Manuscript
Accepted for publication in *Communication Methods and Measure, 2008*

Systematic and Random Disagreement and the Reliability of Nominal Data

Klaus Krippendorff
The Annenberg School for Communication
University of Pennsylvania
kkrippendorff@asc.upenn.edu

2008.2.10

Abstract

Reliability is an important bottleneck for content analysis and similar methods for generating analyzable data. This is because the analysis of complex qualitative phenomena such as texts, social interactions, and media images easily escape physical measurement and call for human coders to describe what they read or observe. Owing to the individuality of coders, the data they generate for subsequent analysis are prone to errors not typically found in mechanical measuring devices. However, most measures that are designed to indicate whether data are sufficiently reliable to warrant analysis do not differentiate among kinds of disagreement that prevent data from being reliable. This paper distinguishes two kinds of disagreement, systematic disagreement and random disagreement, and suggests measures of them in conjunction with the agreement coefficient α (alpha) (Krippendorff, 2004a, pp. 211-256). These measures, previously proposed for interval data (Krippendorff, 1970), are here developed for nominal data. Their importance lies in their ability to not only aid the development of reliable coding instructions but also warn researchers about two kinds of errors they face when using imperfect data.

Reliability – Systematic and Random Disagreements

In psychometrics it is common to distinguish between two kinds of measurement errors, systematic and random. Nunnally and Bernstein (1994, p. 213) point out that systematic errors are unimportant when studying individual differences while random errors have the undesirable effect of hiding existing relationships. Their observation pertains to responses by subjects, for example, to psychological tests. However, the situation is quite different when human coders are instructed to generate analyzable data, for example, from textual matter in content analysis, field observations in ethnographic studies, open-ended interviews in survey research, or focus groups in market assessments. In such situations, coders are not the subjects of research but the means through which phenomena of interest are turned into reliable data.

In the context of generating analyzable data, reliability is the extent to which researchers can rest assured that the distinctions they draw within the data at their disposal also distinguish among the phenomena that coders had actually experienced by observation, hearing, or reading. Such assurances can be obtained only when substantial agreement is observed among several independent coders describing the same set of phenomena following the same coding instructions under various unrelated circumstances. Unlike in studies of individual behaviors and personal attributes, when the reliability of data is the issue, *any* disagreement observed among

coders diminishes their reliability, i.e., the ability of researchers to know what they are analyzing through them.

Since there is no objective ground for determining the accuracy of data – which is the motivation for relying on coders' judgments – the term "error" is a misnomer. Therefore this paper speaks of systematic disagreements and random disagreements and proposes measures of them.

- *Systematic disagreements* exhibit some regularity and are more or less predictable. For example, when one coder favors one candidate for political office, the other favors another and both are ask to code candidates' appearance on television, they are likely to disagree but in ways that can be explained. Or, when coding instructions contain ambiguities that affect some categories more so than others, disagreement can be isolated and related to individual categories. Unless coding tasks are fairly mechanical, coder idiosyncrasies and preferences invariably enter individual judgments and bias the data.
- *Random disagreements***,** much like random noise in a communication channel, exhibit no regularity whatsoever and therefore cannot be predicted or explained. They cannot be considered biased, but fogged or blurred.

While a preponderance of agreement among coders always is the deciding criterion for accepting data as reliable, the two kinds of disagreements can affect the subsequent analyses of imperfect data unequally and knowing their respective extent is therefore important for two reasons.

First, *systematic disagreements tend to encourage Type I errors*, raising the likelihood of rejecting the null hypothesis when it should have been accepted. Systematic disagreements reveal unwarranted structures in the data, structures that are due to the coding or measuring process and extraneous to the phenomena supposedly coded. For example, coders who are prejudiced in favor of a particular hypothesis are likely to interpret the phenomena they categorize in support of that hypothesis, generating biased data. Ignorant of this distortion, the analysts of such data can be led to invalid conclusions.[1] Measuring the agreement among diverse coders establishes the reliability of data but would not distinguish between whether coders are merely variable in their conceptions of the coding task or systematically prejudiced. A separate measure of systematic disagreement could warn the researcher about the extent of this error.

Random disagreements, by contrast, *tend to encourage Type II errors*, increasing the likelihood of accepting the null-hypothesis when it should have been rejected. Random disagreement resembles the static in radio transmissions and can be visualized as obscuring an image intended to represent phenomena of interest. Randomness in data makes it difficult for researchers to identify relationships that would be apparent absent such disagreements.

Second, some systematic disagreements can be corrected provided their cause is identifiable. For example, when two coders consistently misread their coding instruction or apply two different version of it to the phenomena coded, disagreements are to some extent predictable, hence systematic. When discovered while a coding instrument is under development, a measure of that systematic disagreement can guide researchers to its source, for example, to the coder who is misreading the coding instructions and needs to be informed about it. When discovered

[1] If researchers have a stake in the hypotheses they are pursuing, coding their own data can easily create self-fulfilling hypotheses. Therefore, it is generally advisable to test reliability with coders who do not know the purpose of their coding efforts or have no or diverse preconceptions.

after data were coded, the data generated by the systematically disagreeing coder could be pulled, reexamined, and where obvious corrected.

