa}_2 = 0.67$ with standard error $s.e. = 0.04$. Based on Table 4, one would naturally qualify both agreement coefficients as being "Substantial" since 0.67 falls within the 0.61-0.8. Considering that the estimated agreement coefficient 0.67 is calculated with a margin of error, the validity of such a statement comes into question. A more elaborate approach for interpreting $\widehat{\kappa}_1$ and $\widehat{\kappa}_2$ and which is based on the notion of cumulative interval membership probability (CIMP) is summarized in Table 5.

The "Cumulative Probability" Approach to Benchmarking

It follows from Table 5 that the interval membership probability (IMP) and cumulative probability are calculated for each agreement coefficients $\widehat{\kappa}_1$ and $\widehat{\kappa}_2$. The interval probability represents the Normality-based probability that the "true" agreement coefficient κ belongs to the interval in question, and is calculated based on $\widehat{\kappa}_1$ and an arbitrary interval (a, b) as follows:

$$\begin{aligned} P(a \leq \kappa_1 \leq b) &= P\big[(\widehat{\kappa}_1 - b)/se(\widehat{\kappa}_1) \leq Z \leq (\widehat{\kappa}_1 - a)/se(\widehat{\kappa}_1)\big], \\ &= \Phi\big[(\widehat{\kappa}_1 - a)/se(\widehat{\kappa}_1)\big] - \Phi\big[(\widehat{\kappa}_1 - b)/se(\widehat{\kappa}_1)\big], \end{aligned} \tag{1}$$

where Φ is the cumulative distribution function of the standard Normal distribution. The Normality of many agreement coefficients has been demonstrated by Gwet (2008*b*). Using the $\widehat{\kappa}_1$ coefficient and looking at the Cumulative Probability column, it appears that the "true" agreement coefficient is Almost Perfect with a small probability probability 0.179, but is Substantial or Almost Perfect with a cumulative probability of 0.666. It will be Moderate or better with a much higher cumulative probability of 0.95. In this case we recommend to consider the agreement $\widehat{\kappa}_1$ as moderate. The general rule consists of retaining the highest interval whose CIMP equals or exceeds the threshold of 0.95. Using the same approach with the second agreement coefficient $\widehat{\kappa}_2$, we conclude that it is substantial. As one would notice, the standard error of the agreement coefficient plays a pivotal in the benchmarking process based on the cumulative probability method. Therefore, it is essential to be able to compute the correct standard error of various agreement coefficients under the specific study design that was implemented.

TABLE 4

Table 4: Landis and Koch Kappa's Benchmark Scale

Kappa Statistic	Strength of Agreement
< 0.0	Poor
0.0 to 0.20	Slight
0.21 to 0.40	Fair
0.41 to 0.60	Moderate
0.61 to 0.80	Substantial
0.81 to 1.00	Almost Perfect

TABLE 5

4 Inter-Rater Reliability under the FC$_1$ Design

In this section, we consider the FC$_1$ experimental design where each of the r raters in the target rater population must rate all the n participating subjects. We also discuss the weighted versions of 3 agreement coefficients: (1) Percent

Table 5: Benchmarking Agreement Coefficients

		$\widehat{\kappa}_1$ Coefficient		$\widehat{\kappa}_2$ Coefficient	
Interpretation	Intervals	Interval Probability	Cumulative Probability	Interval Probability	Cumulative Probability
Almost Perfect	0.8-1	0.179	0.179	0.001	0.001
Substantial	0.6-0.8	0.487	0.666	0.959	**0.960**
Moderate	0.4-0.6	0.284	**0.950**	0.040	1
Fair	0.2-0.4	0.035	0.985	0	1
Slight	0-0.2	0	0.986	0	1
Poor	0	0	1	0	1

agreement, (2) Fleiss' generalized kappa, and (3) Gwet's AC$_2$. The notion of weights was introduced by Cohen (1968) to account for partial agreement. Let r_{ik} be the number of raters who classified subject i into category k, and $r_{ik}^\star$ the weighted number of raters who classified subject i into category k (or other categories representing partial agreement with k), and defined as follows:

$$r_{ik}^\star = \sum_{l=1}^{q} w_{kl} r_{il}, \tag{2}$$

where w_{kl} is the weight value associated with categories k and l. Some of the equations presented here were discussed by Gwet (2008*a*, 2014).

- *Percent Agreement*
  All agreement coefficients discussed in this section share the same percent agreement p_a given by:

$$p_a = \frac{1}{n}\sum_{i=1}^{n} p_{a|i}, \text{ where } p_{a|i} = \sum_{k=1}^{q} \frac{r_{ik}(r_{ik}^\star - 1)}{r(r-1)}. \tag{3}$$

The variance of the percent agreement under the FC$_1$ design is given by,

$$v(p_a) = \frac{1-f}{n}\frac{1}{n-1}\sum_{i=1}^{n}(p_{a|i} - p_a)^2. \tag{4}$$

- *Fleiss' Generalized Kappa*
  Fleiss(1971) proposed a generalized version of Cohen's two-rater kappa coefficient, which may be used with 3 raters or more, and is defined as follows:

$$\widehat{\kappa}_{\text{F}} = \frac{p_a - p_e}{1-p_e},\ \text{where } p_e = \sum_{k,l} w_{kl}\pi_k\pi_l \text{ and } \pi_k = \frac{1}{n}\sum_{i=1}^{n} r_{ik}/r. \tag{5}$$

  The variance of Fleiss' generalized kappa coefficient under the FC_1 design is given by,

$$v\big(\widehat{\kappa}_{\text{F}}\big|\text{FC}_1\big) = \frac{1-f}{n}\frac{1}{n-1}\sum_{i=1}^{n}\big(\kappa^{\star}_{\text{F}|i} - \widehat{\kappa}_{\text{F}}\big)^2,$$

  where $\kappa^{\star}_{\text{F}|i}$ is calculated as follows:

$$\begin{aligned}\kappa^{\star}_{\text{F}|i} &= \kappa_{\text{F}|i} - 2(1-\widehat{\kappa}_{\text{F}})(p_{e|i} - p_e)/(1-p_e),\\ &\textit{and } \kappa_{\text{F}|i} = (p_{a|i} - p_e)/(1-p_e).\end{aligned}$$

