**Postprint.** Warrens, M. J. (2012). A family of multi-rater kappas that can always be increased and decreased by combining categories. *Statistical Methodology, 9,* 330-340.

http://dx.doi.org/10.1016/j.stamet.2011.08.008

**Author.** Matthijs J. Warrens
Institute of Psychology
Unit Methodology and Statistics
Leiden University
P.O. Box 9555, 2300 RB Leiden
The Netherlands
E-mail: warrens@fsw.leidenuniv.nl

# A family of multi-rater kappas that can always be increased and decreased by combining categories

Matthijs J. Warrens, Leiden University

**Abstract.** Cohen's kappa is a popular descriptive statistic for measuring agreement between two raters on a nominal scale. Various authors have generalized Cohen's kappa to the case of $m \geq 2$ raters. We consider a family of multi-rater kappas that are based on the concept of $g$-agreement ($g = 2, 3, \ldots, m$), which refers to the situation in which it is decided that there is agreement if $g$ out of $m$ raters assign an object to the same category. For the family of multi-rater kappas we prove the following existence theorem: In the case of three or more categories there exists for each multi-rater kappa $\kappa(m, g)$ two categories such that, when combined, the $\kappa(m, g)$ value increases. In addition, there exist two categories such that, when combined, the $\kappa(m, g)$ value decreases.

**Key words.** Inter-rater reliability; Cohen's kappa; Schouten-type inequality; Hubert's kappa; Mielke, Berry and Johnston's kappa.

**Acknowledgment.** The author thanks two anonymous reviewers for their helpful comments and valuable suggestions on an earlier version of this article.

# 1 Introduction

In various fields of science, including behavioral sciences and the biomedical field, it is frequently required that a group of subjects is classified into a set of mutually exclusive (nominal) categories, such as psychodiagnostic classifications (Fleiss 1981; Zwick 1988). Because there is often no golden standard, researchers require that the classification task is performed by multiple raters. The agreement of the ratings is then taken as an indicator of the quality of the category definitions and the raters' ability to apply them. The most popular measure for summarizing agreement between two raters is Cohen's (1960) kappa, denoted by $\kappa$ (Kraemer, 1979; Brennan & Prediger, 1981; Schouten, 1986; Zwick, 1988; Kraemer, Periyakoil & Noda, 2002; Hsu & Field, 2003; Warrens 2008a, 2010a,b). The value of Cohen's $\kappa$ is 1 when perfect agreement between the two raters occurs, 0 when agreement is equal to that expected under independence, and negative when agreement is less than expected by chance.

The number of categories used in various classification schemes varies from the minimum number of two to five in many practical applications. Ratings are usually summarized in a square agreement table of size $k \times k$, where $k$ is the number of categories. It is sometimes desirable to combine some of the categories (Warrens, 2010c), for example, when two categories are easily confused (Schouten, 1986), and then calculate the $\kappa$ value of the collapsed $(k-1) \times (k-1)$ agreement table. Schouten (1986) presented a necessary and sufficient condition for the $\kappa$ to increase when two categories are combined, and showed that it depends on which categories are combined whether the value of $\kappa$ increases or decreases. Using the condition presented in Schouten (1986), Warrens (2010c) showed that for a nontrivial table with $k \geq 3$ categories there exist two categories such that, when the two are merged, the $\kappa$ value of the collapsed $(k-1) \times (k-1)$ agreement table is higher than the original $\kappa$ value, that is, the $\kappa$ value increases, and that there exist two categories such that, when combined, the $\kappa$ value decreases.

The popularity of Cohen's $\kappa$ has led to the development of many extensions (Nelson & Pepe, 2000; Kraemer et al., 2002) including kappas for groups of raters (Vanbelle & Albert, 2009a,b) and weighted kappas for ordinal categories (Vanbelle & Albert, 2009c; Warrens, 2011a,b). Cohen's $\kappa$ has also been extended to the important case of multiple raters (Hubert, 1977; Conger 1980; Von Eye & Mun, 2006; Mielke, Berry & Johnston, 2007, 2008; Warrens, 2010d). With multiple raters there are several views in the literature on how to define agreement (Hubert, 1977; Conger, 1980; Popping, 2010). For example, simultaneous agreement (or $m$-agreement) refers to the situation in which it is decided that there is only agreement if all $m$ raters assign an object to the same category (see for example Warrens, 2009). Hubert (1977, p. 296) refers to this type of agreement as DeMoivre's definition of agreement. In contrast, pairwise agreement (or 2-agreement)

refers to the situation in which it is decided that there is already agreement if only two raters categorize an object consistently. Conger (1980) argued that agreement among raters can actually be considered to be an arbitrary choice along a continuum ranging from $m$-agreement to 2-agreement. $g$-agreement with $g \in \{2, 3, \ldots, m\}$ refers to the situation in which it is decided that there is agreement if $g$ out of $m$ raters assign an object to the same category (Conger, 1980).

In this paper we consider a family of multi-rater kappas for nominal categories that are based on the concept of $g$-agreement ($g \in \{2, 3, \ldots, m\}$). Various multi-rater kappas proposed in the literature belong to this family. Given $m \geq 2$ raters we can formulate $m - 1$ multi-rater kappas, one based on 2-agreement, one based on 3-agreement, and so on, and one based on $m$-agreement. The kappa statistic for $m$ raters that is based on $g$-agreement is denoted by $\kappa(m, g)$. We prove the following existence theorem for the family of multi-rater kappas: In the case of three or more categories there exist for each $\kappa(m, g)$ two categories such that, when combined, the $\kappa(m, g)$ value increases. In addition, there exist two categories such that, when combined, the $\kappa(m, g)$ value decreases. The paper is organized as follows. The family of multi-rater kappas is introduced in the next section. In Section 3 we present a sufficient and necessary condition for $\kappa(m, g)$ to increase when two categories are combined. In Section 4 we present the existence theorem. Section 5 contains numerical illustrations of the existence theorem. Section 6 contains a conclusion.

