PSYCHOMETRIKA—VOL. 73, NO. 3, 407–430
SEPTEMBER 2008
DOI: 10.1007/S11336-007-9054-8

VARIANCE ESTIMATION OF NOMINAL-SCALE INTER-RATER RELIABILITY WITH RANDOM SELECTION OF RATERS

KILEM LI GWET

SR. STATISTICAL CONSULTANT, STATAXIS CONSULTING

Most inter-rater reliability studies using nominal scales suggest the existence of two populations of inference: the population of subjects (collection of objects or persons to be rated) and that of raters. Consequently, the sampling variance of the inter-rater reliability coefficient can be seen as a result of the combined effect of the sampling of subjects and raters. However, all inter-rater reliability variance estimators proposed in the literature only account for the subject sampling variability, ignoring the extra sampling variance due to the sampling of raters, even though the latter may be the biggest of the variance components. Such variance estimators make statistical inference possible only to the subject universe. This paper proposes variance estimators that will make it possible to infer to both universes of subjects and raters. The consistency of these variance estimators is proved as well as their validity for confidence interval construction. These results are applicable only to fully crossed designs where each rater must rate each subject. A small Monte Carlo simulation study is presented to demonstrate the accuracy of large-sample approximations on reasonably small samples.

Key words: inter-rater reliability, AC_1 coefficient, kappa statistic, agreement coefficient.

1. Introduction

We consider the problem of evaluating the extent of agreement between r raters who must classify n subjects into one of q nominal response categories during a reliability experiment. Fleiss (1971) proposed one of the most popular multiple-rater agreement coefficients in use today. Fleiss's statistic is often referred to as the generalized kappa coefficient, although it generalizes the π-statistic of Scott (1955) rather than the κ-statistic of Cohen (1960). Fleiss also proposed a variance estimator of his generalized kappa under the assumption of no agreement between raters. This variance estimator, although useful for testing the null hypothesis of no agreement among raters, cannot be used for confidence interval construction, and therefore does not measure the precision of the observed agreement coefficient. To solve this problem, Gwet (2008) proposed a nonparametric variance estimator of Fleiss's coefficient, which is consistent and valid for confidence interval construction. However, Gwet's proposal estimates the variance due to the sampling of subjects only, and does not account for the variance due to the sampling of the rater universe. In this paper, I propose a variance estimation procedure that accounts for both sources of variability. In addition to Fleiss's statistic, the new variance estimation procedure will be applied to other agreement coefficients proposed in the literature. Although the problem of raters as an additional source of sampling variability is common for metric-scale agreement statistics, where it is addressed using analysis of variance models as discussed by McGraw and Wong (1996), it has not been specifically addressed for nominal-scale agreement statistics.

Inter-rater reliability is regularly used in medical and social research to evaluate the reliability of rating systems. To classify patients into various mental disease categories, psychiatrists, for example, may develop a protocol that the nurses will use routinely. The nurses in this example are often referred to as raters, judges, or observers. Their level of agreement represents

Requests for reprints should be sent to Kilem Li Gwet, Sr. Statistical Consultant, STATAXIS Consulting, 20315 Marketree Place, Montgomery Village, MD 20886, USA. E-mail: gwet62@hotmail.com

a measure of the reliability of the classification system. Early attempts to develop a measure of agreement go back to the works of Goodman and Kruskal (1954) who suggested the use of the observed proportion of agreement corrected for chance-agreement. Janson and Vegelius (1979) later indicated that the Goodman–Kruskal proposal may at times yield unpredictable results. Scott (1955), working in the context of conjoint analysis, proposed a two-rater agreement coefficient referred to as the π-statistic, which Fleiss (1971) later generalized to the case of multiple raters. Cohen (1960), criticizing Scott's statistic for its underlying assumption of marginal homogeneity, proposed the well-known kappa statistic. Other authors such as Bennett, Alpert, and Goldstein (1954), Holley and Guilford (1964), or Maxwell (1977) advocated the use of an agreement coefficient similar in its form to Scott's π-statistic and Cohen's κ-statistic with the difference that the chance-agreement probability is taken to be the inverse of the number of response categories. These authors developed the same estimator independently giving it different names.

More recently, Kraemer et al. (2002) provided an extensive overview of the kappa statistic. They discussed several versions of the intraclass kappa as well as its application in various medical research contexts. Advances have also been made in the area of confidence interval construction for the intraclass kappa coefficient. Nam (2000) proposed a likelihood score method for constructing a confidence interval of the intraclass kappa for binary data, which is more efficient than the chi-square goodness-of-fit approach of Donner and Eliasziw (1992). On the other hand, Zou and Klar (2005) proposed a noniterative procedure for obtaining a confidence interval for the multivariate intraclass kappa statistic using a modified Wald-type procedure. This provides an alternative confidence interval estimation procedure to the iterative procedure of Bartfay and Donner (2001).

In addition to Fleiss, several other authors have proposed multiple-rater agreement coefficients. Light (1971) introduced measures of agreement conditionally upon a specific category. Landis and Koch (1977), as well as Conger (1980), are other important contributions in the area of multiple-rater agreement coefficient. Important extensions of the kappa statistic to ordinal and interval data were proposed by Berry and Mielke (1988) and Janson and Olsson (2001, 2004). These new indices express the kappa statistic as a function of the Euclidean distance between pairs of multidimensional classification vectors. The use of the Euclidean distance allows for a natural generalization of kappa from nominal to ordinal and to interval data. Of particular interest is Janson and Olsson's (2004) proposal that releases the requirement of having the same group of raters rate each subject. Simon (2006) proposed an extension of the kappa statistic in the field of sequential observation data, where agreement is measured on the proper identification of the observation units to be rated as well as on their classification into categories. Simon (2006) also discussed some of kappa's paradox features. While acknowledging the limitations of kappa for analyzing clinical diagnostic data, Simon (2006) argued that the paradoxes are not problematic for analyzing observation data where all response categories have the same value (weight). Other authors, such as Schuster (2004) or Schuster and Smith (2006), addressed the inter-rater reliability problem within the framework of theoretical statistical models and have been able to gain further insight into the nature of agreement among raters.

Although Scott's π (pi) and Cohen's κ (kappa) statistics are currently used in many scientific fields, these two agreement measures have been severely criticized as they are often inconsistent with the observed level of agreement among raters. Feinstein and Cicchetti (1990) presented two troubling paradoxes associated with these measures, where low chance-corrected agreement measures are coupled with a high observed agreement. Cicchetti and Feinstein (1990) discussed alternative indexes in an effort to address these paradoxes. Zwick (1988) studied this paradox issue and recommended a two-phase approach to inter-rater agreement assessment where a marginal homogeneity test will be performed first, followed with the kappa calculation if the homogeneity of the marginals is demonstrated. Byrt, Bishop, and Carlin (1993) attempted to improve

on the kappa estimator by removing its dependency upon the trait prevalence. Their proposed PABAK estimator was shown to be equivalent to the G-index of Holley and Guilford (1964). The difficulties created by the kappa paradoxes led some authors such as Uebersax and Grove (1990, 1993) to consider alternative approaches for quantifying the rater agreement based on latent models, which unlike kappa will take rater characteristics into consideration. Another informative overview of the kappa problem was given by Cook (1998). In another attempt to address the paradox problem, an alternative and more stable (i.e., small variance) multiple-rater agreement coefficient referred to as the AC_1 statistic was proposed by Gwet (2008). Gwet (2008) also proposed a nonparametric and consistent estimator for estimating its variance and showed its validity for confidence interval construction with a Monte Carlo simulation study. Note that Brennan and Prediger (1981) also suggested interesting alternative agreement measures. Although some authors minimized the seriousness of the kappa's paradoxes as a conceptual flaw of kappa, its heavy dependency on trait prevalence remains a major problem to practitioners.

Although Fleiss (1971) proposed his multiple-rater agreement coefficient within a framework where each subject is rated by different randomly selected group of raters, its variance estimator was derived under the assumption of no agreement among raters beyond chance. That is Fleiss's proposed variance estimator is valid only if the "true" multiple-rater kappa is 0. Moreover, Fleiss did not take into consideration the variability due to the sampling of raters when deriving the variance. Consequently, Fleiss's variance estimator estimates the variance of the multiple-rater kappa due the sampling of subjects only, and when the underlying agreement coefficient is assumed to be 0.

The variance estimators that Gwet (2008) proposed for his AC_1 statistic as well as for Fleiss's kappa were also developed to estimate the variability due to the sampling of raters alone. Unlike Fleiss's variance estimator, the validity of Gwet's variance estimators is guaranteed whether the agreement coefficient is 0 or not. Consequently, the conclusions of a reliability experiment obtained with existing variance estimation methods cannot be extended beyond the group of participating raters. Resolving this limitation is the purpose of this paper.

In most practical settings, raters who are potential users of a classification system will not be selected to participate in the reliability experiment. In order to still infer to the population of subjects as well as to the whole population of raters, the variability due to the sampling of raters must be taken into account. For this problem to be resolved, a new variance estimator must be derived, its consistency for estimating the "true" variance proved, and the asymptotic normality of the rater agreement coefficient established. Asymptotic normality guarantees the validity of confidence intervals for large samples.

Section 2 presents a real-life example to show that the magnitude of inter-rater reliability is very dependent upon the specific sample of raters who participated in the reliability experiment. Section 3 describes the general framework within which the main results are obtained. Section 4 presents the main results for Gwet's AC_1 statistic as well as for Fleiss's generalized kappa. The proofs of these results are presented in the technical appendix. Section 5 illustrates the proposed variance estimators using the example of Section 2. Section 6 present a small Monte Carlo experiment to demonstrate the validity of the large-sample theory discussed in Sections 4, and 7 is devoted to concluding remarks.

2. An Example

In a study conducted at the department of Educational Psychology and Leadership studies of the University of Victoria (Canada), nine psychologists in the field of forensic mental health rated 40 computer-modified images of adolescents on the five-item Tanner physical development

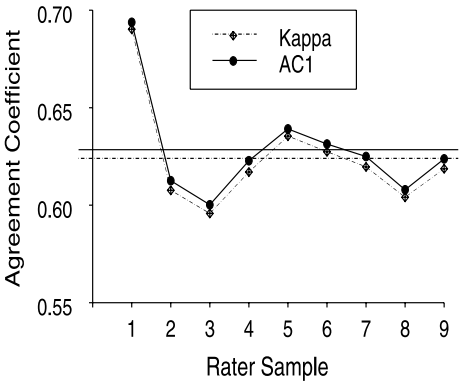

FIGURE 1.
Kappa and AC_1 coefficients for all nine rater subsamples for size 8.

TABLE 1.
Inter-rater reliability estimates for the nine subsamples of eight raters.

#	AC_1	Kappa	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8
1	0.694	0.690	1	2	3	4	5	6	7	8
2	0.613	0.608	2	3	4	5	6	7	8	9
3	0.600	0.596	1	2	3	4	5	6	7	9
4	0.623	0.617	1	2	3	4	5	6	8	9
5	0.639	0.635	1	2	3	4	5	7	8	9
6	0.631	0.627	1	2	3	4	6	7	8	9
7	0.625	0.620	1	2	3	5	6	7	8	9
8	0.608	0.604	1	2	4	5	6	7	8	9
9	0.624	0.619	1	3	4	5	6	7	8	9

scale. The Tanner physical development scale is a sexual maturity rating scale used to measure the development of secondary sex characteristics.

