A COMPARISON OF MULTI-WAY SIMILARITY COEFFICIENTS FOR BINARY SEQUENCES

Matthijs J. Warrens
Leiden University, Institute of Psychology, Unit Methodology and Statistics
P.O. Box 9555, 2300 RB Leiden, Email: warrens@fsw.leidenuniv.nl

ABSTRACT

The paper compares three formulations of *n*-way (for groups of size $n \geq 2$) similarity coefficients for binary sequences. Properties that the similarity coefficients may have in general, not just for specific data, are discussed, and it is investigated how the different *n*-way formulations are related. Using the *n*-way Bennani-Heiser coefficients, the similarity between *m* sequences ($2 \leq m \leq n$) is always equal to or greater than the similarity between the *m* sequences and $n - m$ other sequences. *n*-Way coefficients based on 2-way information lack several of the properties that the Bennani-Heiser coefficients possess. For example, with the former coefficients it is possible to have zero similarity between two objects, but positive similarity between the two objects and a third object.

Keywords: *Multi-way coefficients; n-Way measures; Simple matching coefficient; Jaccard coefficient; Dice coefficient; Bennani-Heiser coefficients*.

1. INTRODUCTION

Sequences of binary scores occur in various fields of data analysis and classification. Generally speaking, a sequence corresponds to an object or individual and the binary scores reflect the presence or absence of certain attributes of the object [2,18]. An object may be a person that may or may not possess certain traits, or a location where certain species types do or do not occur. In many cases one wants to determine the amount of similarity (agreement, resemblance) between two binary sequences. The classification literature contains a vast amount of similarity coefficients that can be used to quantify the similarity between binary sequences [3,23,31,32,35]. Popular examples are the simple matching coefficient [29] and the Jaccard coefficient [21]. We do not consider coefficients that measure association between two binary variables in this paper. An example of an association coefficient is the phi coefficient. Pairwise similarity coefficients play a central role in data analysis and classification. Individual coefficients can be used for summarizing parts of a research study, while coefficient matrices can be used as input for multivariate data analysis techniques like component analysis [15,18] or cluster analysis [1,30,34].

Coefficients that reflect the similarity between two sequences are here called 2-way coefficients. 2-Way coefficients only allow comparison of two sequences at a time. Let *n* be a positive integer. Multi-way or *n*-way coefficients (for groups of size $n \geq 2$) may be used to compare *n* objects at a time [7,11,36]. For example, the 2-way Jaccard coefficient [21] measures the number of species types that are found together in two locations, relative to the total number of species types that are found in the two locations. The 3-way Jaccard coefficient [4,7,19] measures the number of species types that are found together in three locations, relative to the total number of species types that are founding the three locations. Hence, *n*-way coefficients can be used if one wants to know the degree of resemblance between 3, 4 or *n* objects. For the free sorting method, Daws [8] showed that reduction of a distribution over all subset patterns to 2-way similarity implies loss of information about how the individuals have classified the objects [19]. Furthermore, similar to 2-way coefficients, *n*-way coefficients may be used as input in several methods of multi-way data analysis, including three-way multidimensional scaling and three-way hierarchical cluster analysis [4,7,19,22]. Some *n*-way coefficients that are widely used in practice are the multi-rater versions of Cohen's kappa [5] proposed in [14,24,25]. Kappa is a popular descriptive statistic for assessing inter-rater reliability on a nominal scale [38,40,41,42,43].

In the classification literature, *n*-way similarity coefficients are usually defined as functions of the 2-way or pairwise information [26]. If one is interested in the similarity between three or more objects at a time, an intuitive and appealing option in statistics is taking the average of all the pairwise coefficients that can be formed between the objects. For example, Conger [6] showed that the multi-rater extension of Cohen's kappa proposed in Light [24] is the arithmetic mean of the $n(n-1)/2$ pairwise kappas that can be calculated with *n* raters [25]. For binary sequences, Warrens [33] studied a family of *n*-way coefficients that preserve the relations between coefficients with respect to correction for chance. This family includes the multi-rater kappa proposed in [6,20,25].

De Rooij [9] showed that *n*-way coefficients that are functions of 2-way coefficients do not give more information than is already present in the 2-way coefficients, that is, no higher order relations are given by these *n*-way

coefficients. Following Heiser and Bennani [19], Warrens [36] formulated *n*-way coefficients for binary sequences that generalize basic characteristics of 2-way coefficients. In contrast to other *n*-way coefficients from the literature, the *n*-way coefficients proposed in [36] are not functions of the 2-way information, but can be considered coefficients of simultaneous similarity [6,26].

In this paper, we compare three formulations of *n*-way similarity coefficients for binary sequences that can be found in the literature. Furthermore, we discuss properties that the similarity coefficients may have in general, not just for certain data, and investigate how the *n*-way formulations are related. The paper is organized as follows. In the next section we introduce the 2-way coefficients. Basic definitions of *n*-way coefficients are presented in Section 3. Three classes of *n*-way generalizations of the 2-way coefficients are considered in Sections 4 to 6. In Section 7, we consider up to four *n*-way generalizations of the Dice similarity coefficient. Section 8 contains a discussion.

Bennani-Heiser coefficients and some of their properties are discussed in Section 4. Using Bennani-Heiser coefficients, the similarity between m sequences ($2 \leq m \leq n$) is never smaller than the similarity between the m sequences and $n - m$ additional sequences. An analogous condition for dissimilarities is considered a desirable property for distance functions [4,22]. Sections 5 and 6 contain *n*-way coefficients that are functions of the 2-way information. These coefficients lack some of the properties of the coefficients discussed in Section 4. For example, with these coefficients it is possible to have zero similarity between two objects, but positive similarity between the two objects and a third object.

2. WAY SIMILARITY COEFFICIENTS

We will use the symbol S to denote a similarity coefficient. A 2-way similarity coefficient $S^{(2)}$ on a nonempty set of objects E, is a function from the Cartesian product $E \times E$ to the real unit interval [0,1] that is symmetric, $S(i,j) = S(j,i)$, and satisfies $S(i,j) \leq S(i,i) = 1$ for all i, j in E. In this paper, the objects i and j are binary sequences (profiles, score patterns) of the same finite length u, where $u \geq 1$ is a positive integer. Many coefficients can be defined using the four dependent proportions a_{ij}, b_{ij}, c_{ij} and d_{ij} presented in Table 1. Instead of proportions, Table 1 may also be defined on counts or frequencies; proportions are used here for notational convenience. Table 1 is a cross-classification of two binary sequences. It is also called a 2×2 table [32,33].

Table 1: Bivariate proportions table for binary sequences.

