# INTRARATER RELIABILITY

KILEM L. GWET
STATAXIS Consulting,
Gaithersburg, Maryland

The notion of intrarater reliability will be of interest to researchers concerned about the reproducibility of clinical measurements. A rater in this context refers to any data-generating system, which includes individuals and laboratories; intrarater reliability is a metric for rater's self-consistency in the scoring of subjects. The importance of data reproducibility stems from the need for scientific inquiries to be based on solid evidence. Reproducible clinical measurements are recognized as representing a well-defined characteristic of interest. Reproducibility is a source of concern caused by the extensive manipulation of medical equipment in test laboratories and the complexity of the judgmental processes involved in clinical data gathering. Grundy (1) stresses the importance of choosing a good laboratory when measuring cholesterol levels to ensure their validity and reliability. This article discusses some basic methodological aspects related to intrarater reliability estimation. For continuous data, the intraclass correlation (ICC) is the measure of choice and will be discussed in the section entitled "Intrarater reliability for continuous scores." For nominal data, the kappa coefficient of Cohen (2) and its many variants are the preferred statistics, and they are discussed in the section entitled "nominal scale score data." The last section is devoted to some extensions of kappa-like statistics aimed at intrarater reliability coefficients for ordinal and interval data.

## 1 INTRARATER RELIABILITY FOR CONTINUOUS SCORES

A continuous clinical measurement, such as blood pressure, will be considered reproducible if repeated measures taken by the same rater under the same conditions show a rater variation that is negligible compared with the subject variation. ICC is the ratio of the between-subject variation (BSV) to the total variation [i.e., the sum of the BSV and the within-subject variation (WSV)], and it is the statistical measure most researchers adopted for quantifying the intrarater reliability of continuous data. Note that ICC reaches its maximum value of 1 when WSV (i.e., the average variation for a subject) reaches its lower bound of 0, a situation indicating that any variation in the data is because the subjects are different and not because the rater is being inconsistent. Used previously as a measure of reliability by Ebel (3) and Barko (4), the ICC has proved to be a valid measure of raters' self-consistency.

Shrout and Fleiss (5) discuss various forms of the ICC as a measure of inter-rater reliability, which quantifies the extent of agreement between raters as opposed to intrarater reliability used to measure self-consistency. The selection of a particular version of the Shrout-Fleiss ICCs is dictated by the design adopted for the intrarater reliability study. Lachin (6) also discusses various techniques for evaluating the quality of clinical trial data, which includes the ICC among others.

### 1.1 Defining Intrarater Reliability

Ratings in a typical intrarater reliability study that involves $m$ subjects and $n$ replicates per subject are conveniently organized as shown in Table 1. The entry $y_{ij}$ represents the $i$th replicate score that the rater assigned to subject $j$. This table may be transposed if the number of subjects is very large. The relationship between the ICC and the analysis of variance (ANOVA) techniques motivated the proposed disposition of rows and columns of Table 1.

The WSV is the average of the $m$ subject-level variances $s_j^2$ calculated over the $n$ replicates. More formally, the WSV is defined as follows:

$$WSV = \frac{1}{m}\sum_{j=1}^{m} S_j^2,$$

$$\text{where } S_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_{ij} - \overline{y}_{\cdot j})^2 \qquad (1)$$

**Table 1. Scores Assigned to *m* Subjects with *n* Replicates per Subject**

| | Subject | | | | | |
|---|---|---|---|---|---|---|
| Observation | 1 | 2 | ⋯ | j | ⋯ | m |
| 1 | $y_{11}$ | $y_{12}$ | ⋯ | $y_{1j}$ | ⋯ | $y_{1m}$ |
| 2 | $y_{21}$ | $y_{22}$ | ⋯ | $y_{2j}$ | ⋯ | $y_{2m}$ |
| ⋮ | ⋮ | ⋮ | ⋱ | ⋮ | ⋮ | ⋮ |
| *i* | $y_{i1}$ | $y_{i2}$ | ⋯ | $y_{ij}$ | ⋯ | $y_{im}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋱ | ⋮ |
| *n* | $y_{n1}$ | $y_{n2}$ | ⋯ | $y_{nj}$ | ⋯ | $y_{nm}$ |

and $\overline{y}_{.j}$ is the average of the $n$ replicate observations related to subject $j$. The BSV is obtained by taking the variance of the $m$ subject-level mean scores trimmed of the fraction of the WSV it contains. It is formally defined as follows:

$$BSV = BSV_o - \frac{WSV}{n},$$

$$\text{where } BSV_o = \frac{\sum_{j=1}^{m}(\overline{y}_{.j} - \overline{y}_{..})^2}{m-1} \qquad (2)$$

and $\overline{y}_{..}$ is the overall mean of all $mn$ observations.

Using Equations (1) and (2), we define the intrarater reliability coefficient $\hat{\gamma}$ (read gamma hat with the hat indicating an estimation of the "true" parameter $\gamma$ to be defined later) as follows:

$$\hat{\gamma} = \frac{BSV}{BSV + WSV} \qquad (3)$$

To illustrate the calculation of the ICC, let us consider the cholesterol level data of Table 2. Table 2 data represent cholesterol levels taken from 10 individuals who participated in the 2005 National Health and Nutrition Examination Survey (7). For the sake of illustration, I assume that the data was collected on two occasions (times 1 and 2) by the same laboratory.