The primary objective of this study was to evaluate the extent of agreement between psychologists about matching the images with one of the five Tanner developmental stages. The nine participating psychologists, numbered from 1 to 9, had various levels of experience with the Tanner scales, from some familiarity to very familiar. The rating data obtained from this study is given in Table 4.

This reliability study is based on a sample of 40 images and on a sample of nine raters. Using the whole sample of subjects (images in this case) and the whole sample of raters, we evaluated inter-rater reliability using the kappa and AC_1 statistics and obtained the following estimates:

$$AC_1 = 0.63 \quad \text{and} \quad \text{kappa} = 0.62.$$

My goal is to show how sensitive the inter-rater reliability can be to the specific sample of raters that is used in this study. To this end, I have considered all nine subsamples of eight raters that can be formed from the original full sample of nine raters. For each of the subsamples, I calculated the corresponding kappa and AC_1 statistics.

It appears from the results shown in Figure 1 and in Table 1 that the absence of rater 9 alone from the sample yields a much higher inter-rater reliability (around 70%). The lowest agreement level (about 60%) is obtained with sample 3, which is obtained by removing rater 8 from the full sample. This example indicates that depending on the raters included in the rater sample, the impact on raters agreement may be as high as 10%.

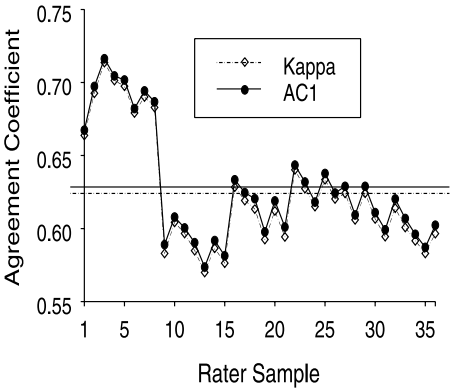

FIGURE 2.
Kappa and AC_1 coefficients for all 36 rater subsamples for size 7.

Using all 36 subsamples of size 7, which are obtained by removing two raters from the full rater sample, we calculated 36 AC_1 and kappa coefficients. The results depicted in Figure 2 indicate that depending on the seven raters used in the rater sample, inter-rater reliability can be as low as 59% and as high as 72%. It is unlikely that this real-life example represents the worst possible scenario that may occur in practice.

I deliberately used the University of Victoria study in this example to illustrate the sensitivity of AC_1 and kappa statistics to the sampling of raters, even though the Tanner scale is ordinal and the use of other methods is often indicated for assessing the extent of agreement among raters. This is not problematic since the ordinal nature of the Tanner scale does not affect the (rater) sampling distributions of the agreement coefficients under investigation.

Consequently, different sample of raters of the same size may yield very different agreement coefficients among raters. This is due to the fact that the observed magnitude of inter-rater reliability does not simply represent raters ability to use a given classification system, but also represents the characteristics of the specific rater and subject samples retained for the study. Some authors have recommended a careful analysis of heterogeneity within the rater sample, and have questioned the use of an overall measure of agreement if the rater is not homogeneous. In fact, an heterogeneous rater sample, like any other heterogeneous sample, will increase the variance of the statistics without affecting their validity. However, this issue should be addressed by evaluating precisely the variance due to the sampling of raters as proposed in this paper.

3. The Framework

Although inter-rater reliability is commonly treated in terms of variance components models using the generalizability theory where the model parameters are the targets for inference, I have adopted the randomization approach for inference in this paper, which is based on the principles of finite-population sampling. The parameter that is subject to inference is the finite-population value of the agreement coefficient γ. Cochran (1977) provided a good account of the finite-population sampling theory. Although this approach to statistical inference differs from the classical approach often used in the inter-rater reliability literature, the results obtained are comparable if the finite populations are assumed to be infinitely large.

Throughout this paper, we assume that at the time the reliability experiment is designed, the researcher knows (at least approximately) the number N of subjects and R of raters that are targeted and has a way to contact them eventually for their participation to the study. Due to

practical and cost considerations, not all N subjects nor all R raters are expected to participate in the study. They merely define the target populations (or populations of inference). In some instances, these population sizes may be unknown and would be approximated using appropriate methods that are beyond the scope of this paper, or alternatively would be assumed infinitely large. This last assumption will generally lead to more familiar results.

It is also assumed that of the N population subjects, the researcher decides to select n for participation to the study, while including r of the R population raters. The reliability experiment will then be implemented with a sample of r raters who must each classify every one of the n subjects into one of q possible response categories. That is, the reliability experiment is based on a fully crossed design. Let $f_n = n/N$ and $f_r = r/R$ denote, respectively, the subject and rater sampling fractions (i.e., the proportion of the population that the sample size represents). Let r_{ik} be the number of sample raters who classified sample subject i into category k following the experiment. If all R population raters had participated in the experiment, an unknown number R_{ik} of them would have classified subject i into category k. However, R_{ik}, which is the quantity the researcher is interested in, is a population parameter that cannot be observed, therefore must be approximated from the data.

Let us define some concepts and introduce some notations that will be used regularly in Section 4. A rater randomly selected from the rater population will classify a randomly selected subject into category k with a probability denoted by π_k. When estimated from the samples of raters and subjects, this probability will be denoted by $\widehat{\pi}_k$. The probability π_k and its estimator $\widehat{\pi}_k$ are defined as follows:

$$\pi_k = \frac{1}{N}\sum_{i=1}^{N}\frac{R_{ik}}{R}, \quad \text{and} \quad \widehat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n}\frac{r_{ik}}{r}.$$

Obtaining rater agreement coefficients often involve the computation of the probability that two randomly selected raters classify a randomly selected subject into the same category. When calculated from the subject and rater populations, this probability will be denoted by P_a and its sample-based estimator by p_a. Then,

$$P_a = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{q}\frac{R_{ik}(R_{ik}-1)}{R(R-1)}, \quad \text{and} \quad p_a = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{q}\frac{r_{ik}(r_{ik}-1)}{r(r-1)}.$$

Two raters who agree about the classification of a subject, may reach that agreement by chance without actually having the same look at the subject. It is widely accepted in the inter-rater reliability literature that the impact of agreement by chance must be minimized when evaluating the extent of agreement among raters. Therefore, the probability of occurrence of this event is essential for obtaining a precise assessment of raters' agreement level. Practitioners extensively use the method proposed by Fleiss (1971) for computing the chance-agreement probability. However, Gwet (2008) proposed an alternative approach for obtaining the same probability that was shown to have better statistical properties. Using population data, Fleiss's and Gwet's chance-agreement probabilities, denoted respectively as $P_{e\pi}$ and $P_{e\gamma}$, are given by

$$P_{e\pi} = \sum_{k=1}^{q}\pi_k^2, \quad \text{and} \quad P_{e\gamma} = \frac{1}{q-1}\sum_{k=1}^{q}\pi_k(1-\pi_k).$$

These two parameters can be estimated from the samples using the following estimators:

$$p_{e\pi} = \sum_{k=1}^{q}\widehat{\pi}_k^2, \quad \text{and} \quad p_{e\gamma} = \frac{1}{q-1}\sum_{k=1}^{q}\widehat{\pi}_k\big(1-\widehat{\pi}_k\big).$$

The generalized kappa agreement statistic $\widehat{\gamma}_\pi$ suggested by Fleiss (1971) for estimating inter-rater reliability is given by $\widehat{\gamma}_\pi = (p_a - p_{e\pi})/(1 - p_{e\pi})$, while the AC_1 estimator of Gwet (2008) is defined as $\widehat{\gamma}_1 = (p_a - p_{e\gamma})/(1 - p_{e\gamma})$. Gwet (2008) indicated that, conditionally upon the sample of raters (i.e., no variance is expected from the sampling of raters), a consistent estimator of the variance of $\widehat{\gamma}_1$ is given by

$$v_G(\widehat{\gamma}_1) = \frac{1 - f_n}{n} \frac{1}{n-1} \sum_{i=1}^{n} \left(\widehat{\gamma}^*_{1|i} - \widehat{\gamma}_1\right)^2, \tag{1}$$

where $\widehat{\gamma}^*_{1|i} = \widehat{\gamma}_{1|i} - 2(1 - \gamma_1)(p_{e\gamma|i} - p_{e\gamma})/(1 - p_{e\gamma})$, $\widehat{\gamma}_{1|i} = (p_{a|i} - p_{e\gamma})/(1 - p_{e\gamma})$, $p_{a|i}$ and $p_{e\gamma|i}$ being defined as

$$p_{a|i} = \sum_{k=1}^{q} \frac{r_{ik}(r_{ik} - 1)}{r(r-1)} \quad \text{and} \quad p_{e\gamma|i} = \frac{1}{q-1} \sum_{k=1}^{q} \widehat{\pi}_k \left(1 - \frac{r_{ik}}{r}\right). \tag{2}$$

Note that a variance estimator is said to be consistent if it tends to estimate the correct variance for large sample sizes. That is, its bias and standard error go to 0 as the number of sample subjects increases. Similarly, conditionally upon the sample of raters, a consistent estimator of the variance of $\widehat{\gamma}_\pi$ is given by

$$v_G(\widehat{\gamma}_\pi) = \frac{1 - f_n}{n} \frac{1}{n-1} \sum_{i=1}^{n} \left(\widehat{\gamma}^*_{\pi|i} - \widehat{\gamma}_\pi\right)^2, \tag{3}$$

where $\widehat{\gamma}^*_{\pi|i} = \widehat{\gamma}_{\pi|i} - 2(1 - \widehat{\gamma}_\pi)(p_{e\pi|i} - p_{e\pi})/(1 - p_{e\pi})$, $\widehat{\gamma}_{\pi|i} = (p_{a|i} - p_{e\pi})/(1 - p_{e\pi})$, $p_{e\gamma|i}$ being defined as

$$p_{e\pi|i} = \sum_{k=1}^{q} \widehat{\pi}_k \frac{r_{ik}}{r}. \tag{4}$$

Note that the variance estimator Fleiss (1971) proposed for $\widehat{\gamma}_\pi$, when adapted to the finite-population context, is given by

$$v_F(\widehat{\gamma}_\pi) = \frac{1 - f_n}{n} \frac{1}{r(r-1)} \frac{p_{e\pi} - (2r-3)p_{e\pi}^2 + 2(r-2)\sum_{k=1}^{q} \widehat{\pi}_k^3}{(1 - p_{e\pi})^2},$$

the final population correction factor being the only difference between this formula and Fleiss's. The variance estimators of (1) and (3) work well for estimating the variances of the AC_1 and π estimators conditionally upon the rater sample, as indicated in the Monte Carlo simulation presented by Gwet (2008). Fleiss's variance estimator $v_F(\widehat{\gamma}_\pi)$, on the other hand, was derived under the assumptions that there was no agreement among raters beyond chance, that each subject is rated by a different randomly selected group of raters, and that these ratings follow a multinomial distribution with probabilities $\pi_1, \ldots, \pi_q$. Fleiss's estimator also ignores the variability due the sampling of raters like Gwet's estimator. It will primarily be useful for testing the hypothesis that there is no agreement among raters, and should not be used to quantify the precision of the π statistic nor to construct confidence intervals. Its validity also relies heavily upon the multinomial distribution assumed for the ratings. Gwet's estimator can be used either for hypothesis testing, for quantifying the precision of the π statistic, or for constructing confidence intervals, and is not based on any particular assumption. Therefore, the two variance estimators will be equal only under the assumption that there is no agreement among raters beyond chance, and if there are three raters or more involved it may be necessary to further assume that the rating of subjects are independent. That is, $v_G(\widehat{\gamma}_\pi)$ and $v_F(\widehat{\gamma}_\pi)$ will be equal only if the conditions that ensure the

validity of Fleiss's variance estimator are satisfied. Because the variance estimators $v_G(\widehat{\gamma}_1)$ and $v_G(\widehat{\gamma}_\pi)$ both consider the sample of raters to be fixed and not subject to sampling variability, the resulting significance tests or confidence intervals will be valid only for the specific group of raters who participated in the reliability experiment. However, it is a fact that researchers have a natural tendency to generalize their findings to bigger groups of raters. Consequently there is a need to have a variance estimator that will simultaneously account for both the subject and the rater sampling variability. We will not consider generalizing Fleiss's variance estimator to account for the sampling of raters because of its limited applicability.