	Sequence j		
Sequence i	Value 1	Value 2	Total
Value 1	a_{ij}	b_{ij}	p_i
Value 2	c_{ij}	d_{ij}	q_i
Total	p_j	q_j	1

In Table 1, a_{ij}, b_{ij}, c_{ij} and d_{ij} are joint proportions, whereas p_i and q_i are marginal proportions. If value 1 and value 2 in Table 1 are, respectively, 1 and 0, then a_{ij} is the proportion of 1s that i and j share in the same positions, and d_{ij} can be interpreted as the proportion of 0s that i and j share in the same positions. More precisely, if sequences i and j are the ratings of u individuals by two observers on the presence or absence of a trait, or the presence/absence codings of u species types in two locations, then a_{ij} and d_{ij} can be interpreted as, respectively, the proportion of positive matches and negative matches. Instead of two binary sequences, Albatineh et al. [1] consider two methods for clustering data, and a_{ij} is the proportion of data points that were placed in the same cluster according to methods i and j. Quantity p_i is the proportionof 1s in sequence i.

If there is no confusion possible, we will use $a^{(2)}$ and $d^{(2)}$ for short, instead of a_{ij} and d_{ij}. Furthermore, many similarity coefficients for binary sequences (or 2×2 tables) are defined as ratios. It may occur that the denominator of a similarity coefficient has zero value, in which case the value of the coefficient is indeterminate [2,35]. In the following we assume that the value of each coefficient S is defined. See Batagelj and Bren [2] and Warrens [35] for robust definitions of similarity coefficients for 2×2 tables.

A straightforward coefficient of similarity is the observed proportion of agreement

$$S_{\mathrm{SM}}^{(2)} = a^{(2)} + d^{(2)}.$$

Coefficient $S_{\mathrm{SM}}^{(2)}$ is also known as the simple matching coefficient [29]. The subscript of S, for example SM in $S_{\mathrm{SM}}^{(2)}$, will be used to distinguish the various coefficients. The capital letters reflect the authors to whom the coefficient or coefficient family can be attributed [1,31,32,33,35,36].

Coefficient $S_{\mathrm{SM}}^{(2)}$ is the main member of the parameter family

$$S_{\mathrm{GL}}^{(2)}(\theta)=\frac{a^{(2)}+d^{(2)}}{a^{(2)}+d^{(2)}+\theta(1-a^{(2)}-d^{(2)})}=\frac{a^{(2)}+d^{(2)}}{\theta+(1-\theta)(a^{(2)}+d^{(2)})},$$

where $\theta > 0$ is used to avoid negative values. Coefficient $S_{\mathrm{SM}}^{(2)}=S_{\mathrm{GL}}^{(2)}(1)$. The Gower-Legendre family $S_{\mathrm{GL}}^{(2)}(\theta)$ was first studied in Gower [16] and Gower and Legendre [18, p. 13]. The numerator of $S_{\mathrm{GL}}^{(2)}(\theta)$ is equal to coefficient $S_{\mathrm{SM}}^{(2)}$, whereas the denominator is θ plus $(1-\theta)$ times coefficient $S_{\mathrm{SM}}^{(2)}$.

A binary sequence can be either a nominal or an ordinal variable. In the latter case a 1 is `more' in a sense than a 0, for example, species presence/absence in ecology. Coefficient $S_{\mathrm{SM}}^{(2)}$ is a popular coefficient if the sequences are nominal. If the data are ordinal, popular choices are the Jaccard coefficient [21]

$$S_{\mathrm{J}}^{(2)}=\frac{a^{(2)}}{1-d^{(2)}},$$

and the Dice-Sørenson coefficient [12,27]

$$S_{\mathrm{D}}^{(2)}=\frac{2a^{(2)}}{p_i+p_j}=\frac{2a^{(2)}}{1+a^{(2)}-d^{(2)}}.$$

Coefficients $S_{\mathrm{J}}^{(2)}$ and $S_{\mathrm{D}}^{(2)}$ are members of parameter family

$$S_{\mathrm{FG}}^{(2)}(\theta)=\frac{a^{(2)}}{a^{(2)}+\theta(1-a^{(2)}-d^{(2)})}=\frac{a^{(2)}}{(1-\theta)a^{(2)}+\theta(1-d^{(2)})},$$

where $\theta > 0$. Coefficient $S_{\mathrm{J}}^{(2)}=S_{\mathrm{FG}}^{(2)}(1)$, and coefficient $S_{\mathrm{D}}^{(2)}=S_{\mathrm{FG}}^{(2)}(\frac{1}{2})$. The Fichet-Gower family $S_{\mathrm{FG}}^{(2)}(\theta)$ was first studied in Fichet [13] and Gower [16]. The numerator of $S_{\mathrm{FG}}^{(2)}(\theta)$ is equal to the proportion of positive matches $a^{(2)}$. The denominator of $S_{\mathrm{FG}}^{(2)}(\theta)$ is more complicated.

A main reason for studying parameter families $S_{\mathrm{GL}}^{(2)}(\theta)$ and $S_{\mathrm{FG}}^{(2)}(\theta)$ is the following property.

Property 1. As noted in [16,18], any two members of parameter family $S_{\mathrm{GL}}^{(2)}(\theta)$, or two members of $S_{\mathrm{FG}}^{(2)}(\theta)$, are globally order equivalent [28]. If two coefficients are order equivalent, they are interchangeable with respect to an analysis method that is invariant under ordinal transformations. Let us show the property for $S_{\mathrm{FG}}^{(2)}(\theta)$. Let $a_1^{(2)}$ and $a_2^{(2)}$, and $d_1^{(2)}$ and $d_2^{(2)}$, denote two versions of respectively $a^{(2)}$ and $d^{(2)}$. We have

$$\frac{a_1^{(2)}}{(1-\theta)a_1^{(2)}+\theta\left(1-d_1^{(2)}\right)}\geq\frac{a_2^{(2)}}{(1-\theta)a_2^{(2)}+\theta\left(1-d_2^{(2)}\right)}\quad\Leftrightarrow\quad\frac{a_1^{(2)}}{1-d_1^{(2)}}\geq\frac{a_2^{(2)}}{1-d_2^{(2)}}. \tag{1}$$

Since inequality (1) does not depend on θ, two members of $S_{\mathrm{FG}}^{(2)}(\theta)$ are globally order equivalent.

3. *n*-WAY SIMILARITY COEFFICIENTS

In this paper, we consider three approaches of formulating *n*-way similarity coefficients for binary sequences. A 3-way similarity coefficient $S^{(3)}$ on a set of objects E is a function from the Cartesian product $E\times E\times E$ to the real unit interval [0,1] that is symmetric, $S(i,j,k)=S(i,k,j)=S(j,i,k)=S(j,k,i)=S(k,i,j)=S(k,j,i)$, and satisfies $S(i,j,k)\leq S(i,i,i)=1$ for all i,j,k in E. The definition of a *n*-way similarity coefficient is analogous: a function $S^{(n)}\colon E^n\rightarrow[0,1]$, that satisfies multi-way symmetry [39], and obtains its maximum of unity if the *n* objects are equal. In this paper we sometimes compare a *n*-way coefficient to one of its special cases, for example, a *m*-way coefficient where $2\leq m\leq n$. Throughout the paper it is assumed that the set of *m* objects is a subset of the set with *n* objects. Thus, $S^{(m)}$ reflects the similarity between *m* objects, and $S^{(n)}$ reflects the similarity between the same *m* objects and $n-m$ additional objects. See also Property 2 at the end of this section.