Although Equations (1)–(3) can be used to obtain the ICC from Table 2 data, the more convenient approach for computing the ICC will generally be to use an ANOVA procedure either from Microsoft Excel (Microsoft Corporation, Redmond, WA), or from a standard statistical package such as SPSS (SPSS Inc., Chicago, IL) or SAS (SAS Institute Inc., Cary, NC). The ANOVA analysis will produce two mean squares (MS) known as the Mean Square for Treatments ($MS_T$), which is also referred to as the Mean Square for the model, and the Mean Square for Error ($MS_E$). The ICC can be expressed as a function of the two mean squares as follows:

$$\hat{\gamma} = \frac{MS_T - MS_E}{MS_T + (n-1)MS_E} \qquad (4)$$

Using MS Excel's Analysis ToolPak and the cholesterol data, I created the output shown in Table 3 known as the ANOVA table, in which the column labeled "MS" contains the two Mean Squares ($MS_T = 1323.56$, and $MS_E = 18.25$) needed to compute the ICC. Therefore, the intrarater reliability associated with

**Table 2. Total Cholesterol Measures (in mg/dL) taken on 10 subjects with 2 Replicates per Subject**

| | Subjects | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Time | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 152 | 202 | 160 | 186 | 207 | 205 | 160 | 188 | 147 | 151 |
| 2 | 155 | 210 | 156 | 200 | 214 | 209 | 163 | 189 | 146 | 153 |

Source: Reference 7.

**Table 3. ANOVA Table Created with MS Excel from Table 2 Cholesterol Data**

| ANOVA | | | | | | |
|---|---|---|---|---|---|---|
| *Source of Variation* | *SS* | *df* | ***MS*** | *F* | *P-value* | *F crit* |
| Between Groups | 11,912.05 | 9 | **1323.56** | 72.52 | 6.4E-08 | 3.02 |
| Within Groups | 182.5 | 10 | **18.25** | | | |
| Total | 12,094.55 | 19 | | | | |

Table 2 data is computed as follows:

$$\hat{\gamma} = \frac{1,323.56 - 18.25}{1,323.56 + (2-1) \times 18.25} = 0.973$$

Consequently, the subject factor accounts for 97.3% of the total variation observed in the cholesterol data of Table 2, which is an indication of high intrarater reliability. This estimation gives us a sense of the reproducibility of cholesterol levels. However, it also raises questions about its accuracy and the steps that could be taken to improve it. These issues can only be addressed within a well-defined framework for statistical inference.

### 1.2 Statistical Inference

The primary objective of this section is to present a framework for statistical inference that will help answer the following fundamental questions about the intrarater reliability estimate:

- Is the obtained ICC sufficiently accurate?
- Can the obtained ICC be considered valid?
- Have we used a sufficiently large number of replicates?
- Have we used a sufficiently large number of subjects?
- Can the data be collected by multiple raters?

These questions can be addressed only if this theoretical framework provides the following two key components:

1. The definition of a population parameter $\gamma$ (i.e., gamma without the hat) that represents the "true" unknown intrarater reliability being measured
2. Methods for evaluating the precision of proposed statistics $\hat{\gamma}$ with respect to the parameter of interest $\gamma$

Let us consider the simple scenario in which $y_{ij}$ results from the additive effect of a common score $\mu$, the subject effect $t_j$, and an error $\varepsilon_{ij}$ committed on the $i$th replicate score of subject $j$. This relation is mathematically expressed as follows:

$$y_{ij} = \mu + t_j + \epsilon_{ij}, i = 1, \cdot n \text{ and } j = 1, \cdots m \quad (5)$$

This example is a single-factor ANOVA model that goes along with the following assumptions:

- The $m$ subjects that participate in the intrarater reliability study form a random sample selected from a larger population of subjects of interest. Moreover, $t_j$ is a normally distributed random variable with mean 0 and variance $\sigma_t^2 > 0$.
- The error $\varepsilon_{ij}$ is a normally distributed random variable with mean 0, and variance $\sigma_{\epsilon}^2 > 0$, and it is independent of $t_j$.

A small value for $\sigma_{\epsilon}^2$ will lead to a small variation between replicate scores, which in turn should lead to a high intrarater reliability. Therefore, the theoretical parameter that represents the intrarater reliability $\gamma$ is defined as follows:

$$\gamma = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_{\epsilon}^2} \quad (6)$$

which is one of the parameters studied by McGraw and Wong (8). It follows from Equation (5) that the denominator of $\gamma$ is the total variation in the scores, and Equation (6) is the most popular form of ICC found in the

statistical literature and represents the portion of total variance that is accounted for by the between-subject variance. The statistic $\hat{\gamma}$ of Equation (3) is a sample-based approximation of $\gamma$ widely accepted in the statistical literature.

We will now present a method for evaluating the precision of the statistic $\hat{\gamma}$. How close is $\hat{\gamma}$ to $\gamma$? To answer this question we will construct a 95% confidence interval around $\gamma$; that is a range of values expected to contain the unknown $\gamma$ with 95% certainty.

Constructing a 95% confidence interval for $\gamma$ requires the calculation of the 2.5th and 97.5th percentiles of the F distribution with $m-1$ and $m(n-1)$ degrees of freedom. These two percentiles are denoted by $F_{0.975,m-1,m(n-1)}$ and $F_{0.025,m-1,m(n-1)}$, respectively, where $0.975 = 1-(1-0.95)/2$ and $0.025 = (1-0.95)/2$. Although textbooks' statistical tables provide these percentiles, they are also readily obtained from MS Excel as follows: "*=FINV(0.025,m−1,m*(n−1))*" gives the 97.5th percentile, whereas "*FINV(0.975,m−1,m*(n−1))*" for the 2.5th percentile.

Let F be defined as follows:

$$F_o = \frac{BSV_o}{WSV/n} = \frac{MS_T}{MSE} \tag{7}$$

The 95% confidence interval for $\gamma$ is obtained by noting that:

$$P(L_{95} \leq \gamma \leq U_{95}) = 0.95$$

$$L_{95} = \frac{F_0/F_{0.025,m-1,m(n-1)} - 1}{(m-1) + F/F_{0.025,m-1,m(n-1)}} \text{ and}$$

$$\mathrm{U}_{95} = \frac{F_0/F_{0.975,m-1,m(n-1)} - 1}{(m-1) + F/F_{0.975,m-1,m(n-1)}} \tag{8}$$

The 95% confidence interval width is given by:

$$W_{95} = U_{95} - L_{95} \tag{9}$$

The design of an intrarater reliability study must aim at minimizing the magnitude of the confidence interval width. The optimal values for the number $m$ of subjects and the number $n$ of replicates per subject are those that minimize $W_{95}$ of Equation (9).