Let us consider, for example, a reliability study that led to an inter-rater reliability coefficient of $\widehat{\gamma} = 70\%$. Can a researcher claim that the extent of agreement among raters is greater than 60%? The correct answer to this question depends on which group of raters is being targeted for inference. If the researcher targets only the group of raters that participated in the experiment, then one may conclude that 70% is significantly greater than 60% if $0.60 > 0.7–1.96\sqrt{v(\widehat{\gamma})}$, where $v(\widehat{\gamma})$ is given either by (1) or by (3) depending on the statistic being used. This statement assumes that, conditionally on the rater sample, the asymptotic distribution of $\widehat{\gamma}$ is normal. Note that one of the objectives of this paper is to establish the more general result of asymptotic normality of the unconditional distribution of $\widehat{\gamma}$.

If the researcher likes to infer to a group of raters larger than the one that participated in the reliability experiment, then neither (1) nor (3) provide the correct method for evaluating the variance of the inter-rater reliability coefficient. As shown in the Monte Carlo simulation study presented in Section 6, the estimators in (1) and (3) underestimate the correct variance when the participating raters only represent a sample of a bigger universe of potential raters.

The exact variance of the inter-rater reliability coefficient that accounts for the two sources of variability due to the sampling of raters and subjects is difficult to obtain. Instead, we will derive a large sample approximation and will propose a consistent variance estimator (i.e., a variance estimator that converges to the "true" unconditional variance as the subject and rater sample sizes increase). Asymptotic normality for the AC_1 and π statistics will be proved as well, showing thereby the validity of confidence intervals based on the proposed variance estimators.

4. Large-Sample Results

This section presents the proposed unconditional variance estimators for the AC_1, and Fleiss's generalized kappa statistics, as well as Theorem 1, which establishes their asymptotic normality, the main result of this paper. Three lemmas containing intermediary results needed to prove Theorem 1 are also presented. Throughout this section, the probability distributions of the agreement coefficients are defined with respect to both the sampling of subjects as well as the sampling of raters.

In order to avoid redundancy, we will use a general expression for the agreement coefficient that encompasses several known agreement statistics, including the AC_1 coefficient, the generalized kappa statistic, and possibly other agreement coefficients. Let $\widehat{\gamma}$ be the inter-rater reliability statistic whose large-sample properties are being investigated. We assume that $\widehat{\gamma} = (p_a - p_e)/(1 - p_e)$, where p_a is defined as in Section 3 and the chance-agreement probability p_e defined as

$$p_e = \sum_{k=1}^{q} \widehat{\pi}_k f(\widehat{\pi}_k), \quad \text{where } f(x) = ax + b, \tag{5}$$

$f(\cdot)$ being a real-valued function defined in the interval $(0, 1)$, a, and b two real numbers. The chance-agreement probabilities $p_{e\gamma}$ and $p_{e\pi}$ used, respectively, for the AC_1 and Fleiss's kappa are special cases of p_e. For $f(x) = x$ (i.e., $a = 1$, and $b = 0$), one obtains Fleiss's generalized

kappa, while $f(x) = (1-x)/(q-1)$ (i.e., $a = -1/(q-1)$, and $b = 1/(q-1)$) yields Gwet's AC_1 coefficient. For the remaining part of this paper, we will often use p_e or $\widehat{\gamma}$ to refer to both agreement coefficients studied in this paper and f' is the first-order derivative of function f. The population-level chance-agreement probability P_e is expressed as in (5) with $\widehat{\pi}_k$ being replaced by π_k.

Let $p_{ik} = r_{ik}/r$ and $P_{ik} = R_{ik}/R$ be, respectively, the sample and population proportions of raters to classify subject i into category k. In order to derive the unconditional variance of the agreement coefficient $\widehat{\gamma}$, which accounts for both the sampling of subjects and that of raters, I approximated $\widehat{\gamma}$ by a linear function of the p_{ik}'s. The validity of this approximation for large rater and subject samples is established in Lemma 1.

Let us introduce the following notations for a sample rater α, and a sample subject i:

$\Rightarrow \widehat{\pi}_k^{(\alpha)}$: proportion of sample subjects that rater α classified into category k,

$\Rightarrow \alpha_i$: response category into which rater α classified subject i,

$$\Rightarrow p_a^{(\alpha)} = \frac{1}{n}\sum_{i=1}^{n} p_{i\alpha_i}, \quad p_e^{(\alpha)} = \big(1-\widehat{\gamma}\big)\sum_{k=1}^{q}\widehat{\pi}_k f\big(\widehat{\pi}_k^{(\alpha)}\big), \quad \widehat{\gamma}_{(\alpha)} = \frac{p_a^{(\alpha)} - p_e^{(\alpha)}}{1-p_e},$$

$$p_{a|i} = \sum_{k=1}^{q}\frac{r_{ik}(r_{ik}-1)}{r(r-1)}, \quad \widehat{\gamma}_i = \frac{p_{a|i} - p_e}{1-p_e}, \quad \widehat{\gamma}_i{}^* = \widehat{\gamma}_i - 2a\frac{1-\widehat{\gamma}}{1-p_e}\sum_{k=1}^{q}\widehat{\pi}_k\big(p_{ik} - \widehat{\pi}\big).$$

Note that $p_a^{(\alpha)}$ can be seen as the probability that a randomly chosen rater and rater α agree (i.e., classify a randomly chosen subject into the same category). On the other hand, $p_e^{(\alpha)}$ is the probability that a randomly chosen rater and rater α agree by chance, while $\widehat{\gamma}_{(\alpha)}$ estimates the extent of agreement between rater α and the other raters in the rater universe.

The estimator that we propose for computing the variance of the agreement coefficient $\widehat{\gamma}$ is given by

$$v(\widehat{\gamma}) = v_s(\widehat{\gamma}) + v_r(\widehat{\gamma}), \quad \text{where,} \quad \begin{cases} v_s(\widehat{\gamma}) = \frac{1-f_n}{n}s_\gamma^{*2}, \text{ and } s_\gamma^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(\widehat{\gamma}_i^* - \widehat{\gamma})^2, \\ v_r(\widehat{\gamma}) = 4\frac{1-f_r}{r}\frac{1}{r}\sum_{\alpha=1}^{r}(\widehat{\gamma}_{(\alpha)} - \widehat{\overline{\gamma}}_{(\bullet)})^2. \end{cases} \tag{6}$$

The variance estimator $v(\widehat{\gamma})$ of $\widehat{\gamma}$ given by (6) has two components. The first component $v_s(\widehat{\gamma})$ is the variance due to the sampling of subjects, and the second component $v_r(\widehat{\gamma})$ estimates the variance due to the sampling of raters. It appears that the rater component of the variance will be small if the agreement coefficient γ is high or the number of raters in the sample is high. If the rater sample is small or the agreement coefficient is low (below 70% say), then it is strongly recommended to compute the rater component of the variance. The Monte Carlo simulation of Section 6 further discusses the importance of this variance component.

It should be noted that unconditional variance estimators $v(\widehat{\gamma}_1)$ and $v(\widehat{\gamma}_\pi)$ for the AC_1 and generalized kappa statistics are obtained directly from (6) using the appropriate f function. Both variance estimators are included in the Monte Carlo experiment presented in Section 6.

The next theorem contains the main result of this paper regarding the asymptotic normality of the AC_1 and the generalized kappa estimators when both the subject and rater sample sizes are large. Let f_r and f_n be, respectively, the rates at which the rater and subject populations are sampled. Throughout this section, the following three conditions are assumed to be satisfied:

(a) $\lim_{r\to\infty} f_r = \nu_0$, for $0 < \nu_0 < 1$,
(b) $p_e = \theta_e + o_p(1)$, for $0 \le \theta_e < 1$, and
(c) $\lim_{n\to\infty} f_n = f_0$, for $0 < f_0 < 1$.

Condition (a) indicates that the rater universe is sampled at a positive rate excluding the case where all population raters are sampled. Condition (b) guarantees that chance-agreement probability is smaller than 1, at least for large samples, which ensures the existence of the inter-rater reliability coefficient. Condition (c) states that the subject population is sampled at a positive rate and excludes the case where all population subjects are sampled with certainty.

Theorem 1. *If conditions* (a), (b) (*for* $\widehat{\gamma}_\pi$ *only*), *and* (c) *are satisfied, then as* $r \to \infty$ *and* $n \to \infty$ *the statistic* $\widehat{\gamma}$ *converges weakly to the normal distribution. That is, for* γ *given by* $\gamma = (P_a - P_e)/(1 - P_e)$, *we have that*

$$\frac{\widehat{\gamma} - \gamma}{\sqrt{v(\widehat{\gamma})}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

If the rater and subject samples are "sufficiently large", Theorem 1 guarantees that the proposed variance estimator can be used to construct symmetric normality-based confidence intervals for γ with a coverage rate that is close to its nominal value. This result is confirmed by the simulation study of Section 6. However, for small to moderate samples, the agreement coefficient $\widehat{\gamma}$ will have a skewed distribution unless the "true" agreement coefficient γ is 0.5. In this case, alternative approaches such the bootstrap confidence intervals may be used. This skewness will disappear as the rater and subject sample sizes increase. However, the Monte Carlo study results reported in Table 5 indicate that for an agreement coefficient around 63%, valid confidence intervals can be obtained for rater sample sizes as small as 5, and subject sample size as small as 20. I would even expect subject sample sizes around 15 to still yield valid confidence intervals, although the precision of the of statistics may not be very good.

To prove Theorem 1, I will first approximate the agreement coefficient $\widehat{\gamma}$ by a linear function of the $\boldsymbol{p}_i$'s where $\boldsymbol{p}_i^\top = (p_{i1}, \ldots, p_{iq})$. This linear approximation will be in the form $\widehat{\gamma} = \gamma + \overline{\lambda} + \textit{Remainder}$, where $\overline{\lambda}$ is the sample mean of the λ_i's, $\lambda_i = \boldsymbol{\Phi}_i^\top(\boldsymbol{p}_i - \boldsymbol{P}_i)$ (for some vectors $(\boldsymbol{\Phi}_i)_{i\geq 1}$), and the "remainder" being a random variable, which can be neglected based on its stochastic order of magnitude. It follows from the multivariate central limit theorem (see Rao, 2002) that the ratio $\overline{\lambda}/\sqrt{V(\overline{\lambda})}$ converges in distribution to $\mathcal{N}(0, 1)$. The proof will be completed by demonstrating that the remainder converges in probability to 0, that the ratio $v(\widehat{\gamma})/V(\overline{\lambda})$ converges in probability to 1, and by evoking Slutsky's theorem (see Rao, 2002). Slutsky's theorem stipulates that if two random sequences X_n and Y_n converge in law to X, and in probability to a constant a, respectively, then the sequence $g(X_n, Y_n)$ converges in law to $g(X, a)$ for any continuous function g. Only a few lemmas needed to prove Theorem 1 are presented in the main text, the complete proofs can be found in Appendix.