In the literature, *n*-way similarity coefficients are usually defined as functions of the 2-way information. In the case of binary sequences, one may typically obtain the necessary 2-way information by constructing all $n(n-1)/2$ pairwise 2×2 tables between the *n* sequences. The coefficients discussed in Sections 5 and 6 are based on the positive and negative matches a_{ij} and d_{ij}.

The Bennani-Heiser coefficients discussed in Section 4 are not functions of the 2-way information. For these coefficients we must extend the concept of the 2-way or bivariate 2×2 table from Section 2 to a multi-way or *n*-way contingency table. In the 2-way case, the positive and negative matches $a^{(2)}$ and $d^{(2)}$ are the elements of the main diagonal of the 2×2 table. Quantities $a^{(2)}$ and $d^{(2)}$ can be interpreted as the proportions of 1s and 0s that two sequences share in the same positions. For *n* binary sequences we define the proportions:

$a^{(n)}$ = proportion of 1s that *n* sequences share in the same positions;
$d^{(n)}$ = proportion of 0s that *n* sequences share in the same positions.

The quantities $a^{(n)}$ and $d^{(n)}$ are the elements of the main diagonal of the *n*-way contingency table. Quantities $a^{(n)}$ and $d^{(n)}$ have an important property that will repeatedly be used in Sections 5, 6 and 7.

Property 2. We have $a^{(m)} \geq a^{(n)}$ and $d^{(m)} \geq d^{(n)}$ for $2 \leq m \leq n$, that is, the proportion of 1s (0s) that *m* sequences share in the same positions is always equal or greater than the proportion of 1s (0s) that the *m* sequences and $n - m$ other sequences share in the same positions.

4. BENNANI-HEISER COEFFICIENTS

Bennani-Heiser coefficients are *n*-way similarity coefficients that can be defined using only the quantities $a^{(n)}$ and $d^{(n)}$ defined in Section 3. These *n*-way formulations generalize certain basic characteristics of the corresponding 2-way versions. Warrens [36] gave the following generalizations of coefficients $S_{\mathrm{SM}}^{(2)}$, $S_{\mathrm{J}}^{(2)}$ and $S_{\mathrm{D}}^{(2)}$:

$$S_{\mathrm{SM1}}^{(n)} = a^{(n)} + d^{(n)}, \qquad S_{\mathrm{J1}}^{(n)} = \frac{a^{(n)}}{1 - d^{(n)}} \quad \text{and} \quad S_{\mathrm{D1}}^{(n)} = \frac{2a^{(n)}}{1 + a^{(n)} - d^{(n)}}.$$

Warrens [36] also gave the following generalizations of parameter families $S_{\mathrm{GL}}^{(2)}(\theta)$ and $S_{\mathrm{FG}}^{(2)}(\theta)$:

$$S_{\mathrm{GL1}}^{(n)}(\theta) = \frac{a^{(n)} + d^{(n)}}{\theta + (1 - \theta)(a^{(n)} + d^{(n)})}$$

and

$$S_{\mathrm{FG1}}^{(n)}(\theta) = \frac{a^{(n)}}{(1 - \theta)a^{(n)} + \theta(1 - d^{(n)})}.$$

The 3-way coefficient $S_{\mathrm{SM1}}^{(3)}$ and 3-way parameter family $S_{\mathrm{FG1}}^{(3)}(\theta)$ were first formulated in Bennani-Dosse [4] and Heiser and Bennani [19]. It should benoted that the function $1 - S_{\mathrm{J1}}^{(n)}$ was already used in Cox et al. [7, p. 200]. The latter function is also studied in [39].

Jaccard [21] studied the distribution of species of plants in three different Alpine districts. Coefficient $S_{\mathrm{J}}^{(2)}$ can be interpreted as the number of species types common to two districts, divided by the total number of species types in the two districts. The interpretation of coefficient $S_{\mathrm{J1}}^{(3)}$ is analogous to that of $S_{\mathrm{J}}^{(2)}$: the number of species types common to three districts, divided by the total number of species types in the three districts.

Cox et al. [7] pointed out that *n*-way coefficients may detect similarity where 2-way coefficients fail. We consider the following example.

Example 1. Suppose we have the following four binary sequences on ten attributes.

objects	attributes									
i	0	1	0	0	1	0	0	1	0	0
j	0	0	0	0	1	1	0	0	0	1
k	0	0	0	1	1	0	1	0	0	0
l	1	1	1	1	0	1	1	1	1	1

The 2-way Jaccard coefficient compares the number of positions where a 1 occurs in both sequences to the total number of positions where a 1 occurs in one of the sequences. The 3-way Jaccard coefficient [4,7,19] compares the number of positions where a 1 occurs in all three sequences to the total number of positions where a 1 occurs in one of the three sequences. For these data, the six 2-way Jaccard coefficients are all equal ($S_{\mathrm{J}}^{(2)} = 1/5$), giving no discriminative information about the objects. However, the 3-way Jaccard coefficient between objects *i*, *j* and *k* ($S_{\mathrm{J1}}^{(3)} = 1/7$) differs from the other three 3-way coefficients ($S_{\mathrm{J1}}^{(3)} = 0$). We may conclude that object *l* is different from *i*, *j* and *k*.

One may also argue that the wrong 2-way coefficient has been specified for analyzing these data. Due to Property 1, coefficient $S_{\mathrm{J}}^{(2)}$ cannot be replaced by another member of family $S_{\mathrm{FG}}^{(2)}(\theta)$, since the six 2-way coefficients between

the four objects are also equal for this other coefficient. This can be seen from replacing the inequality signs in (1) by an equality sign.

To obtain a different outcome of the 2-way data analysis, one should use a different coefficient, for example $S_{\mathrm{SM}}^{(2)}$. The 2-way simple matching coefficient compares the number of positions where either a 1 or 0 occurs in both sequences to the total number of positions. For these data, the three 2-way simple matching coefficients between *i*, *j* and *k* are $S_{\mathrm{SM}}^{(2)}(i,j) = S_{\mathrm{SM}}^{(2)}(i,k) = S_{\mathrm{SM}}^{(2)}(j,k) = 3/5$, whereas the three 2-way simple matching coefficients between *i*, *j* and *k* on the one hand and *l* on the other, are $S_{\mathrm{SM}}^{(2)}(i,l) = S_{\mathrm{SM}}^{(2)}(j,l) = S_{\mathrm{SM}}^{(2)}(k,l) = 1/5$. Again, we conclude that object *l* is different from *i*, *j* and *k*.

Any two members of parameter family $S_{\mathrm{GL}}^{(2)}(\theta)$ or two members of $S_{\mathrm{FG}}^{(2)}(\theta)$ are globally order equivalent (Property 1). The *n*-way generalizations $S_{\mathrm{GL1}}^{(n)}(\theta)$ and $S_{\mathrm{FG1}}^{(n)}(\theta)$ preserve Property 1.