  Moreover, $p_{e|i}$ is given by,

$$p_{e|i} = \sum_{k=1}^{q}\pi^{\star}_k r_{ik}/r \text{ with } \pi^{\star}_k = \sum_{l=1}^{q} w_{kl}\pi_l. \tag{6}$$

- *Gwet's $AC_{\mathbf{2}}$ Coefficient*
  Gwet (2008*a*) proposed the AC_2 coefficient as a paradox-resistant alternative agreement statistic. This coefficient is known to be less sensitive to trait prevalence and marginal homogeneity than other coefficients, and is defined as follows:

$$\widehat{\kappa}_{\text{G}} = \frac{p_a - p_e}{1-p_e},\ \text{where } p_e = \frac{T_w}{q(q-1)}\sum_{k=1}^{q}\pi_k(1-\pi_k) \text{ and } T_w = \sum_{k,l} w_{kl}. \tag{7}$$

  Its variance under the FC_1 design is given by,

$$v\big(\widehat{\kappa}_{\text{G}}\big|\text{FC}_1\big) = \frac{1-f}{n}\frac{1}{n-1}\sum_{i=1}^{n}\big(\kappa^{\star}_{\text{G}|i} - \widehat{\kappa}_{\text{G}}\big)^2, \tag{8}$$

where $\kappa^{\star}_{\text{G}|i}$ is obtained as follows:

$$\begin{aligned}\kappa^{\star}_{\text{G}|i} &= \kappa_{\text{G}|i} - 2(1-\widehat{\kappa}_{\text{G}})(p_{e|i}-p_e)/(1-p_e),\\ &\textit{and } \kappa_{\text{G}|i} = (p_{a|i}-p_e)/(1-p_e).\end{aligned} \tag{9}$$

The i^{th}-element $p_{e|i}$ of the percent chance agreement is defined as follows:

$$p_{e|i} = \frac{T_w}{q(q-1)}\sum_{k=1}^{q}(1-\pi_k)r_{ik}/r. \tag{10}$$

The FC_2 design does not affect the way agreement coefficients are calculated. However, the variance under the FC_2 design has two components: a subject component and a rater component. The subject component is the variance under the FC_1 design, and the rater component is calculated using the jackknife method previously advocated by Kraemer (1979). For the AC_2 coefficient for example, the variance under the FC_2 design is defined as follows:

$$v\big(\widehat{\kappa}_{\text{G}}\big|\text{FC}_2\big) = v\big(\widehat{\kappa}_{\text{G}}\big|\text{FC}_1\big) + v_{\text{R}}\big(\widehat{\kappa}_{\text{G}}\big), \tag{11}$$

where $v\big(\widehat{\kappa}_{\text{G}}\big|\text{FC}_1\big)$ the subject variance component is given by equation 8, and the rater variance component given by,

$$v_{\text{R}}\big(\widehat{\kappa}_{\text{G}}\big) = \frac{r-1}{r}\sum_{g=1}^{r}\Big(\widehat{\kappa}_{\text{G}}^{(-g)} - \widehat{\kappa}_{\text{G}}\Big)^2, \tag{12}$$

with r being the number of raters. Note that $\widehat{\kappa}_{\text{G}}^{(-g)}$ represents the AC_2 coefficient evaluated after removing all rater g's ratings from the sample.

5 Inter-Rater Reliability under the PC_2 Design

In this section, we consider inter-rater reliability experiments that aim at quantifying the extent of agreement among 3 raters or more (i.e. $r \geq 3$). Under the PC_2 design, we assume that the selection of pairs of raters is done with replacement, which opens up the possibility that the same pair might be selected more than once and have to score 2 subjects or more.

Cohen (1960), and Fleiss, Cohen, and Everitt (1969) made their derivations under the assumption that a single pair of raters rated all n subjects. The same can be said about Scott's Pi. Although the form taken by many multiple-rater agreement coefficients under the $\textsc{pc}_2$ design may look similar to their two-rater versions, they are different. Using the names Cohen's kappa, or Scott's Pi, or Gwet's AC$_2$ under the PC$_2$ design when different pairs of raters rate different subjects can be seen as an abuse of language.

The variance under the PC$_2$ design must be carefully derived and explained. For the sake of clarity, we will first present the exact variance (also referred to as population or theoretical variance) before showing how it can be calculated from actual ratings. The exact variance is the sum of two components: (1) the subject variance due to the sampling of subjects, and (2) the rater variance due to the random assignment of pairs of raters to subjects.

- *Fleiss' Generalized Kappa*
  Fleiss' generalized kappa under the $\textsc{pc}_2$ design looks like Scott's Pi and is formulated as follows:

$$\widehat{\kappa}_{2\text{F}} = \frac{p_a - p_e}{1 - p_e}, \text{ where } p_a = \sum_{k=1}^{q}\sum_{l=1}^{q} w_{kl}p_{kl},\; p_e = \sum_{k,l}^{q} w_{kl}\pi_k\pi_l. \tag{13}$$