Let us introduce the following notations:

- α_i: Category into which rater α classified subject i.
- $P_{i\alpha_i}$: Probability for a randomly chosen rater to classify subject i into category α_i (i.e., the category where rater α classified subject i).
- $\pi_k^{(\alpha)}$: Probability that rater α classify a randomly chosen subject into category k.
- $P_a^{(\alpha)}$ and $P_e^{(\alpha)}$, respectively, represent the probability that a randomly chosen rater and rater α agree, and the probability that the two raters agree by chance. The resulting agreement coefficient is $\gamma_{(\alpha)} = (P_a^{(\alpha)} - P_e^{(\alpha)})/(1 - P_e)$.
- $P_{a|i}$: Probability that two randomly selected raters classify subject i into the same category.
- $P_a^{(\alpha)} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} P_{i\alpha_i}, \quad P_e^{(\alpha)} = (1-\gamma)\sum_{k=1}^{q}\pi_k f\big(\pi_k^{(\alpha)}\big), \quad P_{a|i} = \sum_{k=1}^{q}\frac{R_{ik}(R_{ik}-1)}{R(R-1)},$

- $\gamma_i = (P_{a|i} - P_e)/(1 - P_e)$, $\gamma_i^* = \gamma_i - 2a\dfrac{1-\gamma}{1-P_e}\displaystyle\sum_{k=1}^{q}\pi_k(P_{ik} - \pi_k)$, $\pi_k^{(\alpha)} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\eta_{ik}^{(\alpha)}$, where $\eta_{ik}^{(\alpha)} = 1$, if rater α classifies subject i into category k, and $\eta_{ik}^{(\alpha)} = 0$ otherwise.

Lemma 1. *Suppose that conditions* (a), (b), *and* (c) *are satisfied. Then, as* $n, r \to \infty$, *we have*

$$\widehat{\gamma} = \gamma + \frac{1}{n}\sum_{i=1}^{n}\lambda_i + O_p(1/n) + O_p(1/r),$$

$$\lambda_i = (\gamma_i^* - \gamma) + \sum_{k=1}^{q}\Phi_{ik}(p_{ik} - P_{ik}), \quad \textit{and} \quad \Phi_{ik} = \frac{2[P_{ik} - a(1-\gamma)\pi_k]}{1 - P_e}.$$

Lemma 1 stipulates that the agreement coefficient $\widehat{\gamma}$ can be expressed as the summation of a linear function of the vectors $\boldsymbol{p}_i$'s and a remainder made up of two random variables, one of which has the same stochastic magnitude order as $1/n$ (i.e. is $O_p(1/n)$), and the other has the same stochastic magnitude order as $1/r$ (i.e. is $O_p(1/r)$). Section A.2 of Appendix gives a more detailed account of the O_p notations.

Lemma 2 gives an expression of the "true" unconditional variance of $\overline{\lambda}$.

Lemma 2. *Suppose that conditions* (a), (b), *and* (c) *are satisfied. Then,*

$$V(\overline{\lambda}) = \frac{1 - f_n}{n}S_\gamma^{*2} + \frac{4(1 - f_r)}{r}\frac{1}{R}\sum_{\alpha=1}^{R}\big(\gamma_{(\alpha)} - \overline{\gamma}_{(\bullet)}\big)^2 + A + B,$$

where S_γ^{*2}, A, *and* B *are given by*

$$S_\gamma^{*2} = \frac{1}{N-1}\sum_{i=1}^{N}\big(\gamma_i^* - \gamma\big)^2, \quad B = \frac{(1-f_r)}{r(R-1)}\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\Psi_{ij},$$

$$A = \frac{(1-f_n)(1-f_r)R}{nr(R-1)}\left(\frac{1}{N}\sum_{i=1}^{N}\Psi_{ii} - \frac{1}{N(N-1)}\sum_{i\neq j}^{N}\sum^{N}\Psi_{ij}\right),$$

$$\Psi_{ij} = \frac{1}{R}\sum_{\alpha=1}^{R}\big(\Phi_{i\alpha_i} - \overline{\Phi}_{i(\bullet)}\big)\big(\Phi_{j\alpha_j} - \overline{\Phi}_{j(\bullet)}\big), \quad \textit{and} \quad \overline{\Phi}_{i(\bullet)} = \frac{1}{R}\sum_{\alpha=1}^{R}\Phi_{i\alpha_i}.$$

Lemma 3, which is stated without proof, says that the two random variables $(\widehat{\gamma} - \gamma)/\sqrt{v(\widehat{\gamma})}$ and $\overline{\lambda}/\sqrt{V(\overline{\lambda})}$ differ by a random variable that converges in probability to 0. Therefore, the two random variables have the same asymptotic distribution.

Lemma 3. *If conditions* (a), (b), *and* (c) *are satisfied, then*

$$\frac{\widehat{\gamma} - \gamma}{\sqrt{v(\widehat{\gamma})}} = \frac{\overline{\lambda}}{\sqrt{V(\overline{\lambda})}} + o_p(1).$$

It is also possible to demonstrate that the proposed variance estimator of $\widehat{\gamma}$ is good for estimating the "true" variance of $\widehat{\gamma}$, in the sense that the ratio $v(\widehat{\gamma})/V(\widehat{\gamma})$ converges in probability to 1.

TABLE 2.
AC_1, kappa statistics, and associated precision measures.

Statistics	AC_1	Kappa
Agreement coefficient	62.9%	62.4%
S.E.[a] due to the sampling of subjects ($\sqrt{v_s(\widehat{\gamma})}$)	4.6%	4.5%
S.E. due to the sampling of raters ($\sqrt{v_r(\widehat{\gamma})}$)	5.5%	5.5%
Unconditional standard error ($\sqrt{v(\widehat{\gamma})}$)	7.3%	7.3%
95% C.I.[b] conditionally upon the rater sample	(53.6%; 72.1%)	(53.4%; 71.4%)
Unconditional 95% confidence interval	(48.2%; 77.5%)	(47.8%; 77.0%)

[a]Standard error.

[b]Confidence interval.

5. Application

In this section, we apply the variance estimation techniques discussed in Section 4 to the experimental data on sexual maturity of Section 2. Because the researchers in this study did not report the size of the target subject population nor that of the rater population, I considered them to be reasonably large by assuming $N = 1{,}000$ for the subject population and $R = 100$ for the rater population. The subject and rater sample sizes are given by $n = 40$ and $r = 9$, respectively, as seen in Section 2.

Table 2 shows various precision measures calculated for the AC_1 and kappa coefficients. The AC_1 and kappa agreement coefficients are 62.9% and 62.4%, respectively. If only the variance due to the sampling of subjects is taken into account as is often done, the standard error of the agreement coefficient will be around 5%. This leads to a substantial underestimation of the overall sampling variability associated with the agreement coefficient. In fact, Table 2 also indicates that the standard error due to the sampling of raters is about 6%, and should be taken into consideration. This yields an overall unconditional standard error for the agreement coefficient of about 7%.

Overlooking the rater component of the variance will lead to confidence intervals that are narrower than they should normally be. As Table 2 suggests, the 95% confidence interval of either agreement coefficient, when based on the sampling of subjects alone (i.e. conditionally upon the current rater sample) is about $(53, 72)$, while the unconditional confidence interval given by $(48, 78)$ is much wider. Based on the conditional confidence interval, one would conclude that the difference between the observed raters agreement and 50% is statistically significant. However, if we use the unconditional confidence interval, this conclusion does not hold anymore.

The next section presents a Monte Carlo simulation aimed at demonstrating the validity of the unconditional confidence interval. We demonstrate the fact that, by sampling simultaneously both the universe of raters and that of subjects, the confidence intervals obtained will contain the "true" agreement coefficient at a coverage rate that is close to its nominal value.

6. Monte Carlo Simulation

The Monte Carlo experiment presented in this section has two objectives: (i) to validate the normality of the AC_1 and the generalized kappa statistics sampling distribution as stated in Theorem 1 for large subject and rater samples, and (ii) to verify that the proposed variance estimators approximate the "true" variance reasonably well. The validation of normality is carried out by comparing the coverage rate of the normality-based confidence interval to its nominal value.

A small difference between the two values would validate Theorem 1. Although validating normality will also validate the proposed variance estimators, the latter will further be explored by computing the magnitude of their biases. The biases of the AC_1 and kappa statistics will also be computed to evaluate their propensity for estimating the correct agreement coefficient.

The bias of an estimator is the difference between its Monte Carlo expectation and the population parameter that is being estimated. The bias of a variance estimator on the other hand, is the difference between its Monte Carlo expectation and the Monte Carlo variance of the agreement coefficient. A small bias is desirable as it indicates that a given estimator or variance estimator has neither a tendency to overestimate the true population parameter nor a tendency to underestimate it. Normality is verified by comparing the nominal coverage rate of confidence intervals (usually 95%) to the Monte Carlo relative to the number of times that the constructed confidence interval contains the true population inter-rater reliability.

To conduct this experiment, I generated a population of $N = 100$ subjects (labeled as $i = 1, \ldots, 100$), and a population of $R = 20$ raters (labeled as $l = 1, \ldots, 20$). One of 5 (i.e., $q = 5$) possible response categories (labeled as $k = 1, \ldots, 5$) was then associated with each subject in such a way that category 1 is assigned about 50% of all population subjects, while each of the remaining four categories is assigned about 12.5% of the population subjects. My primary objective was to create a subject population with high prevalence rate of category 1. The specific percentages of 50% and 12.5% were selected arbitrarily.

The actual number of subjects per category is known only after these subjects are assigned, as this is done randomly. The categories so obtained represent the "correct" classification of subjects or the "gold standard". To complete the construction of population data, I accomplished the following tasks:

- Fifty randomly chosen population subjects were assigned to category 1, while the remaining 50 population subjects were randomly assigned (with the same probability) to one of the four remaining categories.
- I assumed that 80% of the time a rater will classify the subject being rated into the "correct" category. That is, 20% of the time the rating is done randomly, leading to a classification that may or may not be correct. This situation will occur when the rater, not knowing the correct answer, decides to take a leap of faith by assigning a subject to a randomly chosen category. For any subject, the number of correct nonrandom ratings is determined according to the binomial distribution $\mathcal{B}(20, 0.8)$. Consequently, each time a subject is rated, 16 raters on average are expected to perform a correct nonrandom rating.
- After determining their numbers, the specific raters who have performed the correct nonrandom ratings were also determined randomly. After this step, I had a complete population dataset with 100 subjects, 20 raters and their ratings, as well as labels indicating which of the ratings were random. These population data, which are usually not available to the researcher, tell us how each potential rater is expected to classify each subject.

After selecting a sample of subjects and a sample of raters, the rating data needed to compute the agreement coefficient estimates and associated standard errors will come from ratings generated at the population level.

At the population level, the AC_1 and generalized kappa coefficients were evaluated at $\gamma_1 = 0.62$ and $\gamma_\pi = 0.53$, respectively. The population data generated for this study, as well as the Monte Carlo simulation programs written in the SAS IML language, can be obtained from the author.