Property 3. Two members of family $S_{\mathrm{GL1}}^{(n)}(\theta)$, or of $S_{\mathrm{FG1}}^{(n)}(\theta)$, are globally order equivalent. Let us show the property for $S_{\mathrm{GL1}}^{(n)}(\theta)$. Let $a_1^{(n)}$ and $a_2^{(n)}$, and $d_1^{(n)}$ and $d_2^{(n)}$, denote two versions of respectively $a^{(n)}$ and $d^{(n)}$. We have

$$\frac{a_1^{(n)} + d_1^{(n)}}{\theta + (1-\theta)(a_1^{(n)} + d_1^{(n)})} \geq \frac{a_2^{(n)} + d_2^{(n)}}{\theta + (1-\theta)\left(a_2^{(n)} + d_2^{(n)}\right)} \quad \Leftrightarrow \quad a_1^{(n)} + d_1^{(n)} \geq a_2^{(n)} + d_2^{(n)}. \tag{2}$$

Since inequality (2) does not depend on θ, two members of $S_{\mathrm{GL1}}^{(n)}(\theta)$ are globally order equivalent.

The following property of Bennani-Heiser coefficients is perhaps the most distinctive. Property 4 is closely related to Property 2.

Property 4. Bennani-Heiser coefficients satisfy $S^{(m)} \geq S^{(n)}$ for $2 \leq m \leq n$ (see Section 3), that is, the similarity between *m* sequences is always equal to or greater than the similarity between the *m* sequences and *n* – *m* additional sequences (see Example 1). The property characterizes all Bennani-Heiser coefficients in this paper and does not depend on the particular definition of similarity. For example, we have both $S_{\mathrm{SM1}}^{(m)} \geq S_{\mathrm{SM1}}^{(n)}$ and $S_{\mathrm{J1}}^{(m)} \geq S_{\mathrm{J1}}^{(n)}$.

Property 4 has its origin in the axiomatizations of three-way distances presented in [19,22]. Joly and Le Calvé [22] require that a three-way distance between three objects is not smaller than the distance between two of them. This desideratum is translated to Property 4 by transforming a similarity coefficient into a dissimilarity or distance function by taking the complement $1 - S$.

5. ALTERNATIVE *n*-WAY SIMILARITY COEFFICIENTS

Instead of using the quantities $a^{(n)}$ and $d^{(n)}$, which define Bennani-Heiser coefficients, *n*-way similarity coefficients may also be defined using the 2-way information. For example, if a_{ij} is important in the comparison of sequences *i* and *j*, then we may use a_{ij}, a_{ik} and a_{jk} when comparing *i*, *j* and *k*. In this section we consider a class of *n*-way coefficients based on 2-way quantities, that was formulated and investigated in Warrens [32,33].

Warrens [33] introduced the following generalizations of coefficients $S_{\mathrm{SM}}^{(2)}$ and $S_{\mathrm{D}}^{(2)}$ for *n* binary sequences:

$$S_{\mathrm{SM2}}^{(n)} = \frac{2}{n(n-1)} \sum_{i<j}^{n} (a_{ij} + d_{ij})$$

and

$$S_{\mathrm{D2}}^{(n)} = \frac{2\sum_{i<j}^{n} a_{ij}}{(n-1)\sum_{i}^{n} p_i}.$$

The quantity $2/[n(n-1)]$ in $S_{\mathrm{SM2}}^{(n)}$ is used to obtain $0 \leq S_{\mathrm{SM2}}^{(n)} \leq 1$. Coefficient $S_{\mathrm{SM2}}^{(n)}$ is the arithmetic mean of the $n(n-1)/2$ pairwise $S_{\mathrm{SM}}^{(2)} = a_{ij} + d_{ij}$.

Warrens [33] shows that after correction for chance, $S_{\mathrm{SM2}}^{(n)}$ and $S_{\mathrm{D2}}^{(n)}$ become identical. Under the assumption of two different frequency distributions, the cell a_{ij} of the 2×2 table (Section 2) has expectation $E(a_{ij}) = p_i p_j$, where p_i and p_j are the marginal proportions corresponding to the cell a_{ij}. If one uses $E(a_{ij}) = p_i p_j$ for all $n(n-1)/2$ different 2×2 tables, then $S_{\mathrm{SM2}}^{(n)}$ and $S_{\mathrm{D2}}^{(n)}$ become after correction for chance,

$$S_{\mathrm{P}}^{(n)} = \frac{\sum_{i<j}^{n}(a_{ij} - p_i p_j)}{\frac{n-1}{2}\sum_{i}^{n} p_i - \sum_{i<j}^{n} p_i p_j}.$$

Coefficient $S_{\mathrm{P}}^{(n)}$ is the multi-rater generalization of Cohen's kappa [5] that is discussed and studied in [6,20,25,40,41].

The heuristics used for formulating $S_{\mathrm{SM2}}^{(n)}$ and $S_{\mathrm{D2}}^{(n)}$ may also be used for generalizing parameter families $S_{\mathrm{GL}}^{(2)}(\theta)$ and $S_{\mathrm{FG}}^{(2)}(\theta)$. We obtain

$$S_{\mathrm{GL2}}^{(n)}(\theta) = \frac{\frac{2}{n(n-1)}\sum_{i<j}^{n}(a_{ij} + d_{ij})}{\theta + (1-\theta)\frac{2}{n(n-1)}\sum_{i<j}^{n}(a_{ij} + d_{ij})}.$$

and

$$S_{\mathrm{FG2}}^{(n)}(\theta) = \frac{\sum_{i<j}^{n} a_{ij}}{(1-\theta)\sum_{i<j}^{n} a_{ij} + \theta\left(\frac{n(n-1)}{2} - \sum_{i<j}^{n} d_{ij}\right)}.$$

Recall that the numerator of $S_{\mathrm{GL}}^{(2)}(\theta)$ is equal to coefficient $S_{\mathrm{SM}}^{(2)}$, whereas the denominator is θ plus $(1 - \theta)$ times coefficient $S_{\mathrm{SM}}^{(2)}$ (see Section 2). In the family $S_{\mathrm{GL2}}^{(n)}(\theta)$ the coefficient $S_{\mathrm{SM}}^{(2)}$ is replaced by its *n*-way extension $S_{\mathrm{SM2}}^{(n)}$ in both the numerator and the denominator. The family $S_{\mathrm{FG2}}^{(n)}(\theta)$ extends $S_{\mathrm{FG2}}^{(2)}(\theta)$ in a similar way.

Using the same heuristics to generalize coefficient $S_{\mathrm{J}}^{(2)}$, we obtain

$$S_{\mathrm{J2}}^{(n)} = \frac{\sum_{i<j}^{n} a_{ij}}{\frac{n(n-1)}{2} - \sum_{i<j}^{n} d_{ij}}.$$

Any two members of the 2-way parameter family $S_{\mathrm{GL}}^{(2)}(\theta)$, or two members of $S_{\mathrm{FG}}^{(2)}(\theta)$, are globally order equivalent (Property 1). The *n*-way generalizations $S_{\mathrm{GL2}}^{(n)}(\theta)$ and $S_{\mathrm{FG2}}^{(n)}(\theta)$ preserve Property 1, similar to $S_{\mathrm{GL1}}^{(n)}(\theta)$ and $S_{\mathrm{FG1}}^{(n)}(\theta)$ from the previous section (Property 3).