## 2 A family of multi-rater kappas

In this section we consider a family of multi-rater kappas. We first introduce Cohen's (1960) $\kappa$.

Suppose that two raters $r_1$ and $r_2$ each independently classify the same set of $w \in \mathbb{N}_{\geq 1}$ objects (individuals, observations) into $k \in \mathbb{N}_{\geq 2}$ nominal (unordered) categories indexed by $c_1, c_2 \in \{1, 2, \ldots, k\}$ that are defined in advance. Let

$$\mathbf{F} = \{f\left(\begin{smallmatrix} r_1 & r_2 \\ c_1 & c_2 \end{smallmatrix}\right)\}$$

be a 2-way contingency table of size $k \times k$ where the element $f\left(\begin{smallmatrix} r_1 & r_2 \\ c_1 & c_2 \end{smallmatrix}\right)$ indicates the number of objects placed in category $c_1$ by rater $r_1$ and in category $c_2$ by rater $r_2$. If we divide the elements of $\mathbf{F}$ by the total number of objects $w$ we obtain the table

$$\mathbf{P} = \{p\left(\begin{smallmatrix} r_1 & r_2 \\ c_1 & c_2 \end{smallmatrix}\right)\}$$

with relative frequencies $p\left(\begin{smallmatrix} r_1 & r_2 \\ c_1 & c_2 \end{smallmatrix}\right) = w^{-1} f\left(\begin{smallmatrix} r_1 & r_2 \\ c_1 & c_2 \end{smallmatrix}\right)$. For notational convenience we will work with table $\mathbf{P}$ instead of $\mathbf{F}$. Table $\mathbf{P}$ contains the 2-agreement between the raters and is therefore also called an agreement table.

The elements of $\mathbf{P}$ add up to 1. Row and column totals

$$p_{c_1}^{r_1} = \sum_{i=1}^{k} p\left(\begin{smallmatrix} r_1 & r_2 \\ c_1 & c_i \end{smallmatrix}\right) \qquad \text{and} \qquad p_{c_2}^{r_2} = \sum_{i=1}^{k} p\left(\begin{smallmatrix} r_1 & r_2 \\ c_i & c_2 \end{smallmatrix}\right)$$

are the marginal totals of $\mathbf{P}$. The marginal total $p_{c_1}^{r_1}$ denotes the proportion of objects assigned to category $c$, by rater $r_1$, and likewise $p_{c_2}^{r_2}$. An example of $\mathbf{P}$ for five categories is presented in Table 1. This $5 \times 5$ table contains the relative frequencies of data presented in Landis and Koch (1977) and originally reported by Holmquist et al. (1967) (see also, Agresti, 1990, p. 367). Two pathologists (pathologists A and B in Landis & Koch, 1977, p. 365) classified each of 118 slides in terms of carcinoma in situ of the uterine cervix, based on the most involved lesion, using the categories 1) Negative, 2) Atypical squamous hyperplasia, 3) Carcinoma in situ, 4) Squamous carcinoma with early stromal invasion, and 5) Invasive carcinoma.

*Insert Table 1 about here.*

Cohen's $\kappa$ for raters $r_1$ and $r_2$ is defined as

$$\kappa = \frac{O - E}{1 - E} = \frac{\sum_{i=1}^{k} \left( p\left(\begin{smallmatrix} r_1 & r_2 \\ c_i & c_i \end{smallmatrix}\right) - p_{c_i}^{r_1} p_{c_i}^{r_2} \right)}{1 - \sum_{i=1}^{k} p_{c_i}^{r_1} p_{c_i}^{r_2}}$$

where

$$O = \sum_{i=1}^{k} p\left(\begin{smallmatrix} r_1 & r_2 \\ c_i & c_i \end{smallmatrix}\right) \qquad \text{and} \qquad E = \sum_{i=1}^{k} p_{c_i}^{r_1} p_{c_i}^{r_2}$$

are called the proportions of observed and expected agreement. Standard errors for $\kappa$ can be found in Fleiss, Cohen and Everitt (1969). For the data in Table 1 we have

$$O = .186 + .059 + .305 + .059 + .025 = .636,$$

$$\begin{aligned} E &= (.220)(.229) + (.220)(.102) + (.322)(.585) + (.186)(.059) + (.051)(.025) \\ &= .273, \end{aligned}$$

and

$$\kappa = \frac{.636 - .273}{1 - .273} = .498.$$

There are several ways to extend Cohen's $\kappa$ for two raters to the case of $m \geq 2$ raters. Here we consider a multi-rater kappa that incorporates the concept of $g$-agreement where $g \in \{2, 3, \ldots, m\}$. Let $p\left(\begin{smallmatrix} r_1 & \cdots & r_g \\ c_1 & \cdots & c_g \end{smallmatrix}\right)$ where