  Unlike the classical Scott's Pi coefficient under the $\textsc{fc}_1$ design, this coefficient is calculated using $2n$ ratings produced not by a single pair of raters, but by several pairs of raters randomly assigned to subjects. Under the $\textsc{fc}_1$ design p_{kl} for example represents the relative number of subjects classified into categories k and l by the only 2 raters who participated in the experiment. Under the $\textsc{pc}_2$ design, p_{kl} represents the relative number of subjects classified into categories k and l by the different pairs of raters who rated them. The exact variance of $\widehat{\kappa}_{2\text{F}}$ under the $\textsc{pc}_2$ design is given by,

$$V\big(\widehat{\kappa}_{2\text{F}}\big|\textsc{pc}_2\big) = V\big(\widehat{\kappa}_{\text{F}}\big|\textsc{fc}_1\big) + \overline{\text{S}}_2/n, \tag{14}$$

  where $V\big(\widehat{\kappa}_{\text{F}}\big|\textsc{fc}_1\big)$ is the variance under the $\textsc{fc}_1$ design of Fleiss' coefficient (c.f. equation 5), and $\overline{\text{S}}_2/n$ the variance component due to the random assignment of raters. These two statistics are defined as follows:

$$V\big(\widehat{\kappa}_{\text{F}}\big|\textsc{fc}_1\big) = \frac{1-f}{n}\frac{1}{N-1}\sum_{i=1}^{N}\big(\kappa^{\star}_{\text{F}|i} - \kappa_{\text{F}}\big)^2, \quad \overline{\text{S}}_2 = \frac{1}{N}\sum_{i=1}^{N}\text{S}_i^2, \tag{15}$$

where the i^{th} linear component $\kappa^{\star}_{\text{F}|i}$ of Fleiss' coefficient is given by equation 4, while S_i^2 is the rater variance induced by the random assignment of a pair of raters to subject i. A mathematical proof of equation 14 can be found in Section 1 of the Supplemental Materials.

Note that $\kappa^{\star}_{\text{F}|i}$ can also be evaluated using only the ratings generated by a single pair of raters named g, in which case it is denoted by $\kappa^{\star(g)}_{\text{F}|i}$. To obtain S_i^2, for a specific subject i one needs to compute a series of values $\kappa^{\star(g)}_{\text{F}|i}$ for all $r(r-1)/2$ pairs of raters that can be formed out of the group of r raters, and to compute the variance of that series.

The theoretical variance under the PC$_2$ design of Fleiss'coefficient given by equation 14 was translated into FC$_1$ design-based measures to facilitate comparison between both designs. This equation shows that the use of the PC$_2$ design results in an increase of variance (i.e. a loss of precision) compared to the FC$_1$ design. Since only one design can be implemented at a time, equation 14 is purely definitional and cannot be used to compute the variance of Fleiss' kappa with actual ratings collected under the PC$_2$ design. This calculation can be done using the following expression:

$$v(\widehat{\kappa}_{2\text{F}}\big|\text{PC}_2) = \frac{1-f}{n}\frac{1}{n-1}\sum_{i,g}\big(\kappa^{\star(g)}_{\text{F}|i} - \widehat{\kappa}_{2\text{F}}\big)^2, \tag{16}$$

where $\kappa^{\star(g)}_{\text{F}|i}$ comes from equation 4 all calculations being based on the specific pair of raters g that was assigned to subject i. Simulation results from section 6 prove the validity of both equations 14 and 16.

- *Gwet's AC*$_\mathbf{2}$
  Under the PC$_2$ design, the weighted AC$_1$ also known as AC$_2$ takes the following form:

$$\widehat{\kappa}_{2\text{G}} = \frac{p_a - p_e}{1-p_e},\ \text{where}\ p_a = \sum_{k=1}^{q}\sum_{l=1}^{q} w_{kl}p_{kl},\ p_e = \frac{t_w}{q(q-1)}\sum_{k=1}^{q}\pi_k(1-\pi_k), \tag{17}$$

where t_w is the summation of all weights. This coefficient is calculated using $2n$ ratings generated by several pairs of raters randomly assigned to subjects. The theoretical variance of Gwet's AC$_2$ coefficient under the PC_2 design is given by (see proof in Section 2 of the Supplementary Materials),

$$V\big(\widehat{\kappa}_{2\text{G}}\big|\text{PC}_2\big) = V\big(\widehat{\kappa}_{\text{G}}\big|\text{FC}_1\big) + \overline{\text{S}}_2/n, \tag{18}$$

where $\widehat{\kappa}_{\text{G}}$ is the multiple-rater version of Gwet's AC_2 coefficient of equation 7 under the FC_1 design, $V\big(\widehat{\kappa}_{\text{G}}\big|\text{FC}_1\big)$ is the variance of AC_2 under the FC_1 design (estimated by equation 8), and $\overline{\text{S}}_2/n$ the variance component due to the random assignment of raters to subjects. These two statistics are defined as follows:

$$V\big(\widehat{\kappa}_{\text{G}}\big|\text{FC}_1\big) = \frac{1-f}{n}\frac{1}{N-1}\sum_{i=1}^{N}\big(\kappa^{\star}_{\text{G}|i} - \kappa_{\text{G}}\big)^2, \text{ and } \overline{\text{S}}_2 = \frac{1}{N}\sum_{i=1}^{N}\text{S}_i^2, \tag{19}$$

where $\kappa^{\star}_{\text{G}|i}$, the i^{th} linear component of Gwet's coefficient is given by equation 9, while S_i^2 is the rater variance induced by the random assignment of a pair of raters to subject i. S_i^2 is obtained for a specific subject i, by calculating a simple variance of the series of values $\kappa^{\star(g)}_{\text{G}|i}$ for all pairs of raters $g = 1\cdots r(r-1)/2$ that can be formed out of the group of r raters.

Again equation 18 is primarily definitional, and the variance should be calculated using the following equation:

$$v\big(\widehat{\kappa}_{2\text{F}}\big|\text{PC}_2\big) = \frac{1-f}{n}\frac{1}{n-1}\sum_{i,g}\big(\kappa^{\star(g)}_{\text{G}|i} - \widehat{\kappa}_{2\text{G}}\big)^2, \tag{20}$$

where all calculations are based on the specific pair of raters g that was assigned to subject i under the PC_2 design. Simulation results from section 6 confirm the validity of both equations 18 and 20.