Property 5. Two members of family $S_{\mathrm{GL2}}^{(n)}(\theta)$, or of $S_{\mathrm{FG2}}^{(n)}(\theta)$, are globally order equivalent. Let us show the property for $S_{\mathrm{FG2}}^{(n)}(\theta)$. Let x_1 and x_2, and y_1 and y_2, denote two versions of respectively $\sum_{i<j}^{n} a_{ij}$ and $\sum_{i<j}^{n} d_{ij}$. We have

$$\frac{x_1}{(1-\theta)x_1 + \theta\left(\frac{n(n-1)}{2} - y_1\right)} \geq \frac{x_2}{(1-\theta)x_2 + \theta\left(\frac{n(n-1)}{2} - y_2\right)} \quad \Leftrightarrow \quad \frac{x_1}{\frac{n(n-1)}{2} - y_1} \geq \frac{x_2}{\frac{n(n-1)}{2} - y_2}. \tag{3}$$

Since inequality (3) does not depend on θ, two members of $S_{\mathrm{FG2}}^{(n)}(\theta)$ are globally order equivalent.

For Bennani-Heiser coefficients (Section 4), the similarity between *m* sequences is never smaller than the similarity between the *m* sequences and *n* – *m* other sequences (Property 4). The following example shows that the *n*-way coefficients considered in this section do not possess this property.

Example 2. Suppose we have three binary sequences on five attributes:

objects	attributes				
i	0	1	0	1	1
j	1	0	1	0	1
k	1	0	1	1	1

For these data the three 2-way simple matching coefficients between the three objects are $S_{\mathrm{SM}}^{(2)}(i,j) = 1/5$, $S_{\mathrm{SM}}^{(2)}(i,k) = 2/5$ and $S_{\mathrm{SM}}^{(2)}(j,k) = 4/5$. The 3-way simple matching coefficient, $S_{\mathrm{SM2}}^{(3)} = 7/15$, is the arithmetic mean of the three 2-way coefficients. Furthermore, the three 2-way Dice coefficients are $S_{\mathrm{D}}^{(2)}(i,j) = 1/3$, $S_{\mathrm{D}}^{(2)}(i,k) = 4/7$ and $S_{\mathrm{D}}^{(2)}(j,k) = 6/7$. The 3-way Dice coefficient $S_{\mathrm{D2}}^{(3)} = 3/5$. Thus, using the coefficients from

this section, the amount of similarity may increase when one increases the number of sequences or objects that are compared.

The coefficient families formulated in this section may be compared to the Bennani-Heiser families from the previous section. It turns out that coefficients from the two different approaches are bounds of one another. Theorem 2 shows how the Gower-Legendre families $S_{\mathrm{GL1}}^{(n)}(\theta)$ and $S_{\mathrm{GL2}}^{(n)}(\theta)$ are related. In the proof of Theorem 2, we use the following lemma.

Lemma 1. *Let x, y and θ be positive real numbers. Then*

$$\frac{x}{\theta + (1-\theta)x} \leq \frac{y}{\theta + (1-\theta)y} \quad \Leftrightarrow \quad x \leq y.$$

Theorem 2. $S_{\mathrm{GL1}}^{(n)}(\theta) \leq S_{\mathrm{GL2}}^{(n)}(\theta)$ *for all $\theta > 0$.*
Proof: Let

$$x = a^{(n)} + d^{(n)}, \quad \text{and} \quad y = \frac{2}{n(n-1)} \sum_{i<j}^{n} (a_{ij} + d_{ij}).$$

Due to Lemma 1, it must be shown that

$$\frac{n(n-1)}{2}\left(a^{(n)} + d^{(n)}\right) \leq \sum_{i<j}^{n} (a_{ij} + d_{ij}). \tag{4}$$

Inequality (4) follows from Property 2, that is, $a_{ij} \geq a^{(n)}$and $d_{ij} \geq d^{(n)}$. ■

Theorem 4 specifies how the Fichet-Gower families $S_{\mathrm{FG1}}^{(n)}(\theta)$ and $S_{\mathrm{FG2}}^{(n)}(\theta)$ are related. The following lemma is used in the proof of Theorem 4.

Lemma 3. *Let x, y, u, v and θ be positive real numbers. Then*

$$\frac{x}{(1-\theta)x + \theta y} \leq \frac{u}{(1-\theta)u + \theta v} \quad \Leftrightarrow \quad \frac{x}{y} \leq \frac{u}{v}.$$

Theorem 4. $S_{\mathrm{FG1}}^{(n)}(\theta) \leq S_{\mathrm{FG2}}^{(n)}(\theta)$ *for all $\theta > 0$.*
Proof: Let $x = a^{(n)}$, $y = 1 - d^{(n)}$, and

$$u = \sum_{i<j}^{n} a_{ij}\,, \quad \text{and} \quad v = \frac{n(n-1)}{2} - \sum_{i<j}^{n} d_{ij}.$$

Due to Lemma 3, it must be shown that

$$\frac{a^{(n)}}{1 - d^{(n)}} \leq \frac{\sum_{i<j}^{n} a_{ij}}{n(n-1)/2 - \sum_{i<j}^{n} d_{ij}}. \tag{5}$$

Inequality (5) follows from Property 2. ■

6. AVERAGES OF 2-WAY COEFFICIENTS

As shown in the previous section, instead of using the quantities $a^{(n)}$ and $d^{(n)}$, which define Bennani-Heiser coefficients, *n*-way similarity coefficients may be functions of the 2-way information. The *n*-way formulations in the previous section preserve relations between 2-way coefficients with respect to correction for chance [33]. As an alternative approach, we could also formulate *n*-way coefficients that are functions of the 2-way coefficients themselves. There are many functions that can be used to obtain a mean value of $n(n-1)/2$ coefficients, for example, the geometric and harmonic means or the root mean square. The arithmetic mean is however the most commonly used and best understood in statistics. Furthermore, in the context of 3-way distances, the arithmetic mean is analogous to the perimeter distance [10,19].