$r_j \in \{1, 2, \ldots, m\}$ and $c_i \in \{1, 2, \ldots, k\}$ denote the proportion of objects placed in category $c_1$ by the rater $r_1$, in category $c_2$ by rater $r_2$, and so on, and in category $c_g$ by rater $r_g$. Furthermore, let $p_{c_i}^{r_j}$ denote the proportion of objects assigned to category $c_i$ by rater $r_j$. The quantities $p\left(\begin{smallmatrix} r_1 & \cdots & r_g \\ c_1 & \cdots & c_g \end{smallmatrix}\right)$ can be seen as the elements of a $g$-dimensional table or $g$-agreement table $\mathbf{P}^{(g)}$. An example of $\mathbf{P}^{(3)}$ for five categories is presented in Table 2. This $5 \times 5 \times 5$ table contains the relative frequencies of classifications of 118 slides by three pathologists (pathologists A, B and C in Landis & Koch, 1977, p. 365). By summing the elements of the table $\mathbf{P}^{(g)}$ over $g - 1$ of the $g$ dimensions we obtain the marginal totals $p_{c_i}^{r_j}$ for rater $r_j$. For example, if we add the five slices in Table 2, that is, if we sum all elements over the direction corresponding to pathologist 3, we obtain Table 1, the $5 \times 5$ cross-classification between pathologists 1 and 2. The other two collapsed tables corresponding to the 3-dimensional table in Table 2 are the two $5 \times 5$ tables in Table 3.

*Insert Tables 2 and 3 about here.*

A $g$-agreement kappa for $m \geq 2$ raters can be defined as

$$\begin{aligned}\kappa(m, g) &= \frac{O(m, g) - E(m, g)}{\binom{m}{g} - E(m, g)} \\ &= \frac{\sum_{i=1}^{k} \sum_{r_1 < \cdots < r_g}^{m} \left( p\left(\begin{smallmatrix} r_1 & \cdots & r_g \\ c_i & \cdots & c_i \end{smallmatrix}\right) - \prod_{j=1}^{g} p_{c_i}^{r_j} \right)}{\binom{m}{g} - \sum_{i=1}^{k} \sum_{r_1 < \cdots < r_g}^{m} \prod_{j=1}^{g} p_{c_i}^{r_j}}.\end{aligned}$$

where

$$\begin{aligned}O(m, g) &= \sum_{i=1}^{k} \sum_{r_1 < \cdots < r_g}^{m} p\left(\begin{smallmatrix} r_1 & \cdots & r_g \\ c_i & \cdots & c_i \end{smallmatrix}\right) \\ E(m, g) &= \sum_{i=1}^{k} \sum_{r_1 < \cdots < r_g}^{m} \prod_{j=1}^{g} p_{c_i}^{r_j}\end{aligned}$$

are the observed and expected $g$-agreement for $m$ raters and

$$\binom{m}{g} = \frac{m!}{g!(m - g)!}.$$

The binomial coefficient $\binom{m}{g}$ is the maximum value of $O(m, g)$. The value of $\kappa(m, g)$ is 1 when perfect agreement between $m$ raters occurs, and 0 when $O(m, g) = E(m, g)$. Standard errors for $\kappa(m, g)$ can be found in Hubert (1977).

We consider some special cases of $\kappa(m,g)$. For $m = g = 2$ we have Cohen's $\kappa = \kappa(2,2)$. For $g = 2$ we obtain

$$\kappa(m,2) = \frac{\sum_{i=1}^{k} \sum_{r_1<r_2}^{m} \left(p\left(\begin{smallmatrix} r_1 & r_2 \\ c_i & c_i \end{smallmatrix}\right) - p_{c_i}^{r_1} p_{c_i}^{r_2}\right)}{\binom{m}{2} - \sum_{i=1}^{k} \sum_{r_1<r_2}^{m} p_{c_i}^{r_1} p_{c_i}^{r_2}}.$$

Coefficient $\kappa(m,2)$ is based on the 2-agreement between the raters. This descriptive statistic was first considered in Hubert (1977, p. 296, 297) and has been independently proposed by Conger (1980). The measure is also discussed in Davies and Fleiss (1982), Popping (1983), Heuvelmans and Sanders (1993) and Warrens (2008b, 2010d). Furthermore, coefficient $\kappa(m,2)$ is a special case of the descriptive statistics proposed in Berry and Mielke (1988) and Janson and Olsson (2001). For the data in Tables 1 and 3 we have

$$\begin{aligned} O(3,2) &= (.186 + .059 + .305 + .059 + .025) + (.161 + .144 + .169 + .042 + .017) \\ &\quad + (.169 + .059 + .271 + .025 + .017) \\ &= .636 + .534 + .542 = 1.712, \end{aligned}$$

$$\begin{aligned} E(3,2) &= (.220)(.229) + (.220)(.102) + (.322)(.585) + (.186)(.059) + (.051)(.025) \\ &\quad + (.220)(.263) + (.220)(.356) + (.322)(.314) + (.186)(.051) + (.051)(.017) \\ &\quad + (.229)(.263) + (.102)(.356) + (.585)(.314) + (.059)(.051) + (.025)(.017) \\ &= .273 + .248 + .283 = .804, \end{aligned}$$

and

$$\kappa(3,2) = \frac{1.712 - .804}{3 - .804} = .413.$$

For $g = m$ we obtain

$$\kappa(m,m) = \frac{\sum_{i=1}^{k} \left(p\left(\begin{smallmatrix} r_1 & \cdots & r_m \\ c_i & \cdots & c_i \end{smallmatrix}\right) - \prod_{j=1}^{m} p_{c_i}^{r_j}\right)}{1 - \sum_{i=1}^{k} \prod_{j=1}^{m} p_{c_i}^{r_j}}.$$