- *Percent Agreement*

  The variance of the percent agreement p_a is higher under the PC_2 design than under the traditional FC_1 design. In fact,

  $$V\big(p_{2a}\big|\text{PC}_2\big) = V\big(p_a\big|\text{FC}_1\big) + \overline{\text{S}}_{2a}/n, \text{ where } \overline{\text{S}}_{2a} = \frac{1}{N}\sum_{i=1}^{N}P_{a|i}(1-P_{a|i}). \tag{21}$$

  Under the FC_1 design, p_a is given by equation 3, and is defined as in equations 13, and 17 under the PC_2 design. If the sampling fraction $f =$

n/N is negligible as is often the case, then $V\big(p_a\big|\text{PC}_2\big) = P_a(1-P_a)/n$, where P_a is the average of all pairwise percent agreement coefficients calculated from the r raters in the sample. The proof of this results can be found in Section 3 of the Supplementary Materials. Under the PC$_2$ design, this variance can be estimated with the following equation:

$$v\big(p_{2a}\big|\text{PC}_2\big) = p_a(1-p_a)/n. \tag{22}$$

6 Monte-Carlo Simulations

The Monte-Carlo experiment presented in this section focuses on the PC$_2$ design and aims at the following two goals:

(i) Demonstrate the validity of the theoretical variances of equations 14, 18, and 21,

(ii) Demonstrate the validity of the estimated variances under the PC$_2$ design (see equations 16, 20, and 22).

To achieve these two goals, we set up a Monte-Carlo experiment as follows:

(a) We considered a hypothetical target population of $N = 2500$ subjects from which small samples of n subjects are selected and rated by r raters using a 3-item Likert scale ($q = 3$). We investigated 9 n values ranging from 10 to 50 by increment of 5, and 2 r values 3, and 5.

(b) We created 2500 ratings for each of the r raters to have a population in the form of a $2500\times r$ matrix of ratings (i.e. we created one population of 7,500 ratings for $r = 3$, and a second population of 12,500 ratings for $r = 5$), the rating being one of the values 1, 2, or 3. These ratings were created randomly under the constraint that they must achieve a percent agreement that exceeds 0.65, an arbitrary number that ensures a certain level of agreement among raters.

(c) For each n value we randomly selected 5,000 different sets of n rows from the original 2,500$\times r$ matrix of ratings. Each selected set contained n rows and r columns and represents a simulated sample under the FC$_1$ design.

(d) To simulate a PC$_2$ sample when $r = 5$ for example, we took each FC$_1$ sample and only kept 2 ratings per subject randomly chosen from the 5 available. This process was implemented independently for each row. A simulated FC$_1$ sample contains $n \times 5$ ratings while a simulated PC$_2$ sample contains $n \times 2$ ratings extracted from the FC$_1$ sample. Note that for each FC$_1$ sample, we simulated a total of 5,000 PC$_2$ samples (or less depending on the maximum number of PC$_2$ samples that can be formed), which are used to compute the rater component of the variance.

(e) To conduct the Monte-Carlo experiment, we have at our disposal 5,000 FC$_1$ samples and about 25,000,000 PC$_2$ samples (i.e. 25,000,000 = 5,000 × 5,000). One may see this as a 5,000 × 5,000 table where each cell contains an $n \times 2$ PC$_2$ sample of ratings that could be used to compute an agreement coefficient. For any given agreement coefficient one can always produce a 5,000 × 5,000 table where each cell contains an agreement coefficient estimate (let us refer to it as the coefficient's *MC table*). The MC table is used to produce various statistics associated with this experiment.

Computing the Simulated "Exact" PC$_2$ Variance

The simulated "exact" or Monte-Carlo (MC) PC$_2$ variance is obtained by summing the MC subject and rater variances. For a given agreement coefficient, the MC subject variance is calculated as the variance of the 5,000 MC table's row averages, while the MC rater variance is the variance of the 5,000 MC table's column averages. For $r = 3$ the MC subject variance accounted for about 68% of total variance. That percentage went down to about 54% when $r = 5$.

Computing the Expectation of the PC$_2$ Variance Estimate

The expected value of the PC$_2$ variance estimate is calculated by averaging all 25,000,000 variance estimates calculated from the 25,000,000 PC$_2$ samples using the appropriate variance equations 16, 20, or 22. To demonstrate the validity of these equations, we need to show that the expected value of the PC$_2$ variance estimates is sufficiently close to the MC variance.

Computing the Theoretical PC$_2$ Variance

Let us consider for example the theoretical variance of Fleiss' coefficient under the PC_2 design that is given by equation 14. Both components of this equation are calculated according to equation 15 using the entire population of ratings created in step (b). Based on the theoretical variance, the subject component of the variance accounted for about 68% of total variance when $r = 3$, and for about 54% when $r = 5$. This confirms what was previously observed when the MC variance was used to prove the validity of equations 14, 18, and 21.

Computing the 95% Confidence Interval Coverage Rate

The interval coverage rate represents the percentage of all 25,000,000 or so PC_2 samples that produced 95% confidence intervals containing the "true" agreement coefficient calculated using the entire population of ratings of step (b).

If follows from both Tables 6 and 7 that the exact variance, the expected variance estimate, and the theoretical variance are all very close. In fact, when the number of raters is 3 (c.f. Table 6) the AC_2 coefficient shows a maximum pairwise difference between these variances that varies from 0.16% for $n = 10$ to 0.01% for $n = 50$. For Fleiss' generalized kappa this maximum pairwise difference varies from 0.70% for $n = 10$ to 0.02% for $n = 50$, while a similar range associated with the percent agreement goes from 0.14% for $n = 10$ to 0.0% for $n = 50$. Table 7 shows similar results when the number of raters is 5. These results prove that all variance expressions presented in the previous section are accurate.

It follows from Tables 6, and 7 that the coverage rates of the 95% confidence intervals vary typically from 80% for small samples to 93% for larger samples. Although the coverage rate improves as the sample size increases, it generally remains below its nominal value of 95% particularly when the sample size is an even number. This can be partially explained by the small number of replicates used in the Monte-Carlo experiment. For example when the sample size is $n = 10$, one can form approximately $2.5 \times 10^{27} = \binom{2500}{10}$ FC_1 samples. Our experiment is limited to 5,000 FC_1 samples. Moreover, when the number of raters is 3 then each FC_1 sample of size 10 could produce 59,049 PC_2 samples (i.e. $3^{10} = 59,049$). This number increases very fast with the FC_1 sample size, and will exceed 3 billion for an FC_1 sample size of 20. Our Monte-Carlo experiment is limited to 5,000 PC_2 samples in order to reduce the execution time of our simulation program, which was written using the SAS/IML© programming language and listed in Section 4 of the Supporting Materials.