## 1.3 Optimizing the Design of the Intrarater Reliability Study

An intrarater reliability study is well designed if the number of observations is determined to minimize the confidence interval length and the experimental error is kept small. This section addresses the following two questions:

1. What is the optimal number $m$ of subjects, and number $n$ of replicates per subject?
2. Can the intrarater reliability study involve two raters or more?

**1.3.1 Sample Size Determination**. Finding the optimal number of subjects and replicates per subject based the confidence interval length is the approach Giraudeau and Mary (9) used to propose guidelines for planning a reproducibility study.

Let $\omega_{95}$ be the expected width of the 95% confidence interval associated with $\gamma$. Note that $\sigma_\epsilon^2 F_o/(\sigma_\epsilon^2 + m\sigma_t^2) = F_o/(1 + m\gamma/(1-\gamma))$ where $F_o$ is defined by Equation (7) and follows the F distribution with $m-1$ and $m(n-1)$ degrees of freedom. Because $W_{95}$ defined by Equation (9) is a function of $F_o$, its expected value $\omega_{95}$ is a function of $\gamma$. The relationship between $\omega_{95}$ and $\gamma$ is depicted in Figs. 1, 2, and 3 for various values of $m$ and $n$. The $\omega_{95}$ values are calculated using a Monte-Carlo simulation approach because of the difficulty to derive a mathematical expression of the probability distribution of $W_{95}$.

For values of $\gamma$ that vary from 0 to 1 by step of 0.05, and for various combinations of *(m,n)* we simulated 10,000 observations from the F distribution with $m-1$ and *m(n−1)* degrees of freedom, and calculated 10,000 confidence intervals using Equation (8). The mean length of the 10,000 intervals was used as an estimate for $\omega_{95}$.

Each of the three figures contains two plots, and each plot shows how different values of $m$ and $n$ affect the relationship between $\gamma$ and $\omega_{95}$ for a fixed total number of observations $mn$. For the two plots of Fig. 1, the total sample sizes $mn$ are 20 and 40. For Fig. 2, the total sample sizes are 60 and 80, whereas Fig. 3's plots are based on the sample sizes of 100 and 120. All three figures tend to indicate

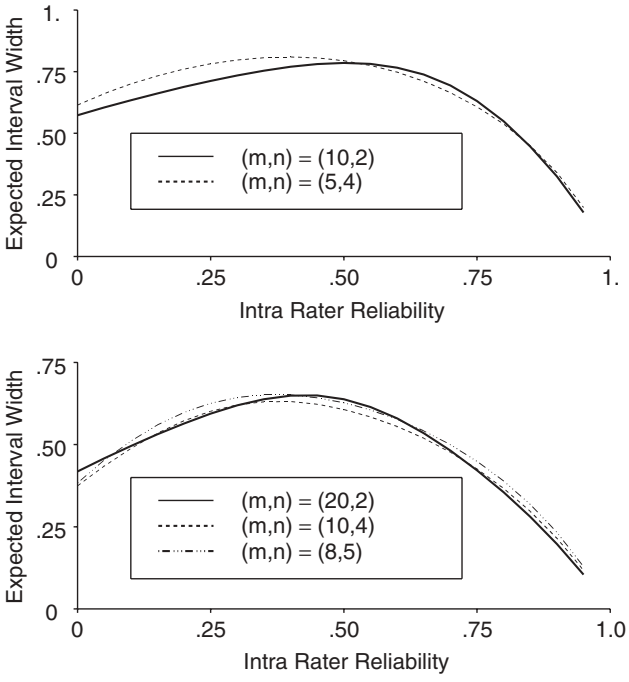


**Figure 1.** Expected width of the 95% confidence interval as a function of $\gamma$ for $m$ and $n$ values that correspond to $mn = 20$ and $mn = 40$.

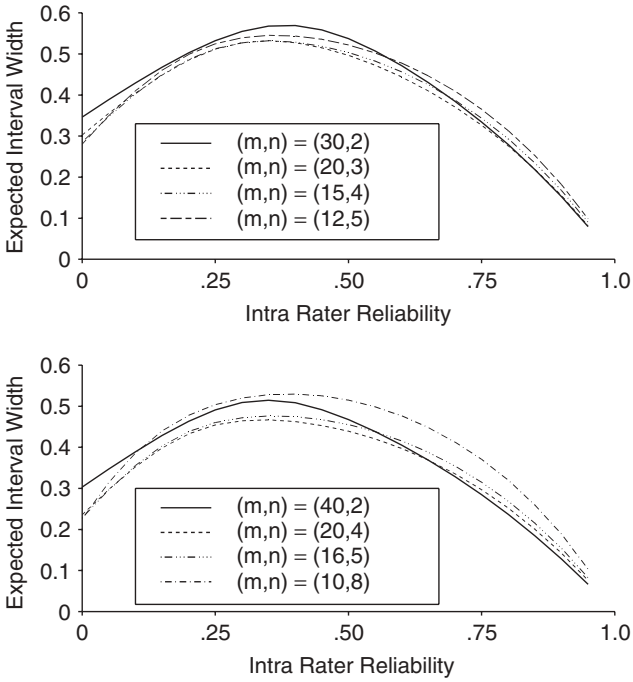


**Figure 2.** Expected width of the 95% confidence interval as a function of $\gamma$ for $m$ and $n$ values that correspond to $mn = 60$ and $mn = 80$.