The simulation consists of selecting successively 10,000 replicate subject samples and 10,000 replicate rater samples of various sizes. The simulation was carried out for the subject sample size values of $n = 20$, 30, 40, 50, and for the rater sample size values of $r = 5$, 7, 9, 11, and 13. When the rater and subject samples are small, the chance-agreement probability $p_{e|\pi}$

TABLE 3.
Relative biases of agreement coefficients, their Monte Carlo standard errors, and efficiency of $\widehat{\gamma}_1$ to $\widehat{\gamma}_\pi$.

n	r	$\sqrt{V_{MC}(\widehat{\gamma}_1)}$ (%)	$\sqrt{V_{MC}(\widehat{\gamma}_\pi)}$ (%)	$e(\widehat{\gamma}_1 \mid \widehat{\gamma}_\pi)$	$RB(\widehat{\gamma}_1)$ (%)	$RB(\widehat{\gamma}_\pi)$ (%)	$V_r / V_{\mathcal{T}}(\widehat{\gamma}_1)$ (%)	$V_r / V_{\mathcal{T}}(\widehat{\gamma}_\pi)$ (%)
	20	7.6	9.1	1.4	0.15	−2.7	23.7	17.8
5	30	5.9	7.0	1.4	0.07	−1.5	26.4	20.0
	40	4.9	5.6	1.3	−0.05	−1.0	29.2	22.3
	50	4.1	4.7	1.3	0.03	−0.6	33.1	25.4
	20	6.5	8.1	1.5	0.19	−2.6	28.4	20.4
7	30	5.1	6.2	1.5	0.25	−1.2	31.6	23.2
	40	4.2	5.0	1.4	0.17	−0.7	34.9	26.0
	50	3.5	4.2	1.4	0.13	−0.4	38.7	29.2
	20	5.7	7.3	1.6	0.20	−2.5	29.5	20.0
9	30	4.5	5.6	1.6	0.06	−1.5	32.7	22.7
	40	3.7	4.5	1.5	0.08	−0.9	35.8	25.3
	50	3.1	3.8	1.5	0.12	−0.5	40.0	28.9
	20	5.1	6.7	1.8	0.30	−2.3	28.5	17.8
11	30	3.9	5.1	1.7	0.15	−1.3	31.3	20.2
	40	3.2	4.2	1.6	0.07	−0.8	34.4	22.9
	50	2.7	3.5	1.6	0.03	−0.5	38.8	26.3
	20	4.5	6.3	1.9	0.26	−2.3	25.6	14.8
13	30	3.5	4.8	1.9	0.15	−1.4	28.1	16.7
	40	2.9	3.8	1.8	0.10	−0.8	31.3	19.1
	50	2.4	3.2	1.7	0.05	−0.6	35.1	21.9

may take a value of 1 leading to a division by 0 in the calculation of the π-statistic. To avoid this problem in the simulation programs, the calculation of the π-statistic was modified slightly in such a way that if $p_{e|\pi} = 1$, then 1 is replaced with 0.99999 to have a defined agreement coefficient.

Table 3 contains the Monte Carlo standard errors and relative biases of agreement coefficients $\widehat{\gamma}_1$ and $\widehat{\gamma}_\pi$. The Monte Carlo relative bias of an agreement coefficient $\widehat{\gamma}$ denoted by $RB(\widehat{\gamma})$ is obtained as follows:

$$RB(\widehat{\gamma}) = \left(\frac{1}{10{,}000} \sum_{s=1}^{10{,}000} \widehat{\gamma}_s - \gamma \right) \Big/ \gamma,$$

where γ is the "true" value of the agreement coefficient calculated at the population level, and $\widehat{\gamma}_s$ is the agreement coefficient estimate obtained from one specific rater-subject pair s of samples (i.e., one pair $s = (s_r, s_s)$ of samples is made up of a rater sample s_r and a subject sample s_s, and the simulation generates 10,000 such pairs). The Monte Carlo variance of $\widehat{\gamma}$ used to obtain standard errors is denoted by $V_{MC}(\widehat{\gamma})$ and given by

$$V_{MC}(\widehat{\gamma}) = \frac{1}{10{,}000} \sum_{s=1}^{10{,}000} \left(\widehat{\gamma}_s - AV(\widehat{\gamma}) \right)^2, \tag{7}$$

where $AV(\widehat{\gamma})$ is the average of all 10,000 estimates $\widehat{\gamma}_s$, $s = 1, \ldots, 10{,}000$. The standard errors so obtained are depicted in Figure 3. This figure indicates that the standard error of an agreement coefficient decreases as r or n increases, giving another indication that the rater sample also may affect the precision of inter-rater reliability substantially.

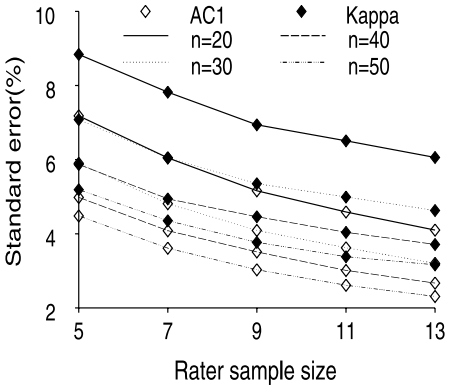

FIGURE 3.
Standard errors of AC_1 and kappa for various rater and subject sample sizes.

Table 3 also shows the relative efficiency of the AC_1 coefficient to Fleiss's generalized kappa. This efficiency is denoted by $e(\widehat{\gamma}_1|\widehat{\gamma}_\pi)$ and is defined as the ratio of $V_{MC}(\widehat{\gamma}_\pi)$ to $V_{MC}(\widehat{\gamma}_1)$. It indicates a striking gain in efficiency for the AC_1 statistic over Fleiss's generalized kappa, which ranges from 40% to 140%. The relative bias of the AC_1 and generalized kappa are often very small even for small sample sizes. It is consistently smaller than 0.30% for the AC_1 statistic, and smaller than 3% for the generalized kappa statistic. While the generalized kappa has a tendency to underestimate the true value, the AC_1 tends to overestimate it very slightly. The last two columns of Table 3 contain the percent of total variance accounted for by the rater variance component for the AC_1 and the generalized kappa, respectively. The rater variance component of the AC_1 statistic represents between 25% and 40% of the total AC_1 variance depending on the sample sizes. That percent varies between 15% and 30% for the generalized kappa. These results show that neglecting the rater variance component may lead to a dramatic understatement of the total variance.

We have studied four variance estimators in this Monte Carlo simulation. These are v_1 the variance estimator of $\widehat{\gamma}_1$ conditionally upon the rater sample given in (1), v_1^* the unconditional variance estimator of $\widehat{\gamma}_1$ given by (6) with function $f(x) = (1-x)/(q-1)$, v_π the variance estimator of Fleiss's generalized kappa $\widehat{\gamma}_\pi$ conditionally upon the rater sample, given by (3), and v_π^* the unconditional variance of $\widehat{\gamma}_\pi$ given by (6) with function $f(x) = x$. The results presented in Table 5 demonstrate the superior performance of the proposed variance estimators v_1^* and v_π^* over v_1 and v_π for performing statistical inference to both subject and rater universes. This superior performance appears in the form of a better coverage rate of the resulting confidence intervals.

Table 5 shows the relative biases of the four variance estimators under investigation as well as the Monte Carlo coverage rates of the corresponding 95% confidence intervals. The relative bias $RB(v)$ of a variance estimator v is defined as the relative difference between its Monte Carlo expectation and the Monte Carlo variance of the agreement coefficient. The Monte Carlo expectation of a variance estimator v is obtained by averaging all 10,000 variance estimates v_s obtained from each replicate pair of samples s. More formally, we have that

$$RB(v) = \left(\frac{1}{10{,}000} \sum_{s=1}^{10{,}000} v_s - V_{MC}(\widehat{\gamma}) \right) \Big/ V_{MC}(\widehat{\gamma}),$$

where the Monte Carlo variance $V_{MC}(\widehat{\gamma})$ is given by (7). To compute the coverage rates of Table 5, a 95% confidence interval was constructed for each variance estimator under study and each

TABLE 4.
Ratings of 40 images of unclothed persons on a five-point sexual maturity rating (or Tanner) scale.

Subject	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7	Rater 8	Rater 9
1	2	2	2	2	2	2	2	2	1
2	5	5	5	5	5	5	5	5	5
3	4	3	3	4	3	3	3	3	4
4	4	3	3	3	3	3	2	3	4
5	1	1	1	1	1	2	1	1	2
6	4	4	4	4	4	4	4	4	5
7	5	5	5	5	5	5	5	5	5
8	1	1	1	1	1	1	1	1	1
9	5	5	5	5	5	5	5	5	5
10	1	1	2	1	1	1	2	1	2
11	2	2	2	2	2	2	2	2	3
12	2	2	3	3	3	2	2	2	3
13	1	1	1	2	2	1	2	1	2
14	2	3	3	3	3	2	3	3	3
15	5	5	5	5	5	5	5	5	5
16	5	5	5	5	5	5	5	5	5
17	3	3	3	4	3	3	3	3	3
18	3	3	3	3	3	3	3	3	4
19	1	2	1	2	1	1	2	1	3
20	1	1	1	1	1	1	1	1	2
21	1	1	1	1	1	1	1	1	1
22	4	4	4	4	3	3	4	4	4
23	2	3	2	3	3	1	3	2	2
24	4	2	3	4	3	2	4	4	2
25	2	4	2	2	2	3	2	2	5
26	5	5	5	5	5	5	5	5	5
27	1	2	1	1	2	1	1	1	1
28	5	5	5	5	5	3	5	5	5
29	2	2	2	2	2	2	2	2	3
30	3	3	3	3	4	3	3	3	3
31	3	4	3	4	3	3	3	3	3
32	3	3	2	3	3	2	2	3	4
33	5	4	5	5	5	4	4	5	5
34	3	4	3	3	4	3	4	4	3
35	1	1	1	1	2	1	1	1	1
36	1	1	1	1	1	1	1	1	1
37	4	4	4	4	4	4	4	4	5
38	5	5	5	5	5	5	5	5	5
39	3	3	3	4	3	3	4	3	5
40	1	1	1	1	1	1	1	1	1

replicate sample s. The coverage rate is then obtained as the relative number of times the interval contains the "true" population agreement coefficient.

It follows from Table 5 that the two unconditional variance estimators v_1^* and v_π^* have relative biases that decrease quickly as the rater and subject sample sizes increase. The variance estimators v_1 and v_π, which are obtained conditionally upon the rater sample, have relative biases that tend to increase with the sample sizes. They both dramatically underestimate the "true" variance, are therefore not consistent estimators, and cannot successfully be used for constructing confidence intervals. Moreover, the seemingly larger relative bias of v_1^* as compared to that

TABLE 5.
Relative biases of agreement coefficient variances and coverage rates of associated confidence intervals.

r	n	$RB(v_1)$ (%)	$RB(v_\pi)$ (%)	$RB(v_1^*)$ (%)	$RB(v_\pi^*)$ (%)	$CR(v_1)$ (%)	$CR(v_\pi)$ (%)	$CR(v_1^*)$ (%)	$CR(v_\pi^*)$ (%)
5	20	−13.9	−13.8	12.9	4.9	91.5	91.4	94.9	94.2
	30	−25.6	−22.6	1.1	−3.3	89.5	90.2	93.8	93.3
	40	−32.7	−28.3	−5.0	−7.8	88.6	89.6	93.3	93.0
	50	−44.6	−39.2	−17.2	−18.6	85.0	86.9	91.6	91.4
7	20	−16.3	−15.6	17.0	6.1	91.0	91.2	95.5	94.5
	30	−22.6	−19.7	13.1	4.6	90.1	91.3	95.3	94.6
	40	−31.4	−23.7	5.4	3.1	88.8	90.6	94.8	94.7
	50	−41.2	−34.9	−4.1	−7.9	86.0	87.8	93.4	93.2
9	20	−13.4	−11.3	22.9	11.0	91.5	92.4	95.9	95.5
	30	−19.7	−15.4	19.3	9.5	91.2	92.2	96.1	95.4
	40	−29.6	−22.6	9.7	3.7	89.3	91.1	95.0	95.0
	50	−37.2	−28.3	4.7	0.7	87.4	90.0	94.7	94.7
11	20	−13.1	−11.8	21.6	7.3	91.6	92.5	96.2	95.3
	30	−18.9	−15.2	18.0	6.3	91.2	92.4	95.9	95.2
	40	−23.9	−18.7	16.0	5.5	90.6	91.9	95.9	95.3
	50	−33.0	−22.8	9.5	4.7	88.9	91.3	95.5	95.1
13	20	−10.9	−9.2	19.8	6.6	92.0	93.3	95.8	95.6
	30	−14.7	−11.6	18.6	6.1	92.2	92.9	96.1	95.5
	40	−20.7	−13.0	15.4	7.5	91.2	93.1	95.9	95.5
	50	−29.2	−21.0	9.1	1.1	89.6	92.2	95.5	95.1

of v_π^*, is due to the often small Monte Carlo variance of $\widehat{\gamma}_1$ that appears in the denominator of the relative bias.