TABLE 6

TABLE 7

7 Sample Size Requirements

This section provides guidelines for determining the optimal number of subjects and raters when designing an inter-rater reliability experiment. The approach retained consists of finding the smallest sample size that yields an error margin not exceeding the prescribed level.

The error margin E associated with an agreement coefficient $\widehat{\kappa}$ is defined under a particular study design $\textsc{d}$ (i.e. $\textsc{fc}_1$ or $\textsc{pc}_2$) as a product of the critical value z_α and the standard error. That is,

$$E = z_\alpha\sqrt{V\big(\widehat{\kappa}\big|\textsc{d}\big)}. \tag{23}$$

For a confidence level as high as 90%, one gets $\alpha = (1 - 0.90)/2 = 0.05$, and the associated critical value $z_{0.05} = 1.645$ represents the 95^{th} percentile of the standard Normal distribution where $95 = 100(1 - \alpha)$.

Agreement coefficients however are complex expressions with several intertwined factors that affect the size of the error margin making it difficult to find a useful upper bound for the variance. Previous known attempts at this problem such as Donner (1999), Altaye et al. (2001*b*), or Donner and Rotondi (2010) rely on theoretical models and the assumption of a binary outcome for simplicity.

To get around this problem, we decided to proceed by experimentation using a pure nonparametric approach which stays as close as possible to what practitioners do. For n subjects, r raters, and q categories, we used the Evolutionary nonlinear algorithm, a standard mathematical optimization method implemented in MS Excel's Solver to obtain the optimal distribution of raters by subject and category. This optimal $n \times r$ table maximizes the variance of a given agreement coefficient. The experiment was repeated for $n = 10$ to $n = 100$ by increment of 5, for $r =$ 2, 3, 4, 5, 6, 7, for $q =$ 2, 3, 4, 5, 6, and 7, and for Fleiss' kappa, Gwet's, AC$_2$, and the percent agreement. The results obtained and shown in Tables 3 through 10 in Section 5 of the Supplementary Materials, were later used as input in the software CurveExpert© to express the maximum variance as a simple function of n, r, and q. That function is then used to compute the optimal number of subjects n.

7.1 Sample Size Requirements for the percent agreement p_a

Our experiment has revealed that if there are fewer raters than categories then the maximum variance of the percent agreement solely depends upon the number of subjects n, does not depend on the particular study design, and is well approximated by the equation $V_{\text{M}}\big(p_a\big) = 1/(4.0081n - 4.0532)$ where $V_{\text{M}}\big(p_a\big)$ denotes the maximum variance under the $\textsc{pc}_2$ design. For the sake of simplicity, one may instead use the equation $V_{\text{M}}\big(p_a\big) = 1/\big[4(n-1)\big]$, which will yield valid upper bounds for the variance. However, if the number of raters exceeds the number of categories, then the upper bounds of the variance depends on the study design D, the number of raters r, and the number of categories q. If $V_{\text{M}}\big(p_a\big|D\big)$ is the maximum variance of p_a under a particular design D, then it can be formulated as follows:

$$V_{\text{M}}\big(p_a\big|D\big) = \begin{cases} 1/(4.0081n - 4.0532) & \text{if } r \leq q \text{ for any design } D, \\ 1/(a_1 n + b_1) & \text{if } r > q \text{ and } D \equiv \textsc{fc}_1, \\ 1/(a_2 n + b_2) & \text{if } r > q \text{ and } D \equiv \textsc{pc}_2, \end{cases} \tag{24}$$

where a_1 and b_1 are 2 parameters related to the $\textsc{fc}_1$ design and given in table 8, whereas a_2 and b_2 are 2 parameters related to the $\textsc{pc}_2$ and given in table 9.

TABLE 8

TABLE 9

Equation 24 could be used to calculate the number of subjects required to achieve a predefined precision level. Let E be the prescribed error margin that must be achieved at a predetermined confidence level (0.90 is one of the standard values often used in the literature, 0.95 being the other). The subject sample size n required to achieve it is given by the following equation:

$$n = \frac{z_\alpha^2/E^2 - b}{a}. \tag{25}$$

Moreover, a and b will take the values a_1 and b_1 or a_2 and b_2 depending on whether the sample size is determined based the $\textsc{fc}_1$ design or the $\textsc{pc}_2$ design.

As an example, Table 10 shows the minimum number of subjects required by the magnitude of the error margin and the number of raters when the number of categories is limited to 2.

TABLE 10

It follows from Table 10 that the number of subjects required to achieve a specified error margin is often close to twice higher under the $\textsc{pc}_2$ design than under the $\textsc{fc}_1$ design, when the number of raters exceeds the number of categories.

7.2 Sample Size Requirements for the AC$_2$

Our experiment has also revealed that when the number of categories reaches 5 or exceeds it, the maximum variance of AC$_2$ becomes less dependent on the number of raters, and is primarily affected by the number of subjects. The maximum variance of AC$_2$ has been been modelled as follows:

$$V_{\text{M}}\big(\widehat{\kappa}_{2\text{G}}\big|D\big) = \begin{cases} 1/(a_1 n + b_1) & \text{if } D \equiv \textsc{fc}_1, \\ 1/(a_2 n + b_2) & \text{if } D \equiv \textsc{pc}_2, \end{cases} \tag{26}$$

where n is the number of subjects, a_1 and b_1 are the parameters associated with the $\textsc{fc}_1$ design, while a_2 and b_2 are the parameters associated with the $\textsc{pc}_2$ design. These parameters are functions of the number of raters r and number of categories q, and are shown in Table 11.