Coefficient $\kappa(m,m)$ is based on the $m$-agreement between the raters, and is thus a coefficient of simultaneous agreement (Hubert, 1977; Popping, 2010). Coefficient $\kappa(m,m)$ is the unweighted kappa proposed in Von Eye and Mun (2006, p. 22), Mielke et al. (2007, 2008) and Berry, Johnston and Mielke (2008). For the data in Table 2 we have

$$O(3,3) = .153 + .034 + .169 + .025 + .017 = .398,$$

$$\begin{aligned} E(3,3) &= (.220)(.229)(.263) + (.220)(.102)(.356) + (.322)(.585)(.314) \\ &\quad + (.186)(.059)(.051) + (.051)(.025)(.017) = .081, \end{aligned}$$

and

$$\kappa(3,3) = \frac{.398 - .081}{1 - .081} = .345.$$

In general, for fixed $m$ raters $\kappa(m,g)$ will produce different values for different values of $g$. For example, for the data in Tables 1, 2 and 3 we have $\kappa(3,2) = .413$ and $\kappa(3,3) = .345$.

# 3 A Schouten-type inequality

In this section we derive a Schouten-type inequality for $\kappa(m,g)$. The inequality is named after Schouten (1986) who was the first to present this type of inequality for Cohen's $\kappa$ for two raters. Lemma 1 is used in the proof of Theorem 1.

**Lemma 1.** *Let $n, h \in \mathbb{N}_{\geq 3}$ and let $a_1, a_2, \ldots, a_n$, $b_1, b_2, \ldots, b_n$, $e_1, e_2, \ldots, e_h$ and $d_1, d_2, \ldots, d_h$ be nonnegative real numbers. Suppose that at least one $a_\ell$, one $b_\ell$, one $e_q$ and one $d_q$ is not zero, and that*

$$\sum_{\ell=1}^{n} b_\ell > \sum_{q=1}^{h} d_q.$$

*Then*

$$\frac{\sum_{\ell=1}^{n} a_\ell - \sum_{q=1}^{h} e_q}{\sum_{\ell=1}^{n} b_\ell - \sum_{q=1}^{h} d_q} < \frac{\sum_{\ell=1}^{n} a_\ell}{\sum_{\ell=1}^{n} b_\ell} \quad \Leftrightarrow \quad \frac{\sum_{q=1}^{h} e_q}{\sum_{q=1}^{h} d_q} > \frac{\sum_{\ell=1}^{n} a_\ell}{\sum_{\ell=1}^{n} b_\ell}.$$

*Proof:* Let $a = \sum_{\ell=1}^{n} a_\ell$, $b = \sum_{\ell=1}^{n} b_\ell$, $e = \sum_{q=1}^{h} e_q$ and $d = \sum_{q=1}^{h} d_q$. Since $a, b, e, d > 0$ and $b - d > 0$ we have $(a-e)/(b-d) < a/b \Leftrightarrow b(a-e) < a(b-d)$ $\Leftrightarrow be > ad \Leftrightarrow e/d > a/b$. $\square$

For Theorem 1 below we assume the following situation. For $m \geq 2$ raters let $\mathbf{P}_\ell^{(g)}$ for $\ell \in \left\{1, 2, \ldots, \binom{m}{g}\right\}$ denote the $\binom{m}{g}$ distinct $g$-agreement tables with $k \geq 3$ categories. Let $\kappa(m,g)$ denote the kappa value corresponding to the $\mathbf{P}_\ell^{(g)}$ for $\ell \in \left\{1, 2, \ldots, \binom{m}{g}\right\}$. Furthermore, let $\kappa^*(m,g)$ denote the kappa values corresponding to $g$-agreement tables that are obtained by combining categories $t$ and $u$ of the $\mathbf{P}_\ell^{(g)}$.

We have the following Schouten-type inequality for $\kappa(m,g)$.

**Theorem 1.** $\kappa^*(m,g) > \kappa(m,g) \Leftrightarrow$

$$\frac{\sum\limits_{r_1<\cdots<r_g}^{m}\left[\sum\limits_{c_i\in\{t,u\}} p\left(\begin{smallmatrix} r_1 & \cdots & r_g \\ c_1 & \cdots & c_g \end{smallmatrix}\right) - p\left(\begin{smallmatrix} r_1 & \cdots & r_g \\ t & \cdots & t \end{smallmatrix}\right) - p\left(\begin{smallmatrix} r_1 & \cdots & r_g \\ u & \cdots & u \end{smallmatrix}\right)\right]}{\sum\limits_{r_1<\cdots<r_g}^{m}\left[\sum\limits_{c_i\in\{t,u\}}\prod\limits_{j=1}^{g} p_{c_i}^{r_j} - \prod\limits_{j=1}^{g} p_t^{r_j} - \prod\limits_{j=1}^{g} p_u^{r_j}\right]} > \frac{\binom{m}{g} - O(m,g)}{\binom{m}{g} - E(m,g)}.$$