In this section we define *n*-way coefficients as the arithmetic mean of the $n(n-1)/2$ pairwise (2-way) coefficients. The arithmetic mean is the most commonly used type of average and is a natural measure of average similarity among *n* objects. Consider the following *n*-way generalization of the simple matching coefficient $S_{\mathrm{SM}}^{(2)}$ for *n* binary sequences:

$$S_{\mathrm{SM3}}^{(n)} = \frac{2}{n(n-1)} \sum_{i<j}^{n} S_{\mathrm{SM}}^{(2)} = \frac{2}{n(n-1)} \sum_{i<j}^{n} (a_{ij} + d_{ij}).$$

Coefficient $S_{\mathrm{SM3}}^{(n)}$ is the arithmetic mean of the $n(n-1)/2$ pairwise coefficients that can be formed given *n* sequences. Note that $S_{\mathrm{SM3}}^{(n)}$ is equivalent to $S_{\mathrm{SM2}}^{(n)}$, the *n*-way generalization of the simple matching coefficient from Section 5.
We consider the following *n*-way generalizations of the Jaccard coefficient $S_{\mathrm{J}}^{(2)}$ and the Dice coefficient $S_{\mathrm{D}}^{(2)}$:

$$S_{\mathrm{J3}}^{(n)} = \frac{2}{n(n-1)} \sum_{i<j}^{n} \frac{a_{ij}}{1 - d_{ij}}$$

and

$$S_{\mathrm{D3}}^{(n)} = \frac{2}{n(n-1)} \sum_{i<j}^{n} \frac{2a_{ij}}{a_{ij} + 1 - d_{ij}}.$$

We also have the following *n*-way generalizations of parameter families $S_{\mathrm{GL}}^{(2)}(\theta)$ and $S_{\mathrm{FG}}^{(2)}(\theta)$:

$$S_{\mathrm{GL3}}^{(n)} = \frac{2}{n(n-1)} \sum_{i<j}^{n} \frac{a_{ij} + d_{ij}}{\theta + (1-\theta)(a_{ij} + d_{ij})}$$

and

$$S_{\mathrm{FG3}}^{(n)} = \frac{2}{n(n-1)} \sum_{i<j}^{n} \frac{a_{ij}}{(1-\theta)a_{ij} + \theta(1 - d_{ij})}.$$

Each *n*-way coefficient and family is simply the arithmetic mean of all $n(n-1)/2$ pairwise coefficients or family functions that can be formed given *n* sequences.
Any two members of the 2-way parameter family $S_{\mathrm{GL}}^{(2)}(\theta)$, or two members of $S_{\mathrm{FG}}^{(2)}(\theta)$, are globally order equivalent (Property 1). The *n*-way generalizations $S_{\mathrm{GL3}}^{(n)}(\theta)$ and $S_{\mathrm{FG3}}^{(n)}(\theta)$ preserve Property 1, similar to families $S_{\mathrm{GL1}}^{(n)}(\theta)$ and $S_{\mathrm{FG1}}^{(n)}(\theta)$ (Property 3) and families $S_{\mathrm{GL2}}^{(n)}(\theta)$ and $S_{\mathrm{FG2}}^{(n)}(\theta)$ (Property 5).

Property 6. Two members of family $S_{\mathrm{GL3}}^{(n)}(\theta)$ and $S_{\mathrm{FG3}}^{(n)}(\theta)$, are globally order equivalent. (See also Properties 3 and 5). The result follows from the fact that the corresponding 2-way coefficient families are globally order equivalent (Property 1).

Example 3. In Example 1 we considered a data matrix for which the six 2-way Jaccard coefficients were all equal, but one 3-way Jaccard coefficient was different. Members of family $S_{\mathrm{GL2}}^{(n)}(\theta)$ and $S_{\mathrm{GL3}}^{(n)}(\theta)$ do not share this characteristic. In fact, for given θ, all *n*-way coefficients are equal if the 2-way coefficients are equal. For $S_{\mathrm{GL3}}^{(n)}(\theta)$ this is by definition. For $S_{\mathrm{GL2}}^{(n)}(\theta)$ this can be seen as follows. If $a_{ij} + d_{ij} = c$, we obtain

$$S_{\mathrm{GL2}}^{(n)} = S_{\mathrm{GL3}}^{(n)} = \frac{c}{\theta + (1-\theta)c},$$

which is a function of θ. Families $S_{\mathrm{GL2}}^{(n)}(\theta)$ and $S_{\mathrm{GL3}}^{(n)}(\theta)$ are thus not suited for detecting possible higher-order relations between the objects that cannot be discovered when one only considers the 2-way information.

Example 4. Suppose we have the following four binary sequences on ten attributes.

objects	attributes									
i	1	1	0	1	0	0	1	0	0	0
j	1	0	1	0	1	0	0	0	1	0
k	0	1	0	0	1	1	0	1	0	0
l	0	0	1	1	0	1	0	0	0	1

For these data the six 2-way Jaccard coefficients between the four objects areal equal ($S_{\mathrm{J}}^{(2)} = 1/7$). In this section the *n*-way coefficients are arithmetic means of the 2-way coefficients. Therefore, $S_{\mathrm{J}}^{(2)} = S_{\mathrm{J3}}^{(3)} = S_{\mathrm{J3}}^{(4)} = 1/7$, that is, all *n*-way coefficients ($n \geq 2$) are equal. The *n*-way coefficients discussed in this section are functions of the 2-way coefficients, and are thus not suited for detecting possible 3-way or higher-order similarity between the objects when the 2-way coefficients give no discriminative information.

For Bennani-Heiser coefficients (Section 4), the similarity between *m* sequences is always equal to or greater than the similarity between the *m* sequences and *n* – *m* other sequences (Property 4). The following example shows that the *n*-way coefficients considered in this section do not possess this property.

Example 5. Consider the data in Example 2. For these data the three 2-way simple matching coefficients between the three objects are $S_{\mathrm{SM}}^{(2)}(i,j) = 1/5$, $S_{\mathrm{SM}}^{(2)}(i,k) = 2/5$ and $S_{\mathrm{SM}}^{(2)}(j,k) = 4/5$. The 3-way simple matching coefficient, $S_{\mathrm{SM2}}^{(3)} = S_{\mathrm{SM3}}^{(3)} = 7/15$, is the arithmetic mean of the three 2-way coefficients. Furthermore, the three 2-way Dice coefficients are $S_{\mathrm{D}}^{(2)}(i,j) = 1/3$, $S_{\mathrm{D}}^{(2)}(i,k) = 4/7$ and $S_{\mathrm{D}}^{(2)}(j,k) = 6/7$. The 3-way Dice coefficient $S_{\mathrm{D3}}^{(3)} = 37/63$, is the arithmetic mean of the three 2-way coefficients. Thus, using the coefficients from this section, the amount of similarity may increase when one increases the number of sequences or objects that are compared.

The parameter families formulated in this section may be compared to the Bennani-Heiser coefficients from Section 4. It turns out that coefficients from the two formulations are bounds of one another. Theorem 5 shows how the Gower-Legendre families $S_{\mathrm{GL1}}^{(n)}(\theta)$ and $S_{\mathrm{GL3}}^{(n)}(\theta)$ are related. Lemma 1 is used in the proof of Theorem 5.

Theorem 5. $S_{\mathrm{GL1}}^{(n)}(\theta) \leq S_{\mathrm{GL3}}^{(n)}(\theta)$ *for all* $\theta > 0$.
Proof: The inequality holds if it can be shown that

$$\frac{a^{(n)} + d^{(n)}}{\theta + (1-\theta)(a^{(n)} + d^{(n)})} \leq \frac{a_{ij} + d_{ij}}{\theta + (1-\theta)(a_{ij} + d_{ij})}. \tag{6}$$

Let $x = a^{(n)} + d^{(n)}$ and $y = a_{ij} + d_{ij}$. Due to Lemma 1, inequality (6) holds if and only if

$$a^{(n)} + d^{(n)} \leq a_{ij} + d_{ij}. \tag{7}$$

Inequality (7) follows from Property 2, that is, $a_{ij} \geq a^{(n)}$ and $d_{ij} \geq d^{(n)}$. ■

Theorem 6 specifies how the Fichet-Gower families $S_{\mathrm{FG1}}^{(n)}(\theta)$ and $S_{\mathrm{FG3}}^{(n)}(\theta)$ are related. Lemma 3 is used in the proof of Theorem 6.