TABLE 11

Suppose that the researcher wants to determine the number of subjects n needed to achieve a specified error margin E. For given r and q values, the optimal n is calculated as follows:

$$n = \frac{z_\alpha^2/E^2 - b(r,q)}{a(r,q)}, \tag{27}$$

where $a(r,q)$ and $b(r,q)$ equal $a_1(r,q)$ and $b_1(r,q)$, or $a_2(r,q)$ and $b_2(r,q)$ of Table 11 depending on the study design. For illustration purposes, Table 12 shows the minimum number of subjects required, by the magnitude of the 90% error margin and the number of raters. This table also reveals that the AC_2 coefficient is not impacted much by the design.

TABLE 12

7.3 Sample Size Requirements for Fleiss' Generalized Kappa

Determining required sample sizes for the generalized kappa of Fleiss or Conger, for Cohen's kappa (Cohen, 1960) or Scott Pi (Scott, 1955) can be quite problematic. In fact, our experiment has revealed that when the number of raters is 2, even with 100 subjects one can find a set of ratings that produces a standard error for Scott's Pi or Cohen's kappa as high as 0.32 resulting in a 95% error margin of 0.63. Such an error margin is unduly high, and makes the agreement coefficient useless. Even with 7 raters and 100 subjects, we were still able to identify a set of ratings that led to a standard error of 0.26 for Fleiss' generalized kappa. This corresponds to a high 95% error margin of 0.51. Consequently, when using Fleiss' generalized kappa or even the basic two-rater Cohen's kappa, it is almost impossible to guarantee a reasonably small error margin at the time the study is being designed. An acceptable error margin can still be obtained in practice with these coefficients when the extent of agreement among raters is moderate or large and spread across several categories.

These facts show that no matter how carefully an inter-rater reliability study is designed, the risk of obtaining a Fleiss' generalized kappa coefficient that is unusable can never be completely eliminated.

TABLE 13

8 Concluding Remarks

The objective of this paper was to raise awareness among researchers of the importance of inter-rater reliability study design by demonstrating the extent to which it can affect how the results are interpreted. We recommended the use of standard error in conjunction with the benchmark scales found in the literature to interpret the magnitude of inter-rater reliability coefficients. Guidelines for sample size determination were provided for selected agreement coefficients.

While the fully-crossed design (FC_1) yields the lower standard error than the two-rater partially-crossed design (PC_2), it is also the most costly with respect to the total number of ratings produced for a given number of subjects. Regarding the PC_2 design, we emphasized the importance of having a systematic approach for assigning raters to subjects. Our experiment revealed that the percent agreement is independent of the design, the number of categories and the number of raters only if the number of raters is smaller than the number of categories. Otherwise, the PC_2 design will require about twice as many subjects than the FC_1 design for the same error margin. On the other hand, the AC_2 coefficient is not affected much by the particular design being used, although the number of subjects required to achieve a given error margin is slightly higher for the PC_2 design than for the FC_1 design.

Methods for determining optimal sample sizes solely based on techniques of mathematical optimization were discussed. These techniques consist of looking for the specific distribution of raters by category and subject, which maximizes the agreement coefficient variance for a given number of subjects and categories. The main advantage of this approach is to not rely on any hypothetical statistical model that may be difficult to validate. Instead, it uses Evolutionary algorithms to find the worst case scenario (i.e. the highest standard error) that could occur in practice, and to recommend the appropriate number of subjects that will keep the associated error margin below a specified threshold. In practice, the optimal number of subjects will generally yield an error margin well below its predicted maximum value. It is because a high extent of agreement among raters further reduces the standard error, and our sample size calculation method does not include a hypothesized magnitude for the agreement coefficient.

A key finding of our research is that unless special and restrictive assumptions are made, the researcher cannot prevent Fleiss'generalized kappa nor Conger's version of it from producing an unduly high error margin. For proponents of the kappa coefficient, further research may be needed to figure out what may be done at the design stage to guarantee an acceptable error margin.

References

[1] Altaye, M., A. Donner, and N. Klar (2001*b*): Inference procedures for assessing interobserver agreement among multiple raters. *Biometrics, 57,* 584588.

[2] Bennett, E. M., Alpert, R., and Goldstein, A. C. (1954). Communications through limited response questioning. *Public Opinion Quarterly, 18,* 303-308.

[3] Brennan, R. L. and Prediger, D. J. (1981). Coefficient Kappa: some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41,* 687-699.

[4] Cicchetti, D. V, and Feinstein, A. R. (1990). High Agreement but low Kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology, 43,* 551-558.

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

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

[7] Conger, A. J. (1980). Integration and Generalization of Kappas for Multiple Raters. *Psychological Bulletin, 88,* 322-328.

[8] Donner, A. (1999). Sample size requirements for interval estimation of the intraclass kappa statistic. *Communications in Statistics - Simulation and Computation, 28,* 415-429.

[9] Donner, A., and Rotondi, M. A.(2010). Sample Size Requirements for Interval Estimation of the Kappa Statistic for Interobserver Agreement Studies with a Binary Outcome and Multiple Raters. *The International Journal of Biostatistics, 6,* (Article 31).

[10] 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.

[11] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters, *Psychological Bulletin, 76,* 378-382.

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

[13] Guttman L. (1945). The test-retest reliability of qualitative data. *Psychometrika, 11,* 81-95.

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

[15] Gwet, K. L. (2008*b*). Variance estimation of nominal-scale inter-rater reliability with random selection of raters. *Psychometrika, 73,* 407-430.

[16] Gwet, K. L. (2014). *Handbook of Inter-Rater Reliability* (4th edn). Advanced Analytics, LLC.

[17] Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology, 8,* 23-34.

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

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

[20] Kraemer, H. C. (1979). Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika, 44,* 461-472.

[21] Landis, J. R., and Koch G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159-174.

[22] Light, R. J. (1971). Measures of response agreement for qualitative data: some generalizations and alternatives, *Psychological Bulletin, 76,* 365-377.

[23] Maxwell, A. E. (1977). Coefficient of agreement between observers and their interpretation. *British Journal of Psychiatry, 130,* 79-83.