The coverage rates of the two unconditional variance estimators are all reasonably close to their nominal values of 95% when the rater sample size is 7 or bigger. Conditionally upon the rater sample, the variance estimators v_1 and v_π yield confidence intervals with poor coverage rates even when the rater and subject sample sizes are large.

7. Concluding Remarks

The goal of this research was to develop statistical procedures for inferring simultaneously to the universe of subjects and that of raters. The procedures discussed by Gwet (2008) are valid for inferring to the universe of subject only for a given sample of raters. The limitations of procedures that are conditional on the rater sample can be serious if the researcher aims to project his findings to a bigger universe of raters.

We have proposed estimators for computing the variances of the AC_1 and Fleiss's generalized kappa statistics that take into account the sampling variability of the rater sample. The proposed unconditional variance estimators were shown to be consistent for estimating the true unconditional variance. Moreover, the AC_1 and generalized kappa sampling distributions were shown under certain conditions to be asymptotically normal when the subject and rater sample sizes increase simultaneously. The Monte Carlo simulation validated this result as seen by the coverage rates of the confidence intervals. In general, the Monte Carlo experiment of Section 6 confirmed the predictions of the asymptotic theory. This experiment showed that the unconditional variance estimators have a relative bias that gets smaller as the subject and rater sample

sizes increase, and that the associated confidence intervals should have a coverage rate close to their nominal value of 95%.

Researchers who are only interested in the raters who participated in the experiment may still want use the variance expressions of (1) and (3). However, many reliability studies show that researchers have a tendency to generalize their results to bigger rater universes. The reader should be aware that the variance estimators presented in this paper are derived under the assumption that there is no missing rating. That is, each rater provided a rating for all subjects in the sample. In practice, this will sometimes not be true, in which case the unconditional variance estimators become rough approximations that may still be useful. A more satisfactory treatment of the problem of missing rating would use the jackknife approach to variance estimation. This will be discussed in another paper.

Appendix

A.1. Asymptotics in Finite Population Sampling

The setup for asymptotics is similar to the one often used in finite population sampling. It was described in detail by Fuller and Isaki (1981) and Isaki and Fuller (1982). The only difference here is that rather than using a single population of inference, as is generally the case in finite population sampling, we have to use a finite population of subjects and a finite population of raters as well.

This setup for asymptotics leads us to consider one sequence of finite subject populations and another sequence of finite rater populations. Without loss of generality, we will index both sequences by t. The two sequences $\{\mathcal{S}_t\}_{t\geq 1}$ (for subject populations) and $\{\mathcal{R}_t\}_{t\geq 1}$ (for rater populations) are assumed to grow bigger and bigger as their index t increases. In other words, we have that $\mathcal{S}_1 \subset \mathcal{S}_2 \subset \cdots \subset \mathcal{S}_t \subset \cdots$ and $\mathcal{R}_1 \subset \mathcal{R}_2 \subset \cdots \subset \mathcal{R}_t \subset \cdots$.

In addition to the assumptions defining the framework for asymptotics, we will need the following general conditions to prove our results:

(a) $\exists\, \nu_0$ $(0 < \nu_0 < 1)$ such that $\lim_{t\to\infty} r_t/R_t = \nu_0$,
(b) $\exists\, \theta_e$ $(0 \leq \theta_e < 1)$ such that $p_{e|t} = \theta_e + o_p(1)$, and
(c) $\exists\, f_0$ $(0 < f_0 < 1)$ such that $\lim_{t\to\infty} n_t/N_t = f_0$.

where r_t and n_t represent, respectively, the rater sample size and the subject sample size taken from the tth populations $\mathcal{R}_t$ and $\mathcal{U}_t$. Similarly, R_t and N_t are, respectively, the sizes of populations $\mathcal{R}_t$ and $\mathcal{U}_t$. The limited number of conditions needed to prove the large-sample results is due to the fact that the variables we are dealing with are probabilities that take values between 0 and 1.

A.2. Magnitude Order of Real Variables, and Stochastic Magnitude Order of Random Variables: the O, o, O_p, and o_p Notations

When studying large-sample approximations in statistics, there are situations where only the order of magnitude of certain terms is relevant. The very nature of these terms is often irrelevant, especially when it can be established that they will converge to 0 as the sample size increases. Special mathematical notations known as O ("big O") and o ("little o") or O_p ("big O p") and o_p ("little o p") are often used to characterize all terms with the same order of magnitude, or with smaller order of magnitude compared to other simpler terms. Below is a short reminder of the definitions of these notations.

Let $(\omega_n)_{n\geq 1}$ be a sequence of numbers.

- The notation $o(\omega_n)$ (i.e. "little o" of omega n) represents any term with a magnitude order smaller than ω_n. That is the ratio $o(\omega_n)/\omega_n$ goes to 0 as n increases. For example, $\sqrt{n} = o(n)$ because $\sqrt{n}/n \to 0$ as $n \to \infty$.
- $O(\omega_n)$ (i.e. "big O" of omega n) represents any term with the same magnitude order as ω_n. That is the ratio $O(\omega_n)/\omega_n$ is bounded as n increases; or there exists a constant K, and an integer n_0 such that $|O(\omega_n)/\omega_n| \leq K$ for $n > n_0$. For example, $3/\sqrt{n} + 45/n^3 = O(1/\sqrt{n})$.
- $o_p(\omega_n)$ (i.e., little *o.p.* of omega n) refers to any random variable whose stochastic order of magnitude is smaller than that of ω_n. That is, as n increases, the ratio $o_p(\omega_n)/\omega_n$ converges in probability to 0, or the probability that $o_p(\omega_n)/\omega_n$ is bounded by an arbitrarily small positive number converges to 0. Mathematically, it means $\forall \epsilon > 0$, $\exists n_\epsilon$ (an integer) such that if $n > n_\epsilon$, then $P(|o_p(\omega_n)/\omega_n| < \epsilon) > 1 - \epsilon$.
- $O_p(\omega_n)$ refers to any random variable with the same stochastic order of magnitude as ω_n. That is, as n increases, the probability that the ratio $O_p(\omega_n)/\omega_n$ is bounded converges to 1. Mathematically, it means $\forall \epsilon > 0$, $\exists K_\epsilon > 0$, and an integer n_ϵ such that if $n > n_\epsilon$, then $P(|O_p(\omega_n)/\omega_n| < K_\epsilon) > 1 - \epsilon$.
- Note that the well-known Chebychev's inequality states that the difference between a random variable X and its mean μ has the same stochastic order of magnitude as the standard deviation of that random variable. According to Chebychev's inequality, for any real number $k > 0$, $P(|X - \mu| < k\sigma) > 1 - 1/k^2$. An interesting consequence of this inequality is that for a sequence of independent and identically distributed random variables $(X_n)_{n \geq 1}$ with mean μ and standard deviation σ, the difference $\overline{X} - \mu$ has the same stochastic order of magnitude as $1/\sqrt{n}$ since the standard deviation of $\overline{X}$ is $\sigma/\sqrt{n}$. This follows from the fact that for all $\epsilon > 0$, $P(\frac{|\overline{X}-\mu|}{1/\sqrt{n}} < \frac{\sigma}{\sqrt{\epsilon}}) > 1 - \epsilon$.
- Any sequence of random variables $(X_n)_{n \geq 1}$ that converges in law (i.e., the sequence of the distribution functions of the X_n's converges to a distribution function), then that sequence of random variables is bounded in probability.

A.3. Proof of Lemma 1

Note that p_a is a multivariate function of $\boldsymbol{p} = (p_{ik}, i = 1, \ldots, n; k = 1, \ldots, q)$. The Taylor series expansion of p_a at $\boldsymbol{P} = (P_{ik}, i = 1, \ldots, n;\ k = 1, \ldots, q)$ leads to the following expression:

$$p_a = \frac{1}{n}\sum_{i=1}^{n} P_{a|i} + \sum_{i=1}^{n}\sum_{k=1}^{q-1}(p_{ik} - P_{ik})\frac{\partial p_a}{\partial x_{ik}}\bigg|_{\boldsymbol{x}=\boldsymbol{P}}$$
$$+ \frac{1}{2}\sum_{i,j=1}^{n}\sum_{k,l=1}^{q-1}(p_{ik} - P_{ik})(p_{jl} - P_{jl})\frac{\partial^2 p_a}{\partial x_{ik}\partial x_{jl}}\bigg|_{\boldsymbol{x}=\boldsymbol{P}(\theta)},$$

where $\boldsymbol{P}(\theta) = \theta\boldsymbol{p} + (1-\theta)\boldsymbol{P}$ for some $\theta \in (0, 1)$. We have that $\partial p_a/\partial x_{ik} = 2r(x_{ik} - x_{iq})/(n(r-1))$, $\partial^2 p_a/\partial x_{ik}^2 = 4r/(n(r-1))$, and if $k \neq l$ we have $\partial^2 p_a/\partial x_{ik}\partial x_{il} = 2r/(n(r-1))$. For $i \neq j$, $\partial^2 p_a/\partial x_{ik}\partial x_{il} = 0$. Thus,

$$p_a = P_{a|\bullet} + \frac{2}{n}\sum_{i=1}^{n}\sum_{k=1}^{q} P_{ik}(p_{ik} - P_{ik}) + A_1, \tag{8}$$

where A_1 is given by

$$A_1 = \frac{2}{n(r-1)}\sum_{i=1}^{n}\sum_{k=1}^{q} P_{ik}(p_{ik} - P_{ik}) + \frac{r}{n(r-1)}\sum_{i=1}^{n}\sum_{k=1}^{q}(p_{ik} - P_{ik})^2,$$

and $P_{a|\bullet} = (P_{a|1} + \cdots + P_{a|n})/n$.