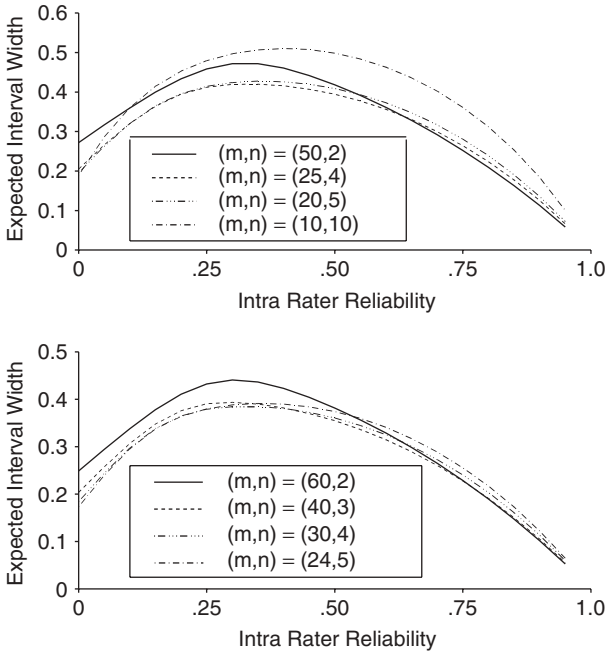


**Figure 3.** Expected width of the 95% confidence interval as a function of $\gamma$ for $m$ and $n$ values that correspond to $mn = 100$ and $mn = 120$.

that for high intrarater reliability coefficients (i.e., greater than 0.5), and a fixed total number of observations $mn$, using 2, 3, or at most 4 replicates per subjects is sufficient to obtain the most efficient intrarater reliability coefficient. Having more than four replicates is likely to lead to a loss of precision. For smaller $\gamma$ values, the recommendation is to use four or five replicates per subject. One would also note that if the "true" value of the intrarater reliability is smaller than 0.80, then its estimation will generally not be very precise.

**1.3.2 Blocking the Rater Factor**. If two raters or more are used in a completely randomized intrarater reliability experiment, the resulting coefficient will be inaccurate. In a completely randomized design, subjects and replicates are assigned randomly to different raters. Consequently, the rater effect will increase the experimental error, which thereby decreases the magnitude of the intrarater reliability coefficient.

This problem can be resolved by designing the experiment so that the rater effect can be measured and separated from the experimental error. A design that permits this method requires each rater to rate all subjects and provide the same number of replicates per subject. Under this design referred to as Randomized Block Design (RBD), the data is gathered by block (i.e., by rater in this case) with random rating within the block, and it would be organized as shown in Table 4, where $\gamma$ is the number of raters.

In Table 4, $y_{ijk}$ represents the $k^{\text{th}}$ replicate observation on subject $i$ provided by rater $j$.

The intrarater reliability coefficient $\hat{\gamma}$ under an RBD design is still defined by Equation (3) with the exception that the within-subject variation (WSV) and the between-subject variation (BSV) are defined as follows:

$$WSV = \frac{1}{mr}\sum_{i=1}^{m}\sum_{j=1}^{r} S_{ij}^2, \text{ and}$$

$$BSV = BSV_o - \frac{\text{WSV}}{\text{nr}}$$

$$\text{where } \mathrm{S}_{ij}^2 = \frac{1}{n-1}\sum_{k=1}^{n}(y_{ijk} - \overline{y}_{ij.})^2 \text{ and}$$

$$BSV_o = \frac{1}{m-1}\sum_{i=1}^{m}(\overline{y}_{i..} - \overline{y}_{...})^2; \qquad (10)$$

$\overline{y}_{ij}$ is the average of all $n$ scores rater $j$ assigned to subject $i$, $\overline{y}_{i..}$ is the average of

**Table 4. Intrarater Reliability Data on *m* Subjects with *r* Raters and *n* Replicates per Subject and per Rater**

| Subjects | Rater 1 | ⋯ | Rater *j* | ⋯ | Rater *r* |
|---|---|---|---|---|---|
| 1 | $y_{111}, \ldots y_{11n}$ | ⋯ | $y_{1j1}, \ldots y_{1jn}$ | ⋯ | $y_{1r1}, \ldots y_{1rn}$ |
| 2 | $y_{211}, \ldots y_{21n}$ | ⋯ | $y_{2j1}, \ldots y_{2jn}$ | ⋯ | $y_{2rl}, \ldots y_{2rn}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| *i* | $y_{ill}, \ldots y_{iln}$ | ⋯ | $y_{ijl}, \ldots y_{ijn}$ | ⋯ | $y_{irl}, \ldots y_{irn}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| *m* | $y_{mll}, \ldots y_{mln}$ | ⋯ | $y_{mjl}, \ldots y_{mjn}$ | ⋯ | $y_{mrl}, \ldots y_{mrn}$ |

$nr$ scores assigned to subject $i$, and $\bar{y}_{...}$ the overall mean scores. Note that the WSV is obtained by averaging the $mr$ sample variances calculated at the cell level. Equation (10) offers the advantage of removing any influence of inter-rater variation when calculating the intrarater reliability.

The number of replicates in an RBD design may vary by rater and by subject. We assumed it to be fixed in this section for the sake of simplicity. Although a single rater is sufficient to carry out an intrarater reliability experiment, the use of multiple raters may be recommended for burden reduction or for convenience.

The techniques and the inferential framework discussed in this section work well for continuous data, such as the cholesterol level, but they are not suitable for nominal data. In the next section, I present some techniques specifically developed for nominal data.

## 2 NOMINAL SCALE SCORE DATA

Although the ICC is effective for quantifying the reproducibility of continuous data, nominal data raise new statistical problems that warrant the use of alternative methods. Rating subjects on a nominal scale amounts to classifying them into one of $q$ possible response categories. The discrete nature of that data has the following two implications:

1. The notion of reproducibility is exact. A response category is reproduced when the initial and the replicate categories are identical, and unlike continuous data, nominal data are not subject to random measurement errors.
2. A rater may classify a subject on two occasions into the exact same category by pure chance with a probability that is non-negligible.

Table 5 shows the distribution of 100 individuals with identified pulmonary abnormalities examined by a medical student on two occasions. On both occasions, the medical student found the same 74 individuals with pulmonary abnormalities and the same 15 individuals without any abnormalities. However, the student disagreed with himself on 11 individuals. These data, which are a modification of the results reported by Mulrow et al. (10), shows how analysts may organize intrarater reliability data, and it is used later in this section for illustration purposes.