*Proof:* Let $n = \binom{m}{g}(k^g - k)$ and $h = \binom{m}{g}(2^g - 2)$. Since $O(m,g)$, $E(m,g)$ and

$$\binom{m}{g} = \sum_{r_1<\cdots<r_g}^{m} 1 = \sum_{r_1<\cdots<r_g}^{m} \sum_{c_i\in\{1,\ldots,k\}} p\left(\begin{smallmatrix} r_1 & \cdots & r_g \\ c_1 & \cdots & c_g \end{smallmatrix}\right)$$

are finite sums, we can choose the $a_\ell$, $b_\ell$, $e_q$ and $d_q$ in Lemma 1 such that

$$\sum_{\ell=1}^{n} a_\ell = \binom{m}{g} - O(m,g) \quad \text{and} \quad \sum_{\ell=1}^{n} b_\ell = \binom{m}{g} - E(m,g)$$

and

$$\sum_{q=1}^{h} e_q = \sum_{r_1<\cdots<r_g}^{m}\left[\sum_{c_i\in\{t,u\}} p\left(\begin{smallmatrix} r_1 & \cdots & r_g \\ c_1 & \cdots & c_g \end{smallmatrix}\right) - p\left(\begin{smallmatrix} r_1 & \cdots & r_g \\ t & \cdots & t \end{smallmatrix}\right) - p\left(\begin{smallmatrix} r_1 & \cdots & r_g \\ u & \cdots & u \end{smallmatrix}\right)\right],$$

$$\sum_{q=1}^{h} d_q = \sum_{r_1<\cdots<r_g}^{m}\left[\sum_{c_i\in\{t,u\}}\prod_{j=1}^{g} p_{c_i}^{r_j} - \prod_{j=1}^{g} p_t^{r_j} - \prod_{j=1}^{g} p_u^{r_j}\right].$$

If we combine categories $t$ and $u$, the observed agreement $O(m,g)$ is increased by $\sum_{q=1}^{h} e_q$ whereas the expected agreement $E(m,g)$ is increased by $\sum_{q=1}^{h} d_q$. We have

$$\kappa^*(m,g) = \frac{O(m,g) - E(m,g) + \sum\limits_{q=1}^{h}(e_q - d_q)}{\binom{m}{g} - E(m,g) - \sum\limits_{q=1}^{h} d_q}.$$

Hence

$$\frac{\sum\limits_{\ell=1}^{n} a_\ell - \sum\limits_{q=1}^{h} e_q}{\sum\limits_{\ell=1}^{n} b_\ell - \sum\limits_{q=1}^{h} d_q} = \frac{\binom{m}{g} - O(m,g) - \sum\limits_{q=1}^{h} e_q}{\binom{m}{g} - E(m,g) - \sum\limits_{q=1}^{h} d_q} = 1 - \kappa^*(m,g).$$

We also have

$$\frac{\sum\limits_{\ell=1}^{n} a_\ell}{\sum\limits_{\ell=1}^{n} b_\ell} = \frac{\binom{m}{g} - O(m,g)}{\binom{m}{g} - E(m,g)} = 1 - \kappa(m,g).$$

Since $\kappa^*(m,g) > \kappa(m,g) \Leftrightarrow 1-\kappa^*(m,g) < 1-\kappa(m,g)$, the result then follows from applying Lemma 1. □

# 4 An existence theorem

In this section we show that the multi-rater coefficient $\kappa(m,g)$ can always be increased and decreased by combining categories.

The following result comes from Warrens (2010c, pp. 674-675). The proof of Lemma 2 was provided by an anonymous reviewer.

**Lemma 2.** *Let $n \in \mathbb{N}_{\geq 2}$ and let $a_1, a_2, \ldots, a_n$, at least 2 non zero and non identical, and $b_1, b_2, \ldots, b_n$ be real nonnegative numbers with $b_s \neq 0$ if $a_s \neq 0$ for all $s \in \{1,\ldots,n\}$ and $b_s \neq a_s$ for at least one $s \in \{1,\ldots,n\}$. Furthermore, let $a = \sum_{s=1}^n a_s$ and $b = \sum_{s=1}^n b_s$. Then there exist indices $i, i' \in \{1,\ldots,n\}$ with $i \neq i'$ such that*

$$\frac{a_i}{b_i} > \frac{a}{b} \quad \text{and} \quad \frac{a_{i'}}{b_{i'}} < \frac{a}{b}.$$

*Proof:* Without loss of generality, let $a_1 > 0$ and $a_2 > 0$ ($a_1 \neq a_2$). (*i*) ($n = 2$) Since $b_1 \neq a_1$ or $b_2 \neq a_2$, we immediately have $a_1/b_1 < a/b$ if $a_2/b_2 > a/b$ and $a_2/b_2 < a/b$ if $a_1/b_1 > a/b$. (*ii*) ($n > 2$) Suppose $a_1/b_1 > a/b$. Then there exists a $s \in \{2,\ldots,n\}$ such that $a_s/b_s < a/b$. Indeed, suppose $a_s/b_s > a/b$, $s \in \{2,\ldots,n\}$. Then $a_s b > b_s a$ and by summation $(a_2 + a_3 + \cdots + a_n)b > (b_2 + b_3 + \ldots + b_n)a$ or $(a - a_1)b > (b - b_1)a$ or $a_1 b < b_1 a$ or $a_1/b_1 < a/b$, which contradicts the starting assumption.
□

In Theorem 2 we assume the same situation that is assumed for Theorem 1. Theorem 2 shows that it is always possible to increase or decrease the value of $\kappa(m,g)$ by merging two categories.