Let us consider p_e as a function of the $q-1$ dimensional vector $\boldsymbol{x} = (x_1, \ldots, x_k, \ldots, x_{q-1})^\top$ (with $0 \leq x_k \leq 1$) defined as

$$p_{e\gamma}(\boldsymbol{x}) = \sum_{k=1}^{q} g(x_k) \quad \text{where } x_q = 1 - \sum_{k=1}^{q-1} x_k.$$

Let $\widehat{\pi}_k^*$ and A_2 be defined as

$$\widehat{\pi}_k^* = \frac{1}{n}\sum_{i=1}^{n} P_{ik} \quad \text{and} \quad A_2 = \frac{1}{2}\sum_{k=1}^{q} g''\big(\pi_k(\theta)\big)\big(\widehat{\pi}_k - \pi_k\big)^2.$$

Using a Taylor series expansion of p_e in the neighborhood of $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k, \ldots, \pi_q)^\top$ leads to the following expression:

$$\begin{aligned} p_e &= P_e + \sum_{k=1}^{q} g'(\pi_k)\big(\widehat{\pi}_k - \pi_k\big) + A_2, \\ &= P_e + \sum_{k=1}^{q} g'(\pi_k)\big[\big(\widehat{\pi}_k - \widehat{\pi}_k^*\big) + \big(\widehat{\pi}_k^* - \pi_k\big)\big] + A_2, \\ &= P_e + \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{q} g'(\pi_k)(p_{ik} - P_{ik}) + \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{q} g'(\pi_k)(P_{ik} - \pi_k) + A_2. \end{aligned} \tag{9}$$

Let A_3 and A_4 be defined as

$$\begin{aligned} A_3 &= \frac{(p_a - p_e)(p_e - P_e)^2}{(1 - P_e(\theta))^3}, \quad \text{where } P_e(\theta) = \theta p_e + (1-\theta)P_e \text{ for } \theta \in (0,1), \\ A_4 &= \frac{(p_e - P_e)(p_a - P_a) - (p_e - P_e)^2}{(1 - P_e)^2}. \end{aligned}$$

The agreement coefficient $\widehat{\gamma}$ can be seen as a function of p_e. A Taylor series expansion of $\widehat{\gamma}$ in the neighborhood of P_e yields

$$\begin{aligned} \widehat{\gamma} &= \frac{p_a - p_e}{1 - P_e} + \frac{(p_a - p_e)(p_e - P_e)}{(1 - P_e)^2} + A_3, \\ &= \gamma + \frac{(p_a - P_{a|\bullet}) + (P_{a|\bullet} - P_a) - (1-\gamma)(p_e - P_e)}{1 - P_e} + A_3 + A_4. \end{aligned}$$

If follows from (8) and (9) that γ can be expressed as follows:

$$\widehat{\gamma} = \gamma + \frac{1}{n}\sum_{i=1}^{n} \lambda_i + R, \quad \text{where } R = \frac{A_1 - (1-\gamma)A_2}{1 - P_e} + A_3 + A_4. \tag{10}$$

To complete the proof of Lemma 2, we must prove that $R = O_p(1/n) + O_p(1/r)$. We have that $(\widehat{\pi}_k - \pi_k)^2 = (\widehat{\pi}_k - \widehat{\pi}_k^*)^2 + (\widehat{\pi}_k^* - \pi_k)^2 + 2(\widehat{\pi}_k - \widehat{\pi}_k^*)(\widehat{\pi}_k^* - \pi_k)$. It follows from the central limit theorem that $(\widehat{\pi}_k^* - \pi_k)^2 = O_p(1/n)$. Moreover, $E(\widehat{\pi}_k - \widehat{\pi}_k^*)^2 = E_{\mathcal{R}} V_{\mathcal{S}}(\widehat{\pi}_k - \widehat{\pi}_k^* | \mathcal{R}) + V_{\mathcal{R}} E_{\mathcal{S}}(\widehat{\pi}_k - \widehat{\pi}_k^* | \mathcal{R})$. After some algebra, one can establish that

$$E\big(\widehat{\pi}_k - \widehat{\pi}_k^*\big)^2 = \xi_R \xi_N \frac{(1-f_n)(1-f_r)}{nr}\big(\sigma_k^2 - \eta_k + \pi_k^2\big) + \xi_R \frac{1-f_r}{r}\big(\sigma_k^2/N + \eta_k - \pi_k^2\big), \tag{11}$$

where $\xi_R = R/(R-1)$, $\xi_N = N/(N-1)$, σ_k^2 and η_k are defined as

$$\sigma_k^2 = \frac{1}{N}\sum_{i=1}^{N} P_{ik}(1-P_{ik}), \qquad \eta_k = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} P_{ik}^{(jk)},$$

$P_{ik}^{(jk)}$ being the proportion of raters who classified subjects i and j into category k. It follows from (11) and Chebychev's inequality that $(\widehat{\pi}_k - \widehat{\pi}_k^*)^2 = O_p(1/r)$. Thus, $A_2 = O_p(1/r) + O_p(1/n)$.

It follows from the central limit theorem that $p_{ik} - P_{ik} = O_p(1/r^{1/2})$. Therefore, one can establish that $A_1 = O_p(1/r^{3/2}) + O_p(1/r)$, $A_3 = O_p(1/r)$, and $A_4 = O_p(1/r)$. □

A.4. Proof of Lemma 2

The unconditional variance of $\overline{\lambda}$ is given by

$$V(\overline{\lambda}) = E_{\mathcal{R}} V_{\mathcal{S}}(\overline{\lambda}|\mathcal{R}) + V_{\mathcal{R}} E_{\mathcal{S}}(\overline{\lambda}|\mathcal{R}).$$

Let vector $\boldsymbol{x}_i = \boldsymbol{p}_i - \boldsymbol{P}_i$. Since $\lambda_i = (\gamma_i^* - \gamma) + \boldsymbol{\Phi}_i^\top \boldsymbol{x}_i$, the conditional variance of $\overline{\lambda}$ given the sample of raters is given by

$$\begin{aligned} V_{\mathcal{S}}(\overline{\lambda}|\mathcal{R}) &= \frac{1-f_n}{n}\frac{1}{N-1}\sum_{i=1}^{N}\big(\lambda_i - \overline{\Lambda}\big)^2, \quad \text{where } \overline{\Lambda} = \frac{1}{N}\sum_{i=1}^{N}\lambda_i \text{ and } f_n = n/N, \\ &= \frac{1-f_n}{n}\left\{\frac{1}{N-1}\sum_{i=1}^{N}\big(\gamma_i^* - \gamma\big)^2 + \frac{1}{N-1}\sum_{i=1}^{N} 2\big(\gamma_i^* - \gamma\big)\boldsymbol{\Phi}_i^\top \boldsymbol{x}_i \right. \\ &\quad \left. + \frac{1}{N-1}\sum_{i=1}^{N}\big(\boldsymbol{\Phi}_i^\top \boldsymbol{x}_i\big)^2 - \frac{N}{N-1}\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\Phi}_i^\top \boldsymbol{x}_i\right)^2\right\}. \end{aligned}$$

Note that $E_{\mathcal{R}}(\boldsymbol{x}_i) = 0$. Therefore,

$$E_{\mathcal{R}} V_{\mathcal{S}}(\overline{\lambda}|\mathcal{R}) = \frac{1-f_n}{n}\left\{S_\gamma^{*2} + \frac{1}{N-1}\sum_{i=1}^{N} E_{\mathcal{R}}\big(\boldsymbol{\Phi}_i^\top \boldsymbol{x}_i\big)^2 - \frac{1}{N(N-1)} E_{\mathcal{R}}\left(\sum_{i=1}^{N}\boldsymbol{\Phi}_i^\top \boldsymbol{x}_i\right)^2\right\}.$$

But, $E_{\mathcal{R}}(\boldsymbol{\Phi}_i^\top \boldsymbol{x}_i)^2 = E_{\mathcal{R}}\boldsymbol{\Phi}_i^\top(\boldsymbol{x}_i \boldsymbol{x}_i^\top)\boldsymbol{\Phi}_i = \boldsymbol{\Phi}_i^\top V_{\mathcal{R}}(\boldsymbol{p}_i)\boldsymbol{\Phi}_i$. Also,

$$\begin{aligned} E_{\mathcal{R}}\left(\sum_{i=1}^{N}\boldsymbol{\Phi}_i^\top \boldsymbol{x}_i\right)^2 &= \sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{\Phi}_i^\top E_{\mathcal{R}}\big(\boldsymbol{x}_i \boldsymbol{x}_j^\top\big)\boldsymbol{\Phi}_j, \\ &= \sum_{i=1}^{N}\boldsymbol{\Phi}_i^\top V_{\mathcal{R}}(\boldsymbol{p}_i)\boldsymbol{\Phi}_i + \sum_{i\neq j}^{N}\sum_{j=1}^{N}\boldsymbol{\Phi}_i^\top \mathrm{COV}_{\mathcal{R}}(\boldsymbol{p}_i, \boldsymbol{p}_j)\boldsymbol{\Phi}_j. \end{aligned}$$

Therefore,

$$\begin{aligned} E_{\mathcal{R}} V_{\mathcal{S}}(\overline{\lambda}|\mathcal{R}) = \frac{1-f_n}{n}\left\{S_\gamma^{*2} + \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\Phi}_i^\top V_{\mathcal{R}}(\boldsymbol{p}_i)\boldsymbol{\Phi}_i \right. \\ \left. - \frac{1}{N(N-1)}\sum_{i\neq j}^{N}\sum_{j=1}^{N}\boldsymbol{\Phi}_i^\top \mathrm{COV}_{\mathcal{R}}(\boldsymbol{p}_i, \boldsymbol{p}_j)\boldsymbol{\Phi}_j\right\}. \end{aligned} \tag{12}$$

To obtain $V_{\mathcal{R}}(\boldsymbol{p}_i)$, one would note that,

$$V_{\mathcal{R}}(p_{ik}) = \frac{P_{ik}Q_{ik}}{r}\frac{R-r}{R-1} = \frac{1-f_r}{r}\frac{R}{R-1}P_{ik}Q_{ik},$$

where $f_r = r/R$, and $Q_{ik} = 1 - P_{ik}$ (see Cochran, 1977). Likewise, one can establish that the covariance of p_{ik} and $p_{ik'}$ for two different categories k and k' is given by

$$\mathrm{COV}_{\mathcal{R}}(p_{ik}, p_{ik'}) = -\frac{1-f_r}{r}\frac{R}{R-1}P_{ik}P_{ik'}.$$

To see this, let us define the following two random variables, ε_l and $\eta_{ik}^{(l)}$, for a rater l, a subject i, and a category k:

$$\eta_{ik}^{(l)} = \begin{cases} 1 & \text{if rater } l \text{ classifies subject } i \text{ into category } k, \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

And $\varepsilon_l = 1$ if rater l has been selected from the rater population for inclusion in the rater sample, and $\varepsilon_l = 0$ otherwise. It follows that

$$\begin{aligned}\mathrm{COV}_{\mathcal{R}}(p_{ik}, p_{ik'}) &= E_{\mathcal{R}}\,p_{ik}p_{ik'} - P_{ik}P_{ik'}, \\ &= \frac{1}{r^2}E_{\mathcal{R}}\left(\sum_{l=1}^{r}\sum_{l'=1}^{r}\eta_{ik}^{(l)}\eta_{ik'}^{(l')}\right) - P_{ik}P_{ik'}, \\ &= \frac{1}{r^2}E_{\mathcal{R}}\left(\sum_{l=1}^{R}\sum_{l'=1}^{R}\eta_{ik}^{(l)}\eta_{ik'}^{(l')}E_{\mathcal{R}}(\varepsilon_l\varepsilon_{l'})\right) - P_{ik}P_{ik'}.\end{aligned}$$