For intrarater reliability experiments based on two replicates per subject, analysts may organize the observations as shown in Table 6, where $m$ is the number of subjects rated, and $m_{kl}$ the number of subjects classified in category $k$ on the first occasion and in category $l$ on the second occasion. If the experiment uses three replicates per subject or more, then a more convenient way to organize the data is shown in Table 7 where $n$ is the number of replicates per subject, and $n_{ik}$ the number of times subject $i$ is classified into category $k$.

### 2.1 Intrarater Reliability: Single Rater and Two Replications

When ratings come from a simple experiment based on a single rater, two replicates per subject, and two categories such as described in Table 5, the kappa coefficient of Cohen (2) or an alternative kappa-like statistic may

**Table 5. Distribution of 100 Subjects with Respect to the Presence of Pulmonary Abnormalities Observed on two Occasions by a Medical Student**

| First Observation | Second Observation | | |
|---|---|---|---|
| | Present | Absent | Total |
| Present | 74 | 1 | 75 |
| Absent | 10 | 15 | 25 |
| Total | 84 | 16 | 100 |

be used to estimate the intrarater reliability coefficient. The medical student who generated Table 5 data could have obtained some of the 89 matches by pure chance because of the small number of response categories limited to two. Consequently, 89% will overestimate student's self-consistency. Cohen's (2) solution to this problem was a chance-corrected agreement measure $\hat{\gamma}_\kappa$, which is known in the literature as kappa, and it is defined as follows:

$$\hat{\gamma}_\kappa = \frac{p_a - p_e}{1 - p_e} \tag{11}$$

where for Table 5 data $p_a = (74 + 15)/100 = 0.89$ is the overall agreement probability, and $p_e = 0.75 \times 0.84 + 0.21 \times 0.16 = 0.6636$ is the chance-agreement probability. Consequently, the kappa coefficient that measures the medical student intrarater reliability is $\hat{\gamma} = 0.673$. According to the Landis and Koch (11) benchmark, a kappa value of this magnitude is deemed substantial.

In a more general setting with $m$ subjects, two replicates per subject, and an arbitrary number $q$ of response categories (see Table 6), the kappa coefficient of Cohen (2) is still defined by Equation (11), except the overall agreement probability $p_a$ and the chance-agreement probability $p_e$ that are respectively defined as follows:

$$p_a = \sum_{k=1}^{q} p_{kk}, \text{ and } p_e = \sum_{k=1}^{q} p_{k+} p_{+k} \tag{12}$$

where $p_{kk} = m_{kk}/m$, $P_{+k} = m_{+k}/m$, and $P_{k+} = m_{k+}/m$ The overall agreement probability is the proportion of subjects classified into the exact same category on both occasions (i.e., the diagonal of Table 6).

The kappa coefficient will at times yield unduly low values when the ratings suggest high reproducibility. Cicchetti and Feinstein (12), as well as Feinstein and Cicchetti (13) have studied these unexpected results known in the literature as the kappa paradoxes. Several alternative more paradox-resistant coefficients are discussed by Brennan and Prediger (14). A Brennan-Prediger alternative denoted by $\hat{\gamma}_{GI}$, which is often referred to as the G-Index (GI) and should be considered by practitioners, is defined as follows:

$$\hat{\gamma}_{GI} = \frac{p_a - 1/q}{1 - 1/q}. \tag{13}$$

Applied to Table 5 data, the Brennan-Prediger coefficient becomes $\hat{\gamma}_{\mathrm{GI}} = (0.89 -$

**Table 6. Distribution of *m* Subjects by Response Category and Replication Number.**

| First-Replication Category | Second-Replication Response Category | | | | | |
|---|---|---|---|---|---|---|
| | 1 | ⋯ | $k$ | ⋯ | $q$ | *Total* |
| *1* | $m_{11}$ | ⋯ | $m_{1k}$ | ⋯ | $m_{1q}$ | $m_{+}$ |
| ⋮ | ⋮ | ⋱ | ⋮ | ⋮ | ⋮ | ⋮ |
| $k$ | $m_{k1}$ | ⋯ | $m_{kk}$ | ⋯ | $m_{kq}$ | $m_{k+}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋱ | ⋮ | ⋮ |
| $q$ | $m_{q1}$ | ⋯ | $m_{qk}$ | ⋯ | $m_{qq}$ | $m_{q+}$ |
| *Total* | $m_{+1}$ | ⋯ | $m_{+k}$ | ⋯ | $m_{+k}$ | $m$ |

$0.5)/(1-0.5) = 0.78$, which is slightly higher than the kappa. Aickin (15) presents an interesting discussion about kappa-like Intrarater reliability coefficients and suggests the use of his $\alpha$ coefficient. The $\alpha$ coefficient is based on a sound theory and uses the maximum likelihood estimation of some of its components obtained with a computationally intensive iterative algorithm. Gwet (16) proposed the $AC_1$ statistic as a simple way to resolve the kappa paradoxes. The $AC_1$ coefficient is defined as follows:

$$\hat{\gamma}_{AC_1} = \frac{p_a - p_e^{(1)}}{1 - p_e^{(1)}} \tag{14}$$

where $p_a$ is defined by Equation (12), and the chance-agreement probability is as follows:

$$p_e^{(1)} = \frac{1}{q-1}\sum_{k=1}^{q} p_k(1-p_k), \text{ where}$$
$$p_k = (p_{k+} + p_{+k})/2 \tag{15}$$

For Table 5 data, $p_1 = (0.75 + 0.84)/2 = 0/795$, $p_2 = (0.25 + 0.16)/2 = 0.205$, and $p_2 = 1-p_1$. Consequently, Gwet's chance-agreement probability is $P_e^{(1)} = 2 \times 0.795 \times 0.205 = 0.32595$. The $AC_1$ statistic is then given by $\hat{\gamma}_{AC_1} = (0.89 - 0.32595)/(1 - 0.32595) \cong 0.84$, which is more in line with the observed extent of reproducibility. Gwet (16) extensively discusses the statistical properties of the $AC_1$ statistic as well as the origins of the kappa paradoxes.