**Theorem 2.** *Assume that one $\mathbf{P}_\ell^{(g)}$ has at least 2 non identical and non zero elements. Then there exist categories $t$ and $u$ such that $\kappa^*(m,g) > \kappa(m,g)$ if $t$ and $u$ are combined. Furthermore, there exist categories $t'$ and $u'$, $t \neq t'$ and/or $u \neq u'$, such that $\kappa^*(m,g) < \kappa(m,g)$ if $t'$ and $u'$ are combined.*

*Proof:* Note that, since

$$\sum_{r_1<\cdots<r_g}^{m} \sum_{c_i \in \{1,\ldots,k\}} p\left(\begin{smallmatrix} r_1 & \cdots & r_g \\ c_1 & \cdots & c_g \end{smallmatrix}\right) = \sum_{r_1<\cdots<r_g}^{m} 1 = \binom{m}{g}$$

and

$$\sum_{r_1<\cdots<r_g}^{m} \sum_{c_i \in \{1,\ldots,k\}} \prod_{j=1}^{g} p_{c_i}^{r_j} = \sum_{r_1<\cdots<r_g}^{m} \prod_{j=1}^{g} \left( \sum_{i=1}^{k} p_{c_i}^{r_j} \right) = \sum_{r_1<\cdots<r_g}^{m} 1 = \binom{m}{g},$$

the $p\left(\begin{smallmatrix} r_1 & \cdots & r_g \\ c_1 & \cdots & c_g \end{smallmatrix}\right)$ and the $\prod_{j=1}^{g} p_{c_i}^{r_j}$ for $r_j \in \{1, 2, \ldots, m\}$ and $c_i \in \{1, 2, \ldots, k\}$, satisfy the criteria of the $a_s$ and $b_s$ of Lemma 2. Since we have finite sums we can choose the $a_s$ and $b_s$ such that for each pair of categories $t$ and $u$, there is a $a_s$ equal to the term

$$\sum_{r_1<\cdots<r_g}^{m} \left[ \sum_{c_i \in \{t,u\}} p\left(\begin{smallmatrix} r_1 & \cdots & r_g \\ c_1 & \cdots & c_g \end{smallmatrix}\right) - p\left(\begin{smallmatrix} r_1 & \cdots & r_g \\ t & \cdots & t \end{smallmatrix}\right) - p\left(\begin{smallmatrix} r_1 & \cdots & r_g \\ u & \cdots & u \end{smallmatrix}\right) \right]$$

and a $b_s$ equal to the term

$$\sum_{r_1<\cdots<r_g}^{m} \left[ \sum_{c_i \in \{t,u\}} \prod_{j=1}^{g} p_{c_i}^{r_j} - \prod_{j=1}^{g} p_t^{r_j} - \prod_{j=1}^{g} p_u^{r_j} \right]$$

for $s \in \left\{1, 2, \ldots, \binom{k}{2}\right\}$. In this case we have

$$\sum_{s=1}^{\binom{k}{2}} a_s = \binom{m}{g} - O(m, g) \quad \text{and} \quad \sum_{s=1}^{\binom{k}{2}} b_s = \binom{m}{g} - E(m, g),$$

and the result follows from application of Lemma 2 and Theorem 1. □

# 5 Numerical illustrations

To illustrate the existence theorem (Theorem 2) we consider the agreement data in Table 2 (and corresponding Tables 1 and 3) for three raters on five categories. For this $5 \times 5 \times 5$ table, denoted by (1)(2)(3)(4)(5), we have $\kappa(3, 2) = .413$ and $\kappa(3, 3) = .345$ (see Section 2). Let the collapsed $4 \times 4 \times 4$ table that is obtained by combining categories 1 and 2 be denoted by (12)(3)(4)(5). The kappa values corresponding to (12)(3)(4)(5) are

$$\kappa(3, 2) = \frac{2.000 - 1.122}{3 - 1.122} = .468 \qquad \text{and} \qquad \kappa(3, 3) = \frac{.517 - .150}{1 - .150} = .432.$$

Thus, both kappa values increase when categories 1 and 2 are merged. The table (12)(3)(4)(5) also illustrates that the increase may be more substantial for one multi-rater kappa compared to another kappa ($.468 - .413 = .055 < .087 = .432 - .345$). If we in addition combine the categories 3 and 4 we obtain the $3 \times 3 \times 3$ table denoted by (12)(34)(5). The kappa values corresponding to (12)(34)(5) are

$$\kappa(3, 2) = \frac{2.305 - 1.373}{3 - 1.373} = 0.573 \qquad \text{and} \qquad \kappa(3, 3) = \frac{.653 - .209}{1 - .209} = .560,$$

which illustrates that the multi-rater kappas can be increased by successively merging categories ($.413 \rightarrow .468 \rightarrow .573$ and $.345 \rightarrow .432 \rightarrow .560$). If we

combine the categories 1, 2 and 3 instead we obtain the $3 \times 3 \times 3$ table denoted by (123)(4)(5). The kappa values corresponding to (123)(4)(5) are

$$\kappa(3,2) = \frac{2.602 - 2.288}{3 - 2.288} = .440 \qquad \text{and} \qquad \kappa(3,3) = \frac{.805 - .651}{1 - .651} = .441,$$

which shows that one kappa value may decrease (from .468 to .440) while another kappa value increases (from .432 to .441). This example also illustrates that it depends on the data which $g$-agreement kappa ($\kappa(3,2)$ or $\kappa(3,3)$) has the highest value.