Since the sampling of raters is carried out without replacement, $E_{\mathcal{R}}(\varepsilon_l\varepsilon_{l'})$ equals the probability to select raters l and l' from the rater population for inclusion in the sample. Thus, $E_{\mathcal{R}}(\varepsilon_l\varepsilon_{l'}) = r(r-1)/[R(R-1)]$. Because the same rater cannot classify the same subject into two different categories, $\eta_{ik}^{(l)}\eta_{ik'}^{(l)} = 0$ for $k \neq k'$. Therefore,

$$\begin{aligned}\mathrm{COV}_{\mathcal{R}}(p_{ik}, p_{ik'}) &= \frac{r-1}{rR(R-1)}\sum_{l=1}^{R}\eta_{ik}^{(l)}\left(\sum_{l'\neq l}^{R}\eta_{ik'}^{(l')}\right) - P_{ik}P_{ik'}, \\ &= -\frac{1-f_r}{r}\frac{R}{R-1}P_{ik}P_{ik'}.\end{aligned}$$

Thus,

$$V_{\mathcal{R}}(\boldsymbol{p}_i) = \frac{1-f_r}{r}\frac{R}{R-1}\big(\mathrm{diag}(\boldsymbol{P}_i) - \boldsymbol{P}_i\boldsymbol{P}_i^{\top}\big).$$

Using the same technique, the covariance of $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$ is obtained as

$$\mathrm{COV}_{\mathcal{R}}(\boldsymbol{p}_i, \boldsymbol{p}_j) = \frac{1-f_r}{r}\frac{R}{R-1}\big(\boldsymbol{P}_{i\bullet}^{(j\bullet)} - \boldsymbol{P}_i\boldsymbol{P}_j^{\top}\big), \quad \text{where } \boldsymbol{P}_{i\bullet}^{(j\bullet)} = \big(P_{ik}^{(jl)}\big)_{1\leq k\leq q, 1\leq l\leq q}.$$

Let $\Psi_i = \boldsymbol{\Phi}_i^{\top}(\mathrm{diag}(\boldsymbol{P}_i) - \boldsymbol{P}_i\boldsymbol{P}_i^{\top})\boldsymbol{\Phi}_i$, and $\Psi_{ij} = \boldsymbol{\Phi}_i^{\top}(\boldsymbol{P}_{i\bullet}^{(j\bullet)} - \boldsymbol{P}_i\boldsymbol{P}_j^{\top})\boldsymbol{\Phi}_j$. It follows that

$$\begin{aligned}E_{\mathcal{R}}V_{\mathcal{S}}\big(\overline{\lambda}|\mathcal{R}\big) &= \frac{1-f_n}{n}\left\{S_{\gamma}^{*2} + \frac{1-f_r}{r}\frac{R}{R-1}\left(\frac{1}{N}\sum_{i=1}^{N}\Psi_i - \frac{1}{N(N-1)}\sum_{i\neq j}^{N}\sum^{N}\Psi_{ij}\right)\right\}, \\ &= \frac{1-f_n}{n}S_{\gamma}^{*2} + A.\end{aligned}$$

The second part of $V(\overline{\lambda})$ is derived as

$$V_{\mathcal{R}}E_{\mathcal{S}}(\overline{\lambda}|\mathcal{R}) = V_{\mathcal{R}}\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\Phi}_i^{\top}\boldsymbol{x}_i\right) = E_{\mathcal{R}}\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\Phi}_i^{\top}\boldsymbol{x}_i\right)^2,$$

$$= E_{\mathcal{R}}\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{\Phi}_i^{\top}\boldsymbol{x}_i\boldsymbol{x}_j^{\top}\boldsymbol{\Phi}_j\right) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{\Phi}_i^{\top}\mathrm{COV}_{\mathcal{R}}(\boldsymbol{p}_i,\boldsymbol{p}_j)\boldsymbol{\Phi}_j,$$

$$= \frac{1-f_r}{r}\frac{R}{R-1}\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\Psi_{ij} = \frac{1-f_r}{r}\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\Psi_{ij} + B,$$

where $\Psi_{ii} = \Psi_i$.

The probabilities $P_{ik}^{(jl)}$ and P_{ik} can be expressed in terms of the binary variables $\eta_{ik}^{(l)}$ of (13) as

$$P_{ik}^{(jk)} = \frac{1}{R}\sum_{\alpha=1}^{R}\eta_{ik}^{(\alpha)}\eta_{jl}^{(\alpha)}, \quad \text{and} \quad P_{ik} = \frac{1}{R}\sum_{\alpha=1}^{R}\eta_{ik}^{(\alpha)}.$$

We have that

$$\Psi_{ij} = \sum_{k=1}^{q}\sum_{l=1}^{q}\Phi_{ik}\Phi_{jl}P_{ik}^{(jl)} - \left(\sum_{k=1}^{q}\Phi_{ik}P_{ik}\right)\left(\sum_{l=1}^{q}\Phi_{jl}P_{jl}\right),$$

$$= \frac{1}{R}\sum_{\alpha=1}^{R}\left[\left(\sum_{k=1}^{q}\Phi_{ik}\eta_{ik}^{(\alpha)}\right)\left(\sum_{l=1}^{q}\Phi_{jl}\eta_{jl}^{(\alpha)}\right)\right]$$

$$- \left\{\frac{1}{R}\sum_{\alpha=1}^{R}\left(\sum_{k=1}^{q}\Phi_{ik}\eta_{ik}^{(\alpha)}\right)\right\}\left\{\frac{1}{R}\sum_{\alpha=1}^{R}\left(\sum_{l=1}^{q}\Phi_{jl}\eta_{jl}^{(\alpha)}\right)\right\},$$

$$= \frac{1}{R}\sum_{\alpha=1}^{R}\Phi_{i\alpha_i}\Phi_{j\alpha_j} - \left(\frac{1}{R}\sum_{\alpha=1}^{R}\Phi_{i\alpha_i}\right)\left(\frac{1}{R}\sum_{\alpha=1}^{R}\Phi_{j\alpha_j}\right),$$

where α_i is the category into which rater α classified subject i.

It follows that

$$\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\Psi_{ij} = \frac{1}{R}\sum_{\alpha=1}^{R}\left(\frac{1}{N}\sum_{i=1}^{N}\Phi_{i\alpha_i}\right)\left(\frac{1}{N}\sum_{j=1}^{N}\Phi_{j\alpha_j}\right) - \left(\frac{1}{R}\sum_{\alpha=1}^{R}\Phi_{(\alpha)}\right)^2,$$

$$= \frac{1}{R}\sum_{\alpha=1}^{R}\Phi_{(\alpha)}^2 - \left(\frac{1}{R}\sum_{\alpha=1}^{R}\Phi_{(\alpha)}\right)^2 = \frac{1}{R}\sum_{\alpha=1}^{R}\left(\Phi_{(\alpha)} - \overline{\Phi}_{(\bullet)}\right)^2,$$

where $\overline{\Phi}_{(\alpha)}$ is the average of all N values $\Phi_{i\alpha_i}$ $(i = 1,\ldots,N)$, and $\overline{\Phi}_{(\bullet)}$ is the average of all R values $\Phi_{(\alpha)}$ $(\alpha = 1,\ldots,R)$. The proof is completed by noting that,

$$\Phi_{(\alpha)} - \Phi_{(\bullet)} = \frac{2}{1-P_e}\left\{\left(P_a^{(\alpha)} - P_a^{(\bullet)}\right) - (1-\gamma)\sum_{k=1}^{q}\left[\pi_k\left(f\left(\pi_k^{(\alpha)}\right) - f(\pi_k)\right)\right]\right\},$$

$$= 2\left(\gamma_{(\alpha)} - \overline{\gamma}_{(\bullet)}\right). \qquad \square$$

References

Bartfay, E., & Donner, A. (2001). Statistical inferences for inter-observer agreement studies with nominal outcome data. *The Statistician*, *50*, 135–146.
Bennet, E.M., Alpert, R., & Goldstein, A.C. (1954). Communications through limited response questioning. *Public Opinion Quarterly*, *18*, 303–308.
Berry, K.J., & Mielke, P.W. Jr. (1988). A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters. *Educational and Psychological Measurement*, *48*, 921–933.
Brennan, R.L., & Prediger, D.J. (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.
Cicchetti, D.V., & Feinstein, A.R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, *43*, 551–558.
Cochran, W.G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.
Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.
Conger, A.J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, *88*, 322–328.
Cook, R.J. (1998). Kappa and its dependence on marginal rates. In P. Armitage & T. Colton (Eds.), *Encyclopedia of biostatistics* (pp. 2166–2168). New York: Wiley.
Donner, A., & Eliasziw, M. (1992). A goodness-of-fit approach to inference procedures for the kappa statistic: Confidence interval construction, significance-testing and sample size estimation. *Statistics in Medicine*, *11*, 1511–1519.
Feinstein, A.R., & Cicchetti, D.V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*, 543–549.
Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*, 378–382.
Fuller, W.A., & Isaki, C.T. (1981). Survey design under superpopulation models. In D. Krewski, J.N.K. Rao, & R. Platek (Eds.), *Current topics in survey sampling* (pp. 199–226). New York: Academic Press.
Goodman, L.A., & Kruskal, W.H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, *49*, 1732–1769.
Gwet, K. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*(1).
Holley, J.W., & Guilford, J.P. (1964). A note on the *G* index of agreement. *Educational and Psychological Measurement*, *24*, 749–753.
Isaki, C.T., & Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, *77*, 89–96.
Janson, H., & Olsson, U. (2001). A measure of agreement for interval or nominal multivariate observations. *Educational and Psychological Measurement*, *61*, 277–289.
Janson, H., & Olsson, U. (2004). A measure of agreement for interval or nominal multivariate observations by different sets of judges. *Educational and Psychological Measurement*, *64*, 62–70.
Janson, S., & Vegelius, J. (1979). On generalizations of the *G* index and the PHI coefficient to nominal scales. *Multivariate Behavioral Research*, *14*, 255–269.
Kraemer, H.C., Periyakoil, V.S., & Noda, A. (2002). Kappa coefficients in medical research. *Statistics in Medicine*, *21*, 2109–2129.
Landis, R.J., & Koch, G.G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, *33*, 363–374.
Light, R.J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin*, *76*, 365–377.
Maxwell, A.E. (1977). Coefficients of agreement between observers and their interpretation. *British Journal of Psychiatry*, *130*, 79–83.
McGraw, K.O., & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46.
Nam, J.M. (2000). Interval estimation of the kappa coefficient with binary classification and an equal marginal probability model. *Biometrics*, *56*, 583–585.
Rao, C.R. (2002). *Wiley series in probability and statistics*. *Linear statistical inference and its applications* (2nd ed.).
Schuster, C. (2004). A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales. *Educational and Psychological Measurement*, *64*, 243–253.
Schuster, C., & Smith, D.A. (2006). Estimating with a latent class model the reliability of nominal judgments upon which two raters agree. *Educational and Psychological Measurement*, *66*, 739–747.
Scott, W.A. (1955). Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly*, *XIX*, 321–325.
Simon, P. (2006). Including omission mistakes in the calculation of Cohen's kappa and an analysis of the coefficient's paradox features. *Educational and Psychological Measurement*, *66*, 765–777.
Uebersax, J.S., & Grove, W.M. (1990). Latent class analysis of diagnostic agreement. *Statistics in Medicine*, *9*, 559–572.
Uebersax, J.S., & Grove, W.M. (1993). A latent trait finite mixture analysis of rating agreement. *Biometrics*, *49*, 823–835.
Zou, G., & Klar, N. (2005). A non-iterative confidence interval estimating procedure for the intraclass kappa statistic with multinomial outcomes. *Biometrical Journal*, *5*, 682–690.
Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin*, *103*, 374–378.

Manuscript received 28 OCT 2004
Final version received 28 OCT 2007
Published Online Date: 17 JAN 2008