[24] Rotondi, M. A., and Donner, A.(2012). A confidence interval approach to sample size estimation for interobserver agreement studies with multiple raters and outcomes. *Journal of Clinical Epidemiology, 65,* 778784.

[25] Scott, W. A. (1955). Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly, XIX,* 321-325.

[26] Wongpakaran, N., Wongpakaran, T., Wedding, D., and Gwet, K. L. (2013). A comparison of Cohens Kappa and Gwets AC_1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Medical Research Methodology, 13(1), 61*

Table 6: Monte-Carlo Simulation Results for $q = 3$ and $r = 3$.

		Sample Size n								
Coefficient	Statistic	10	15	20	25	30	35	40	45	50
AC_2	Exact Variance	0.0284	0.0187	0.0143	0.0117	0.0097	0.0085	0.0075	0.0064	0.0059
	Expected Variance Estimate	0.0288	0.0192	0.0144	0.0117	0.0098	0.0083	0.0073	0.0065	0.0058
	Theoretical Variance	0.0300	0.0200	0.0150	0.0120	0.0100	0.0085	0.0074	0.0066	0.0059
	Interval Coverage Rate	81.9%	92.1%	84.1%	91.7%	87.2%	92.6%	89.3%	93.9%	90.7%
Fleiss	Exact Variance	0.0376	0.0227	0.0165	0.0133	0.0108	0.0093	0.0081	0.0069	0.0063
	Expected Variance Estimate	0.0358	0.0225	0.0163	0.0130	0.0107	0.0090	0.0079	0.0070	0.0062
	Theoretical Variance	0.0306	0.0204	0.0153	0.0122	0.0102	0.0087	0.0076	0.0067	0.0061
	Interval Coverage Rate	80.9%	91.5%	86.9%	91.4%	90.9%	91.9%	91.5%	93.5%	91.7%
P_a	Exact Variance	0.0135	0.0088	0.0066	0.0054	0.0045	0.0039	0.0034	0.0029	0.0027
	Expected Variance Estimate	0.0121	0.0083	0.0063	0.0052	0.0043	0.0037	0.0033	0.0029	0.0026
	Theoretical Variance	0.0134	0.0089	0.0067	0.0053	0.0044	0.0038	0.0033	0.0030	0.0027
	Interval Coverage Rate	81.2%	92.1%	84.0%	91.5%	86.7%	91.5%	89.3%	93.9%	90.7%

Table 7: Monte-Carlo Simulation Results for $q = 3$ and $r = 5$.

		Sample Size n								
Coefficient	Statistic	10	15	20	25	30	35	40	45	50
AC$_2$	Exact Variance	0.0276	0.0191	0.0143	0.0115	0.0097	0.0083	0.0072	0.0065	0.0057
	Expected Variance Estimate	0.0282	0.0189	0.0144	0.0114	0.0096	0.0082	0.0072	0.0064	0.0057
	Theoretical Variance	0.0296	0.0197	0.0148	0.0118	0.0098	0.0084	0.0074	0.0065	0.0059
	Interval Coverage Rate	81.2%	91.4%	84.2%	90.9%	86.6%	91.8%	89.4%	93.2%	90.4%
[-8pt] Fleiss	Exact Variance	0.0361	0.0230	0.0165	0.0130	0.0108	0.0090	0.0078	0.0070	0.0061
	Expected Variance Estimate	0.0349	0.0222	0.0163	0.0127	0.0105	0.0089	0.0077	0.0069	0.0061
	Theoretical Variance	0.0302	0.0201	0.0151	0.0120	0.0100	0.0086	0.0075	0.0067	0.0060
	Interval Coverage Rate	80.5%	90.8%	87.1%	90.7%	92.5%	91.3%	93.5%	92.5%	92.6%
[-8pt] P_a	Exact Variance	0.0131	0.0089	0.0066	0.0053	0.0044	0.0038	0.0033	0.0030	0.0026
	Expected Variance Estimate	0.0119	0.0082	0.0063	0.0051	0.0043	0.0037	0.0032	0.0029	0.0026
	Theoretical Variance	0.0132	0.0088	0.0066	0.0053	0.0044	0.0038	0.0033	0.0029	0.0026
	Interval Coverage Rate	80.7%	91.4%	84.2%	90.8%	85.7%	91.3%	87.9%	93.2%	90.4%

Table 8: Parameters a_1 and b_1 of the maximum variance of p_a under the $\textsc{fc}_1$ design (equation 24), as a function of the number of categories q and the number of raters r

		Number of raters (r)					
q	Parameter	2	3	4	5	6	7
2	a_1	4.0081	9.0184	9.0184	11.1337	11.1337	12.2749
	b_1	-4.0532	-9.1189	-9.1189	-11.2588	-11.2588	-12.4128
3	a_1	4.0081	4.0081	5.7717	6.2627	6.2627	6.9046
	b_1	-4.0532	-4.0532	-5.8366	-6.3331	-6.3331	-6.9822
4	a_1	4.0081	4.0081	4.0081	4.9483	5.3363	5.4555
	b_1	-4.0532	-4.0532	-4.0532	-5.0039	-5.3962	-5.5168
5	a_1	4.0081	4.0081	4.0081	4.0081	4.6012	4.8963
	b_1	-4.0532	-4.0532	-4.0532	-4.0532	-4.6529	-4.9514
6	a_1	4.0081	4.0081	4.0081	4.0081	4.0081	4.4190
	b_1	-4.0532	-4.0532	-4.0532	-4.0532	-4.0532	-4.4686
7	a_1	4.0081	4.0081	4.0081	4.0081	4.0081	4.0081
	b_1	-4.0532	-4.0532	-4.0532	-4.0532	-4.0532	-4.0532

Table 9: Parameters a_2 and b_2 of the maximum variance of p_a under the PC$_2$ design (equation 24), as a function of the number of categories q and the number of raters r