Theorem 6. $S_{\mathrm{FG1}}^{(n)}(\theta) \leq S_{\mathrm{FG3}}^{(n)}(\theta)$ *for all* $\theta > 0$.
Proof: The inequality holds if it can be shown that

$$\frac{a^{(n)}}{(1-\theta)a^{(n)} + \theta(1-d^{(n)})} \leq \frac{a_{ij}}{(1-\theta)a_{ij} + \theta(1-d_{ij})}. \tag{8}$$

Let $x = a^{(n)}$, $y = 1 - d^{(n)}$, $u = a_{ij}$ and $v = 1 - d_{ij}$. Due to Lemma 3, inequality (8) holds if and only if

$$\frac{a^{(n)}}{1-d^{(n)}} \leq \frac{a_{ij}}{1-d_{ij}}. \tag{9}$$

Inequality (9) follows from Property 2. ■

7. DICE COEFFICIENTS

In Warrens [36], a central role is played by the Dice coefficient $S_{\mathrm{D}}^{(2)}$. Thus far, we considered three *n*-way generalizations of $S_{\mathrm{D}}^{(2)}$:

$$S_{\mathrm{D1}}^{(n)} = \frac{2a^{(n)}}{a^{(n)} + 1 - d^{(n)}}, \qquad S_{\mathrm{D2}}^{(n)} = \frac{2\sum_{i<j}^{n} a_{ij}}{\sum_{i<j}^{n} a_{ij} + \frac{n(n-1)}{2} - \sum_{i<j}^{n} d_{ij}}$$

and

$$S_{\mathrm{D3}}^{(n)} = \frac{2}{n(n-1)} \sum_{i<j}^{n} \frac{2a_{ij}}{a_{ij} + 1 - d_{ij}}.$$

The *n*-way Dice coefficient

$$S_{\mathrm{D4}}^{(n)} = \frac{na^{(n)}}{\sum_{i}^{n} p_i},$$

is a fourth generalization of $S_{\mathrm{D}}^{(2)}$considered in Warrens [36]. Coefficient $S_{\mathrm{D4}}^{(n)}$ does not belong to any of the classes considered in Sections 4, 5 or 6. Due to Theorems 4 and 6, we have $S_{\mathrm{D1}}^{(n)} \leq S_{\mathrm{D2}}^{(n)}$ and $S_{\mathrm{D1}}^{(n)} \leq S_{\mathrm{D3}}^{(n)}$, respectively. The *n*-way coefficients $S_{\mathrm{D2}}^{(n)}$ and $S_{\mathrm{D4}}^{(n)}$ are related in the following way.

Proposition 7. $S_{\mathrm{D4}}^{(n)} \leq S_{\mathrm{D2}}^{(n)}$.
Proof: Using the identity

$$(n-1)\sum_{i=1}^{n} p_i = \sum_{i<j}^{n} \left(a_{ij} + 1 - d_{ij}\right),$$

we can write

$$S_{\mathrm{D4}}^{(n)} = \frac{n(n-1)a^{(n)}}{\sum_{i<j}^{n}\left(a_{ij} + 1 - d_{ij}\right)}.$$

Since the denominator of $S_{\mathrm{D2}}^{(n)}$ is equal to the denominator of $S_{\mathrm{D4}}^{(n)}$, we have $S_{\mathrm{D4}}^{(n)} \leq S_{\mathrm{D2}}^{(n)}$ if and only if

$$\frac{n(n-1)a^{(n)}}{2} \leq \sum_{i<j}^{n} a_{ij}. \tag{10}$$

Inequality (10) follows from Property 2. This completes the proof. ■

8. DISCUSSION

Pairwise or 2-way similarity coefficients only allow comparison of two objects at a time. Multi-way coefficients (for groups of size $n \geq 2$) may be used to compare *n* objects at a time [7,11,26,36,40]. In this paper, we compared three definitions of *n*-way similarity coefficients for *n* binary sequences. Furthermore, we discussed properties that the similarity coefficients may have in general, not just for certain data. All three definitions preserve the globally order equivalence of two coefficients (Properties 3, 5 and 6). The Bennani-Heiser coefficients defined in Section 4 possess some properties that the *n*-way coefficients based on 2-way information, considered in Sections 5 and 6, do not exhibit.

First of all, for $2 \leq m \leq n$, the *m*-way similarity of *m* binary sequences is never smaller than the *n*-way similarity between the *m* sequences and $n - m$ other sequences (Property 4). In general, the amount of similarity decreases as *n*, the number of objects compared, increases. Theoretically, this is considered a desideratum in Joly and Le Calvé [22] in the context of distance functions. However, in practice this often means that Bennani-Heiser coefficients have (very) small values for high values of *n* (n = 5, 6) or even moderate values of *n* (n = 3, 4). The *n*-way coefficients from Section 5 are based on the 2-way information and usually have a value that is intermediate of the 2-way similarities between the objects (Example 2). By definition, the value of the arithmetic mean discussed in Section 6 lies between the values of the 2-way coefficients. Furthermore, we showed that the Bennani-Heiser coefficients are bounded from above by both the corresponding *n*-way coefficients in Section 5 as well as the corresponding *n*-way coefficients in Section 6 (Theorems 2, 4,5 and 6). The *n*-way coefficients from Sections 5 and 6 thus always provide higher values.

A main motivation for formulating the Bennani-Heiser coefficients in [36] is that these *n*-way coefficients may be used to detect possible relations between the objects or sequences (Example 1) that cannot be obtained from the pairwise or 2-way information. The *n*-way coefficients from Section 5 and 6 are based on 2-way information. These coefficients provide none or little discriminative information when the 2-way coefficients give no discriminative information (Examples 3 and 4), and are thus not suited for detecting higher-order relations between the objects.

In this paper, the different *n*-way definitions of similarity for binary sequences have only been compared theoretically. For future work it should be investigated whether the various definitions also result in different outcomes in *n*-way data analysis, for example, three-way multidimensional scaling or hierarchical clustering analysis [4,19,22]. We mention the following two studies. Gower and De Rooij [17] demonstrated that 2-way and 3-way multidimensional scaling give very similar results if the 3-way dissimilarities are defined on the 2-way distances (generalized Euclidean distance, perimeter distance). Thus it appears that 3-way coefficients, when defined as functions of the 2-way coefficients, do not give more information than is already present in the 2-way coefficients. In contrast, Cox et al. [7] compared different *n*-way multidimensional scaling analyses (for different *n*) using the complement of the Bennani-Heiser coefficient $S_{\text{J1}}^{(n)}$(Jaccard coefficient). These authors illustrated that *n*-way multidimensional scaling do in fact provide different output and interpretations than ordinary 2-way multidimensional scaling.