### 2.2 Intrarater Reliability: Single Rater and Multiple Replications

Using more than two replicates per subject can improve the precision of an intrarater reliability coefficient. The techniques discussed in this section generalize those of the previous section, and they are suitable for analyzing Table 7 data that involve three replicates or more per subject.

All kappa-like statistics presented in the previous section can still be used with Table 7 data. However, the overall agreement probability $p_a$ is defined as the probability that two replicates random by chosen from the $n$ replicates associated with a randomly selected subject, are identical. More formally $p_a$ is defined as follows:

$$p_a = \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{k=1}^{q}\frac{n_{ik}(n_{ik}-1)}{n(n-1)}\right) \tag{16}$$

Concerning the calculation of chance-agreement probability, several versions have been proposed in the literature, most of which are discussed by Conger (17) in the context of inter-rater reliability, rather than in the context of intrarater reliability. Fleiss (18) suggested that chance-agreement probability be estimated as follows:

$$p_e^{(F)} = \sum_{k=1}^{q} p_k^2 \text{ where } p_k = \frac{1}{m}\sum_{i=1}^{m}\frac{n_{ik}}{n} \tag{17}$$

Note that $p_k$ represents the relative number of times that a subject is classified into category $k$. Fleiss' generalized kappa is then given by:

$$\hat{\gamma}_F = (p_a - p_e^{(F)})/(1 - p_e^{(F)}).$$

Conger (17) criticized Fleiss' generalized kappa statistic for not reducing to Cohen's kappa when the number of replicates is

**Table 7. Frequency Distribution of *mn* Observations by Subject and Response Category.**

| | Response Category | | | | | |
|---|---|---|---|---|---|---|
| Subject | 1 | ⋯ | *k* | ⋯ | *q* | *Total* |
| *1* | $n_{11}$ | ⋯ | $n_{1k}$ | ⋯ | $n_{1q}$ | *n* |
| ⋮ | ⋮ | ⋱ | ⋮ | ⋮ | ⋮ | ⋮ |
| *i* | $n_{i1}$ | ⋯ | $n_{ik}$ | ⋯ | $n_{iq}$ | *n* |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋱ | ⋮ | ⋮ |
| *m* | $n_{n1}$ | ⋯ | $n_{nk}$ | ⋯ | $n_{nq}$ | *n* |
| *Total* | $n_{+1}$ | ⋯ | $n_{+k}$ | ⋯ | $n_{+q}$ | *mn* |

limited to two and proposed the following chance-agreement probability:

$$p_e^{(C)} = \sum_{k=1}^{q} p_k^2 - \sum_{k=1}^{q} s_k^2/n \qquad (18)$$

where $s_k^2$ is the sample variance

$$s_k^2 = \frac{1}{n-1}\sum_{j=1}^{n}(\theta_{jk} - \overline{\theta}_{\cdot k})^2 \qquad (19)$$

where $\theta_{jk} = m_{jk}/m$ is the percent of subjects classified into category $k$ on the $j$th occasion, and $\overline{\theta}_{\cdot k}$ is the average of these $n$ values. To compute the variances $s_k^2, k = 1, \cdots, q$, it could be useful to organize ratings as in Table 8.

Both the Fleiss and Conger versions of kappa are vulnerable to the paradox problem previously discussed, and they yield reasonable intrarater reliability coefficients only when $p_k$, the propensity for classification in category $k$, remain fairly constant from category to category.

The generalized version of the $AC_1$ statistic of Gwet (16) is a more paradox-resistant alternative to kappa, and it is based on Equation (14) with the exception that the chance-agreement probability is defined as follows:

$$p_e^{(1)} = \frac{1}{q-1}\sum_{k=1}^{q} p_k(1-p_k) \qquad (20)$$

where $p_k$ is defined as in Equation (17).

The situation where intrarater reliability data are collected by multiple raters may occur in practice, and it should be dealt with using special methods that eliminate the impact of inter-rater variation. The general approach consists of averaging various probabilities calculated independently for each rater as previously discussed.

### 2.3 Statistical Inference

For an intrarater reliability coefficient to be useful, it must be computed with an acceptable level of precision; this notion can be defined and measured only within a formal framework for statistical inference. This section gives an overview of the main inferential approaches proposed in the literature and provides references for more inquiries.

Several authors have proposed frameworks for statistical inference based on various theoretical models. Kraemer et al. (19) review many models that have been proposed in the literature. Kraemer (20) proposed a model under which the kappa coefficient can be interpreted as an intraclass correlation. Donner and Eliasziw (21), Donner and Klar (22), and Aickin (15) have proposed different models that may be useful in different contexts. This model-based approach poses two important problems for practitioners. The first problem stems from the difficulty of knowing which model is most appropriate for a particular situation. The second problem is the dependency of inferential procedures on the validity of the hypothesized model. Fortunately, a different approach to inference based on finite population sampling and widely used in the social sciences can resolve both problems.

**Table 8. Frequency Distribution of *mn* Observations by Replicate Number and Response Category.**

| | Response Category | | | | | |
|---|---|---|---|---|---|---|
| Replication | 1 | ⋯ | $k$ | ⋯ | $q$ | *Total* |
| *1* | $m_{11}$ | ⋯ | $m_{1k}$ | ⋯ | $m_{1q}$ | $m$ |
| ⋮ | ⋮ | ⋱ | ⋮ | ⋮ | ⋮ | ⋮ |
| $j$ | $m_{j1}$ | ⋯ | $m_{jk}$ | ⋯ | $m_{iq}$ | $m$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋱ | ⋮ | ⋮ |
| $n$ | $m_{n1}$ | ⋯ | $m_{nk}$ | ⋯ | $m_{nq}$ | $m$ |
| *Total* | $m_{+1}$ | ⋯ | $m_{+k}$ | ⋯ | $m_{+q}$ | $mn$ |

The randomization approach or design-based inference is a statistical inference framework in which the underlying random process is created by the random selection of $m$ subjects out of a predefined finite population of $M$ subjects of interest. This approach is described in textbooks such as Kish (23) and Cochran (24), and it has been used extensively in the context of inter-rater reliability assessment by Gwet (16,25). The variances of many intrarater reliability coefficients presented in the second section can be found in Gwet (16,25).