		Number of raters (r)					
q	Parameter	2	3	4	5	6	7
2	a_2	4.0081	4.4434	4.4434	4.6916	4.6916	4.8234
	b_2	-4.0532	-2.0095	-2.0095	-1.7251	-1.7251	-1.6087
3	a_2	4.0081	4.0081	4.0888	4.1350	4.1350	4.2017
	b_2	-4.0532	-4.0532	-2.8568	-2.6664	-2.6664	-2.4633
4	a_2	4.0081	4.0081	4.0081	4.0276	4.0532	4.0623
	b_2	-4.0532	-4.0532	-4.0532	-3.2801	-3.0607	-3.0009
5	a_2	4.0081	4.0081	4.0081	4.0081	4.0117	4.0248
	b_2	-4.0532	-4.0532	-4.0532	-4.0532	-3.5170	-3.3128
6	a_2	4.0081	4.0081	4.0081	4.0081	4.0081	4.0069
	b_2	-4.0532	-4.0532	-4.0532	-4.0532	-4.0532	-3.6611
7	a_2	4.0081	4.0081	4.0081	4.0081	4.0081	4.0081
	b_2	-4.0532	-4.0532	-4.0532	-4.0532	-4.0532	-4.0532

Table 10: Minimum number of subjects required to achieve the prescribed 90% error margin E for the percent agreement p_a when the number of categories is 2, and the number of raters $r = 3, 5, 7$

		FC$_1$ Study Design			PC$_2$ Study Design		
E	$r < q$	$r = 3$	$r = 5$	$r = 7$	$r = 3$	$r = 5$	$r = 7$
0.01	6,752	3,002	2,431	2,206	6,090	5,768	5,611
0.03	751	334	271	246	677	641	624
0.05	271	121	98	89	244	231	225
0.07	139	62	51	46	125	118	115
0.10	69	31	25	23	61	58	56
0.13	41	19	15	14	36	34	34
0.15	31	14	12	11	28	26	25
0.17	24	11	9	9	22	20	20
0.20	18	9	7	7	16	15	14
0.25	12	6	5	5	10	10	9
0.30	9	4	4	3	7	7	7

Table 11: Parameters a_1, b_1, a_2, and b_2 asssociated with the maximum variance of the AC$_2$ agreement coefficient

		FC$_1$ Design		PC$_2$ Design	
r	q	$a_1(r,q)$	$b_1(r,q)$	$a_2(r,q)$	$b_2(r,q)$
2	2	0.7746	-0.6381	0.7746	-0.6381
3	2	1.4231	-1.5276	1.0448	-0.6650
4	2	1.7429	-1.4357	1.1045	-0.4834
5	2	1.8487	-1.7780	1.1529	-0.5404
2	3	1.3463	-1.3040	1.3419	-1.2551
3	3	1.4860	-1.3614	1.4734	-1.3363
4	3	2.0331	-1.9289	1.6377	-1.2217
5	3	2.1826	-2.3794	1.6401	-1.1497
2	4	1.8617	-1.9402	1.8547	-1.8627
3	4	1.8725	-2.0524	1.8563	-1.8809
4	4	1.9675	-1.8709	1.9595	-1.8548
5	4	2.3815	-2.2838	2.0533	-1.7041
2	5	2.2204	-2.2957	2.2141	-2.2266
3	5	2.2286	-2.3576	2.2130	-2.1896
4	5	2.2479	-2.5736	2.2275	-2.3738
5	5	2.3010	-2.2234	2.2950	-2.2046

Table 12: Minimum number of subjects required to achieve the specified 90% error margin for the AC$_2$ agreement coefficient, when the number of categories is $q = 4$

	FC$_1$ Design				PC$_2$ Design		
E	$r=2$	$r=3$	$r=4$	$r=5$	$r=3$	$r=4$	$r=5$
0.05	585	579	551	455	584	553	528
0.08	229	227	216	179	229	217	207
0.10	147	146	138	115	147	139	133
0.15	66	65	62	51	66	62	59
0.20	37	37	35	29	37	35	34
0.25	24	24	23	19	24	23	22

Table 13: Maximum variance of Fleiss' generalized kappa coefficient under the FC$_1$ and PC$_2$ designs as a function of the number of subjects n and number of raters r

		FC$_1$ design			PC$_2$ design		
n	$r = 2$	$r = 3$	$r = 4$	$r = 5$	$r = 3$	$r = 4$	$r = 5$
10	0.1431	0.1171	0.1033	0.0946	0.3091	0.2225	0.1992
15	0.1243	0.1054	0.0949	0.0885	0.2991	0.2135	0.1877
20	0.1170	0.1001	0.0909	0.0853	0.2944	0.2093	0.1824
25	0.1130	0.0971	0.0886	0.0835	0.2917	0.2068	0.1794
30	0.1104	0.0952	0.0871	0.0822	0.2899	0.2052	0.1774
35	0.1086	0.0939	0.0860	0.0813	0.2886	0.2041	0.1760
40	0.1073	0.0929	0.0852	0.0807	0.2877	0.2032	0.1749
45	0.1063	0.0921	0.0846	0.0802	0.2869	0.2026	0.1741
50	0.1055	0.0915	0.0841	0.0797	0.2864	0.2021	0.1735
55	0.1048	0.0910	0.0837	0.0794	0.2859	0.2017	0.1730
60	0.1043	0.0906	0.0834	0.0791	0.2855	0.2013	0.1726
65	0.1038	0.0903	0.0831	0.0789	0.2852	0.2010	0.1722
70	0.1034	0.0900	0.0829	0.0787	0.2849	0.2008	0.1719
75	0.1031	0.0897	0.0827	0.0785	0.2846	0.2005	0.1716
80	0.1028	0.0895	0.0825	0.0784	0.2844	0.2003	0.1714
85	0.1026	0.0893	0.0823	0.0782	0.2842	0.2002	0.1712
90	0.1023	0.0891	0.0822	0.0781	0.2841	0.2000	0.1710
95	0.1021	0.0890	0.0821	0.0780	0.2839	0.1999	0.1708
100	0.1020	0.0888	0.0819	0.0779	0.2838	0.1998	0.1707