In this paper we only considered *n*-way generalizations of the popular simple matching coefficient, the Jaccard and Dice coefficients [36,37], and two *n*-way families that generalize these three coefficients [18]. Some of the ideas presented in this paper can be applied to or may also hold for *n*-way coefficients not studied here. A variety of examples of *n*-way coefficients for binary sequences can be found in [36,40].

9. ACKNOWLEDGEMENT

This research was done while the author was funded by the Netherlands Organisation for Scientific Research, Veni project 451-11-026.

10. REFERENCES

[1]. A. N. Albatineh, M. Niewiadomska-Bugaj, and D. Mihalko. On similarity indices and correction for chance agreement. Journal of Classification, 23:301-313, 2006.
[2]. V. Batagelj and M. Bren. Comparing resemblance measures. Journal of Classification, 12:73-90, 1995.
[3]. F. B. Baulieu. A classification of presence/absence based dissimilarity coefficients. Journal of Classification, 6:233-246, 1989.
[4]. M. Bennani-Dosse. Analyses Métriques á Trois Voies, PhD Dissertation. Université de Haute Bretagne Rennes II, France, 1993.
[5]. J. Cohen. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20:37-46, 1960.
[6]. A. J. Conger. Integration and generalization of kappas for multiple raters. Psychological Bulletin, 88:322-328, 1980.
[7]. T. F. Cox, M. A. A. Cox, and J. A. Branco. Multidimensional scaling of n-tuples. British Journal of Mathematical and Statistical Psychology, 44:195-206, 1991.
[8]. J. T. Daws. The analysis of free-sorting data: Beyond pairwise comparison. Journal of Classification, 13:57-80, 1996.
[9]. M. de Rooij. Distance models for three-way tables and three-way association. Journal of Classification, 19:161-178, 2002.
[10]. M. de Rooij and J. C. Gower. The geometry of triadic distances. Journal of Classification, 20:181-220, 2003.
[11]. J. Diatta. Description-meet compatible multiway dissimilarities. Discrete Applied Mathematics, 154:493-507, 2006.
[12]. L. R. Dice. Measures of the amount of ecologic association between species. Ecology, 26:297-302, 1945.
[13]. B. Fichet. Distances and Euclidean distances for presence-absence characters and their application to factor analysis. In J. de Leeuw, W. J. Heiser, J. J. Meulman, and F. Critchley, editors, Multidimensional Data Analysis, pages 23-46. DSWO Press, Leiden, 1986.
[14]. J. L. Fleiss. Measuring nominal scale agreement among many raters. Psychological Bulletin, 76:378-382, 1971.
[15]. J. C. Gower. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika, 53:325-338, 1966.
[16]. J. C. Gower. Euclidean distance matrices. In J. de Leeuw, W. J. Heiser, J. J. Meulman, and F. Critchley, editors, Multidimensional Data Analysis, pages 11-22. DSWO Press, Leiden, 1986.
[17]. J. C. Gower and M. de Rooij. A comparison of the multidimensional scaling of triadic and dyadic distances. Journal of Classification, 20:115-136, 2003.

[18]. J. C. Gower and P. Legendre. Metric and Euclidean properties of dissimilarity coefficients. Journal of Classification, 3:5-48, 1986.
[19]. W. J. Heiser and M. Bennani. Triadic distance models: Axiomatization and least squares representation. Journal of Mathematical Psychology, 41:189-206, 1997.
[20]. A. P. J. M. Heuvelmans and P. F. Sanders. Beoordelaarsovereenstemming.In P. F. Sanders T. J. H. M. Eggen, editor, Psychometrie in de Praktijk, pages 443-470. Cito Instituut voor Toestontwikkeling, Arnhem, 1993.
[21]. P. Jaccard. The distribution of the flora in the Alpine zone. The New Phytologist, 11:37-50, 1912.
[22]. S. Joly and G. Le Calvé. Three-way distances. Journal of Classification, 12:191-205, 1995.
[23]. M.-J. Lesot, M. Rifqi, and H. Benhadda. Similarity measures for binary and numerical data: A survey. International Journal of Knowledge Engineering and Soft Data Paradigms, 1:63-84, 2009.
[24]. R. J. Light. Measures of response agreement for qualitative data: Some generalizations and alternatives. Psychological Bulletin, 76:365-377, 1971.
[25]. R. Popping. Overeenstemmingsmaten voor nominale data. Rijksuniversiteit Groningen, Groningen, 1983.
[26]. R. Popping. Some views on agreement to be used in content analysis studies. Quality & Quantity, 44:1067-1078, 2010.
[27]. T. Sørenson. A method of stabilizing groups of equivalent amplitude in plant sociology based on the similarity of species content and its application to analyses of the vegetation on Danish commons. Kongelige Danske Videnskabernes Selskab Biologiske Skrifter, 5:1-34, 1948.
[28]. R. Sibson. Order invariant methods for data analysis. Journal of the Royal Statistical Society, Series B, 34:311-349, 1972.
[29]. R. R. Sokal and C. D. Michener. A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin, 38:1409-1438, 1958.
[30]. D. Steinley. Properties of the Hubert-Arabie adjusted Rand index. Psychological Methods, 9:386-396, 2004.
[31]. M. J. Warrens. Bounds of resemblance measures for binary (presence/absence) variables. Journal of Classification, 25:195-208, 2008.
[32]. M. J. Warrens. On association coefficients for 2×2 tables and properties that do not depend on the marginal distributions. Psychometrika, 73:777-789, 2008.
[33]. M. J. Warrens. On similarity coefficients for 2×2 tables and correction for chance. Psychometrika, 73:487-502, 2008.
[34]. M. J. Warrens. On the equivalence of Cohen's kappa and the Hubert-Arabie adjusted Rand index. Journal of Classification, 25:177-183, 2008.
[35]. M. J. Warrens. On the indeterminacy of resemblance measures for binary(presence/absence) data. Journal of Classification, 25:125-136, 2008.
[36]. M. J. Warrens. *k*-Adic similarity coefficients for binary (presence/absence) data. Journal of Classification, 26:227-245, 2009.
[37]. M. J. Warrens. On Robinsonian dissimilarities, the consecutive ones property and latent variable models. Advances in Data Analysis and Classification, 3:169-184, 2009.
[38]. M. J. Warrens. Inequalities between multi-rater kappas. Advances in Data Analysis and Classification, 4:271-286, 2010.
[39]. M. J. Warrens. *n*-Way metrics. Journal of Classification, 27:173-190, 2010.
[40]. M. J. Warrens. A family of multi-rater kappas that can always be increased and decreased by combining categories. Statistical Methodology, 9:330-340, 2012.
[41]. M. J. Warrens. On the equivalence of multi-rater kappas based on 2-agreement and 3-agreement with binary scores. ISRN Probability and Statistics, 2012.
[42]. M. J. Warrens. Cohen's weighted kappa with additive weights. Advances in Data Analysis and Classification, 7:41-55, 2013.
[43]. M. J. Warrens. Conditional inequalities between Cohen's kappa and weighted kappas. Statistical Methodology, 10:14-22, 2013.