The first two sections present various approaches for evaluating the reproducibility of continuous and nominal data. These approaches are not recommended for ordinal or interval data, although ordinal clinical measurements such as the presence (no, possible, probable, definite) of a health condition as read on a radiograph, are commonplace. The objective of the next section is to present a generalization of kappa susceptible for use with ordinal and interval data.

## 3 ORDINAL AND INTERVAL SCORE DATA

Berry and Mielke (26) as well as Janson and Olsson (27,28) have generalized the kappa statistic to handle ordinal and interval data. In addition to being applicable to ordinal and interval data, these extensions can analyze multivariate data of subjects rated on multiple characteristics. Although Berry and Mielke (26) deserve credit for introducing the notions of vector score and Euclidean distance behind these extensions, Janson and Olsson (27) improved and expanded them substantially.

Let us consider a simple intrarater reliability study in which a rater must rate all five subjects ($m = 5$) on two occasions ($n = 2$) on a three-level nominal scale ($q = 3$). If the rater classifies subject 1 into category 2 on the first occasion, then the corresponding score can be represented as a vector $\mathbf{a}_{11} = (0,1,0)$, with the second position of digit "1" indicating the category number where the subject is classified. The vector score associated with the classification of subject 1 into category 3 on the second occasion is $\mathbf{a}_{12} = (0,0,1)$. The squared Euclidean distance between $\mathbf{a}_{11}$ and $\mathbf{a}_{12}$ is obtained by summing all three squared differences between the elements of both vectors and is given by:

$$d^2(\mathbf{a}_{11}, \mathbf{a}_{12}) = (0-0)^2 + (0-1)^2 + (1-0)^2 = 2$$

Following Janon and Olsson (26), Cohen's kappa coefficient can re-expressed as follows:

$$\hat{\gamma}_{JO} = 1 - \frac{\frac{1}{5}\sum_{i=1}^{5} d^2(\mathbf{a}_{i1}, \mathbf{a}_{i2})}{\frac{1}{5^2}\sum_{i=1}^{5}\sum_{j=1}^{5} d^2(\mathbf{a}_{i1}, \mathbf{a}_{j2})} \qquad (21)$$

The kappa coefficient as written in Equation (21) depends solely on several distance functions. Its generalization relies on the distance function's ability to handle ordinal and interval data. If the scoring is carried out on a three-level ordinal scale, then each score will represent a single rank instead of three-dimensional vector of 0s and 1s. If the categories in Table 6 are ordinal, then Equation (21) can be adapted to that data and yield the following more efficient kappa coefficient:

$$\hat{\gamma}_{JO} = 1 - \frac{\sum_{k=1}^{q}\sum_{i=1}^{q} p_{kl}(k-l)^2}{\sum_{k=1}^{q}\sum_{l=1}^{q} p_{k+}p_{+l}(k-l)^2} \qquad (22)$$

To illustrate the use of kappa with ordinal scores, let us consider Table 9 data, which are a modification of the initial chest radiograph data that Albaum et al. (29) analyzed. A radiologist has examined 100 initial chest radiographs on two occasions to determine the presence of a radiographic pulmonary infiltrate. The four levels of the measurement scale for this experiment are "No," "Possible," "Probable," and "Definite." Because classifications of radiographs into the "Probable" and "Definite" categories agree more often than those in the "No" and "Definite" categories, the use of classic kappa of Cohen (2) will dramatically underestimate the intrarater reliability.

Cohen's kappa for Table 9 is given by $\hat{\gamma}_k = (0.57 - 0.3151)/(1 - 0.3151) \approx 0.37$. The generalized version of kappa based on Equation

**Table 9. Distribution of 100 Subjects by Presence of Radiographic Pulmonary inFiltrate and Assessment Time**

| Radiographic Assessment | Radiographic Assessment in Time 2 | | | | |
|---|---|---|---|---|---|
| in Time 1 | No | Possible | Probable | Definite | TOTAL |
| No | 6 | 7 | 2 | 1 | 16 |
| Possible | 2 | 7 | 6 | 2 | 17 |
| Probable | 2 | 4 | 7 | 5 | 18 |
| Definite | 1 | 4 | 7 | 37 | 49 |
| TOTAL | 11 | 22 | 22 | 45 | 100 |

(22) yields an intrarater reliability coefficient of $\hat{\gamma}_{JO} = 1 - 0.89/2.41 = 0.63$. This generalized version of kappa yields an intrarater reliability coefficient substantially higher and accounts for partial agreement in a more effective way.

## 4 CONCLUDING REMARKS

This article introduces the notion of intrarater reliability for continuous, nominal, ordinal, as well as interval data. Although the intraclass correlation coefficient is the measure of choice for continuous data, kappa and kappa-like measures defined by Equations (11), (13), (14), and (18)–(20) are generally recommended for nominal data. The extension of kappa to ordinal data is more efficient than classic kappa when the data is ordinal, and it is an important addition to the kappa literature.

The literature on inter-rater reliability is far more extensive than that on intrarater reliability, particularly for discrete data, which is explained partially by the tendency researchers have to underestimate the importance of data reproducibility. Although many techniques were developed to measure inter-rater reliability, very few specifically address the problem of intrarater reliability. In this article, we have adapted some inter-rater reliability estimation procedures so they can be used for computing intrarater reliability coefficients. Unlike inter-rater reliability experiments that involve multiple raters, multiple subjects, and a single replicate per subject, intrarater reliability experiments typically involve a single rater and several replicates per subject. Consequently inter-rater reliability methods have been modified by considering the replicates as ratings from different independent raters.