The values of the multi-rater kappas can also decrease if two categories are merged. For example, if we combine the categories 2 and 5 we obtain the $4 \times 4 \times 4$ table denoted by (1)(25)(3)(4). The kappa values corresponding to (1)(25)(3)(4) are

$$\kappa(3,2) = \frac{1.712 - .848}{3 - .848} = .402 \qquad \text{and} \qquad \kappa(3,3) = \frac{.398 - .086}{1 - .086} = .342.$$

If we in addition combine the categories 1 and 4 we obtain the $3 \times 3 \times 3$ table denoted by (14)(25)(3). The kappa values corresponding to (14)(25)(3) are

$$\kappa(3,2) = \frac{1.729 - .991}{3 - .991} = .367 \qquad \text{and} \qquad \kappa(3,3) = \frac{.246 - .109}{1 - .109} = .154,$$

which illustrates that the multi-rater kappas can be decreased by successively merging categories ($.413 \rightarrow .402 \rightarrow .367$ and $.345 \rightarrow .342 \rightarrow .154$).

# 6 Conclusion

In this paper we considered a family of multi-rater kappas that extend the popular descriptive statistic Cohen's $\kappa$ for two raters. The multi-rater kappas are based on the concept of $g$-agreement ($g \in \{2, 3, \ldots, m\}$), which refers to the situation in which it is decided that there is agreement if $g$ out of $m$ raters assign an object to the same category. For the family of multi-rater kappas we proved the following existence theorem: In the case of three or more nominal categories there exist for each multi-rater kappa $\kappa(m, g)$ two categories such that, when combined, the $\kappa(m, g)$ value increases. In addition, there exist two categories such that, when combined, the $\kappa(m, g)$ value decreases. The theorem is an existence theorem since it states that there exist categories for increasing (decreasing) the $\kappa(m, g)$ value, although it does not specify which categories these are. The inequality in Theorem 1 can be used to check if the $\kappa(m, g)$ value increases or decreases when two categories are combined. The special case for $m = 2$ raters of this inequality was used in a procedure in Schouten (1986) to find categories that are easily confused. The multi-rater inequality in Theorem 1 can be used for a similar procedure for kappas for multiple raters.

## References


Agresti, A. (1990). *Categorical Data Analysis.* Wiley: New York.

Berry, K. J., & Mielke, P. W. (1988). A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters. *Educational and Psychological Measurement, 48,* 921-933.

Berry, K. J., Johnston, J. E., & Mielke, P. W. (2008). Weighted kappa for multiple raters. *Perceptual and Motor Skills, 107,* 837-848.

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

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.

Davies, M., & Fleiss, J. L. (1982). Measuring agreement for multinomial data. *Biometrics, 38,* 1047-1051.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions.* Wiley: New York.

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

Heuvelmans, A. P. J. M., & Sanders, P. F. (1993). Beoordelaarsovereenstemming. In T.J.H.M. Eggen, P.F. Sanders (eds), *Psychometrie in de Praktijk,* (pp. 443-470). Arnhem: Cito Instituut voor Toestontwikkeling.

Holmquist, N. D., McMahan, C. A., & Williams, O. D. (1967). Variability in classification of carcinoma in situ of the uterine cervix. *Archives of Pathology, 84,* 334-345.

Hsu, L. M., & Field, R. (2003). Interrater agreement measures: Comments on $\text{kappa}_\text{n}$, Cohen's kappa, Scott's $\pi$ and Aickin's $\alpha$. *Understanding Statistics, 2,* 205-219.

Hubert, L. (1977). Kappa revisited. *Psychological Bulletin, 84,* 289-297.

Janson, H., & Olsson, U. (2001). A measure of agreement for interval or nominal multivariate observations. *Educational and Psychological Measurement, 61,* 277-289.

Kraemer, H. C. (1979). Ramifications of a population model for $\kappa$ as a coefficient of reliability. *Psychometrika, 44,* 461-472.

Kraemer, H. C., Periyakoil, V. S., & Noda, A. (2002). Kappa coefficients in medical research. *Statistics in Medicine, 21,* 2109-2129.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics, 33,* 363-374.

Mielke, P. W., Berry, K. J., & Johnston, J. E. (2007). The exact variance of weighted kappa with multiple raters. *Psychological Reports, 101,* 655-660.

Mielke, P. W., Berry, K. J., & Johnston, J. E. (2008). Resampling probability values for weighted kappa with multiple raters. *Psychological Reports, 102,* 606-613.

Nelson, J. C., & Pepe, M. S. (2000). Statistical description of interrater variability in ordinal ratings. *Statistical Methods in Medical Research, 9,* 475-496.

Popping, R. (1983). *Overeenstemmingsmaten voor Nominale Data.* PhD thesis, Rijksuniversiteit Groningen, Groningen.

Popping, R. (2010). Some views on agreement to be used in content analysis studies. *Quality & Quantity, 44,* 1067-1078.

Schouten, H. J. A. (1986). Nominal scale agreement among observers. *Psychometrika, 51,* 453-466.

Vanbelle, S., & Albert, A. (2009a). Agreement between two independent groups of raters. *Psychometrika, 74,* 477-491.

Vanbelle, S., & Albert, A. (2009b). Agreement between an isolated rater and a group of raters. *Statistica Neerlandica, 63,* 82-100.

Vanbelle, S., & Albert, A. (2009c). A note on the linearly weighted kappa coefficient for ordinal scales. *Statistical Methodology, 6,* 157-163.

Von Eye, A., & Mun, E. Y. (2006). *Analyzing Rater Agreement. Manifest Variable Methods.* Lawrence Erlbaum Associates.

Warrens, M. J. (2008a). On the equivalence of Cohen's kappa and the Hubert-Arabie adjusted Rand index. *Journal of Classification, 25,* 177-183.