Several authors, such as Fleiss and Cohen (30), Kraemer (20), and others, have attempted to interpret kappa as well as other kappa-like reliability measures as a form of intraclass correlation under certain conditions. The main justification for this effort stems from the need to link kappa to a population parameter and to create a framework for statistical inference. So far, no clear-cut theory can establish such a link in a broad setting. The connection of kappa to the intraclass correlation is unnecessary to have a good statistical inference framework. A satisfactory solution to this problem is the use of the finite population inference framework discussed in Gwet (16, 25).

## REFERENCES


1. S. M. Grundy, *Second report of the expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel II)*. 1993; National Institutes of Health, NIH Publication No. 93-3095.
2. J. Cohen, A coefficient of agreement for nominal scales. *Educat. Psychol. Measurem*. 1960; **20**: 37–46.
3. R. L. Ebel, Estimation of the reliability of ratings. *Psychometrika* 1951; **16**: 407–424.
4. J. J. Barko, The intraclass correlation coefficient as a measure of reliability. *Psychol. Reports* 1966; **19**: 3–11.
5. P. E. Shrout and J. L. Fleiss, Intraclass correlations: uses in assessing rater reliability. *Psychol. Bull.* 1979; **86**: 420–428.
6. J. M. Lachin, The role of measurement reliability in clinical trials. *Clin. Trials* 2004; **1**: 553–566.
7. National Health and Nutrition Examination Survey. 2005.

8. K. O. McGraw and S. P. Wong, Forming inferences about some intraclass correlation coefficients. *Psychol. Methods* 1996; **1**: 30–46.
9. B. Giraudeau and J. Y. Mary, Planning a reproducibility study: how many subjects and how many replicates per subject for an expected width of the 95 per cent confidence interval of the intraclass correlation coefficient. *Stats. Med.* 2001; **20**: 3205–3214.
10. C. D. Mulrow, B. L. Dolmatch, E. R. Delong, J. R. Feussner, M. C. Benyunes, J. L. Dietz, S. K. Lucas, E. D. Pisano, L. P. Svetkey, B. D. Volpp, R. E. Ware, and F. A. Neelon, Observer variability in the pulmonary examination. *J. Gen. Intern Med.* 2007; **1**: 364–367.
11. J. R. Landis and G. G. Koch, The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**: 159–174.
12. D. V. Cicchetti and A. R. Feinstein, High agreement but low kappa: II. Resolving the paradoxes. *J. Clin. Epidemiol.* 1990; **43**: 551–558.
13. A. R. Feinstein and D. V. Cicchetti, High agreement but low kappa: I. The problems of two paradoxes. *J. Clin. Epidemiol.* 1990; **43**: 543–549.
14. Brennan, RL, and Prediger, DJ. Coefficient kappa: some uses, misuses, and alternatives. *Educat. Psychol Measurem*. 1981; **41**: 687–699.
15. M. Aickin, Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's kappa. *Biometrics* 1990; **46**: 293–302.
16. K. L. Gwet, Computing inter-rater reliability and its variance in the presence of high agreement. *Br. J. Mathemat. Stat. Psychol.* 2008. In press.
17. A. J. Conger, Integration and generalization of kappas for multiple raters. *Psychol. Bull.* 1980; **88**: 322–328.
18. J. L. Fleiss, Measuring nominal scale agreement among many raters. *Psychol. Bull.* 1971; **76**: 378–382.
19. H. C. Kraemer, V. S. Periyakoil, and A. Noda, Kappa coefficients in medical research. *Stats. Med.* 2002; **21**: 2109–2129.
20. H. C. Kraemer, Ramifications of a population model for $\kappa$ as a coefficient of reliability. *Psychometrika* 1979; **44**: 461–472.
21. A. Donner, M. A. Eliasziw, A hierarchical approach to inferences concerning interobserver agreement for multinomial data. *Stats. Med.* 1997; **16**: 1097–1106.
22. Donner, A, Klar, N. The statistical analysis of kappa statistics in multiple samples. *Journal of Clinical Epidemiology.* 1996; **49**(9): 1053–1058.
23. K. Kish, *Survey Sampling.* New York: Wiley, 1965.
24. W. G. Cochran, *Sampling Techniques*, 3rd ed. New York: Wiley, 1977.
25. K. L. Gwet, Variance estimation of nominal-scale inter-rater reliability with random selection of raters. *Psychometrika.* 2008. In press.
26. K. J. Berry and P. W. Mielke Jr., A generalization of Cohen's Kappa agreement measure to interval measurement and multiple raters. *Educat. Psychol. Measurem.* 1988; **48**: 921–933.
27. H. Janson and U. Olsson, A measure of agreement for interval or nominal multivariate observations. *Educat. Psychol. Measurem.* 2001; **61**: 277–289.
28. H. Janson and U. Olsson, A measure of agreement for interval or nominal multivariate observations by different sets of judges. *Educat. Psychol. Measurem.* 2004; **64**: 62–70.
29. M. N. Albaum, L. C. Hill, M. Murphy, Y. H. Li, C. R. Fuhrman, C. A. Britton, W. N. Kapoor, and M. J. Fine, PORT Investigators. Interobserver reliability of the chest radiograph in community-acquired pneumonia. *CHEST* 1996; **110**: 343–350.
30. J. L. Fleiss and J. Cohen, The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educat. Psychol. Measurem.* 1973; **33**: 613–619.

## FURTHER READING

D. C. Montgomery, *Design and Analysis of Experiments*. New York: John Wiley & Sons, 2004.

D. T. Haley, Using a New Inter-rater Reliability Statistic, 2007. Available: http://computing-reports.open.ac.uk/index.php/2007/200716.

ANOVA Using MS Excel. Available: http://highered bcs.wiley.com/legacy/college/mann/0471755303/excel_manual/ch12.pdf.

## CROSS-REFERENCES

Inter-Rater Reliability
Intraclass Correlation Coefficient
Kappa Statistic
Weighted Kappa
Analysis of Variance (ANOVA)