Warrens, M. J. (2008b). On similarity coefficients for $2 \times 2$ tables and correction for chance. *Psychometrika, 73,* 487-502.

Warrens, M. J. (2009). $k$-Adic similarity coefficients for binary (presence/absence) data. *Journal of Classification, 26,* 227-245.

Warrens, M. J. (2010a). Inequalities between kappa and kappa-like statistics for $k \times k$ tables. *Psychometrika, 75,* 176-185.

Warrens, M. J. (2010b). A formal proof of a paradox associated with Cohen's kappa. *Journal of Classification, 27,* 322-332.

Warrens, M. J. (2010c). Cohen's kappa can always be increased and decreased by combining categories. *Statistical Methodology, 7,* 673-677.

Warrens, M. J. (2010d). Inequalities between multi-rater kappas. *Advances in Data Analysis and Classification, 4,* 271-286.

Warrens, M. J. (2011a). Weighted kappa is higher than Cohen's kappa for tridiagonal agreement tables. *Statistical Methodology, 8,* 268-272.

Warrens, M. J. (2011b). Cohen's linearly weighted kappa is a weighted average of $2 \times 2$ kappas. *Psychometrika, 76,* 471-486.

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

Table 1: Relative frequencies of classifications of 118 slides by two pathologists.

| | Pathologist 2 | | | | | |
|---|---|---|---|---|---|---|
| Pathologist 1 | 1 | 2 | 3 | 4 | 5 | Row totals |
| 1 | .186 | .017 | .017 | 0 | 0 | .220 |
| 2 | .042 | .059 | .119 | 0 | 0 | .220 |
| 3 | 0 | .017 | .305 | 0 | 0 | .322 |
| 4 | 0 | .008 | .119 | .059 | 0 | .186 |
| 5 | 0 | 0 | .025 | 0 | .025 | .051 |
| Column totals | .229 | .102 | .585 | .059 | .025 | 1 |

Table 2: Five slices of the 3-dimensional $5 \times 5 \times 5$ table of relative frequencies of classifications of 118 slides by three pathologists.

| | | | Pathologist 2 | | | Category |
|---|---|---|---|---|---|---|
| Pathologist 1 | 1 | 2 | 3 | 4 | 5 | Pathologist 3 |
| 1 | .153 | .008 | 0 | 0 | 0 | Category 1 |
| 2 | .017 | .025 | .034 | 0 | 0 | Total = .263 |
| 3 | 0 | 0 | 0 | 0 | 0 | |
| 4 | 0 | 0 | .017 | 0 | 0 | |
| 5 | 0 | 0 | 0 | 0 | .008 | |
| 1 | .034 | .008 | .017 | 0 | 0 | Category 2 |
| 2 | .025 | .034 | .085 | 0 | 0 | Total = .356 |
| 3 | 0 | .017 | .136 | 0 | 0 | |
| 4 | 0 | 0 | 0 | 0 | 0 | |
| 5 | 0 | 0 | 0 | 0 | 0 | |
| 1 | 0 | 0 | 0 | 0 | 0 | Category 3 |
| 2 | 0 | 0 | 0 | 0 | 0 | Total = .314 |
| 3 | 0 | 0 | .169 | 0 | 0 | |
| 4 | 0 | .008 | .085 | .034 | 0 | |
| 5 | 0 | 0 | .017 | 0 | 0 | |
| 1 | 0 | 0 | 0 | 0 | 0 | Category 4 |
| 2 | 0 | 0 | 0 | 0 | 0 | Total = .051 |
| 3 | 0 | 0 | 0 | 0 | 0 | |
| 4 | 0 | 0 | .017 | .025 | 0 | |
| 5 | 0 | 0 | .008 | 0 | 0 | |
| 1 | 0 | 0 | 0 | 0 | 0 | Category 5 |
| 2 | 0 | 0 | 0 | 0 | 0 | Total = .017 |
| 3 | 0 | 0 | 0 | 0 | 0 | |
| 4 | 0 | 0 | 0 | 0 | 0 | |
| 5 | 0 | 0 | 0 | 0 | 0.17 | |

Table 3: Relative frequencies of classifications of 118 slides by two pairs of pathologists.

| | Pathologist 3 | | | | | |
|---|---|---|---|---|---|---|
| Pathologist 1 | 1 | 2 | 3 | 4 | 5 | Row totals |
| 1 | .161 | .059 | 0 | 0 | 0 | .220 |
| 2 | .076 | .144 | 0 | 0 | 0 | .220 |
| 3 | 0 | .153 | .169 | 0 | 0 | .322 |
| 4 | .017 | 0 | .127 | .042 | 0 | .186 |
| 5 | .008 | 0 | .017 | .008 | .017 | .051 |
| Column totals | .263 | .356 | .314 | .051 | .017 | 1 |

| | Pathologist 3 | | | | | |
|---|---|---|---|---|---|---|
| Pathologist 2 | 1 | 2 | 3 | 4 | 5 | Row totals |
| 1 | .169 | .059 | 0 | 0 | 0 | .229 |
| 2 | .034 | .059 | .008 | 0 | 0 | .102 |
| 3 | .051 | .237 | .271 | .025 | 0 | .585 |
| 4 | 0 | 0 | .034 | .025 | 0 | .059 |
| 5 | .008 | 0 | 0 | 0 | .017 | .025 |
| Column totals | .263 | .356 | .314 | .051 | .017 | 1 |