Under the most favorable circumstances, systematic disagreements can be eliminated completely from the data, in which case their reliability improves by the amount of the systematic disagreement measured. For example, when one coder uses category *c* whenever the other uses category *k* and visa versa, once the mistaken coder is identified, that coder's categories c and k could be recoded to achieve perfect reliability. Thus, the amount of systematic disagreement is a measure of the extent to which reliability may be improved—but only in principle. In practice, finding ways to correct systematic disagreements may not be worth the effort or may be impossible.

Random disagreements, however, cannot be corrected at all. There is little the researcher can do beyond disambiguating the coding instructions, encouraging coders to be more careful in attending to all necessary details, and start the process of coding again.

Figure 1 depicts contingency matrix representations of reliability data with five typical systematic disagreements. Such matrices cross-tabulate one coder's use of categories, here four, with those of another. A ○-cell contains units coded as same by both coders. A ● represents units coded differently. Either can contain positive frequencies. All other cells are unused, empty.

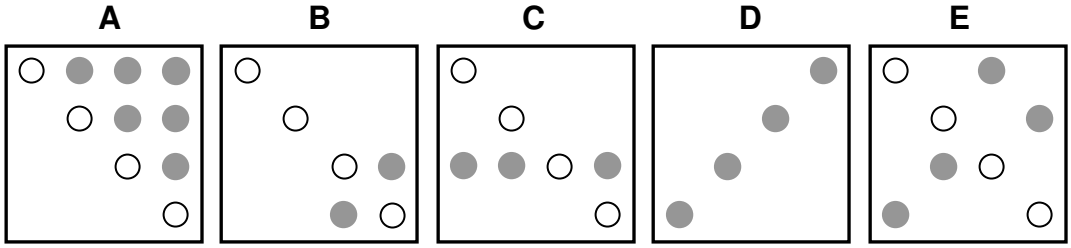

Figure 1. Five typical patterns of systematic disagreements in contingency matrices

Cases A and C exemplify systematic disagreements that result in unequal margins for the two coders. In A, units with mismatching categories occur in one of two off-diagonal triangles, the other is empty. In B, the units of one coder's category are distributed over the categories of the other coders. Either kind of systematic disagreement fails Cohen's (1960) *κ* (kappa) as an index of reliability. While κ responds to this disagreement, instead of reducing κ by a measure of it, contrary to what its proponents claim it does, κ increases when disagreements in the coders' margins are encountered. This peculiarity rewards coders who systematically disagree on their marginal distribution of categories and punishes those who do agree (Brennan & Prediger, 1981; Zwick, 1988).[2] Kappa's response attests to the importance of recognizing not only random disagreements, but systematic disagreements as well and subtracting them from indices of data reliability, not adding them.

[2] Zwick (1988) argues against the use of Cohen's (1960) κ when the margins of contingency matrices are unequal and suggests testing for marginal agreement, following Maxwell (1970) and Fleiss and Everitt's (1971) proposals, based on a procedure suggested by Stuart (1955). I take these proposals as stopgap efforts to save a coefficient that has proven inadequate as a reliability measure and an implicit acknowledgement of the importance of systematic disagreements in reliability data.

Case B shows a situation in which disagreements are limited to two categories, revealing them to be to some extent if not perfectly synonymous, i.e., equal in meaning and indistinguishable by both coders. Here, systematic disagreement is not related to differences in coder conceptions, as in case A, but to the definitions of two categories out of four. Discovering systematic disagreement of this kind justifies lumping the ambiguous categories into one, losing an unreliable distinction but gaining perfectly reliable data, here concerning the remaining three categories.

Case C demonstrates a situation in which all disagreements point to one coder who is uncertain about categorizing one kind of phenomenon that the other coder clearly distinguishes. If coding instructions are still under development, locating the problematic coder and the category that this coder fails to understand enables the researcher to instruct that coder regarding the intended meaning. Once the data are coded, the only way to salvage such data would be to exclude the questionable category from subsequent analyses.[3]

Case D depicts a perfect inversion of the list of categories used by the two coders. Their categorizations are perfectly predictable from each other, however, without agreement. After inquiring about who followed the coding instructions as intended and who worked with an inverted list, this systematic disagreement can be eliminated fully by either instructing one coder of the intended use of categories or recoding that coder's data and achieve perfect reliability.

Case E visualizes a complex pattern of disagreements, explainable in terms of interactions between the coders regarding their use of categories. It is unlikely that such systematic disagreements can be converted into more reliable data, but quantifying their extent can at least warn researchers about spurious structures in the data that could eventually mislead a researcher.

Of course, these five kinds of systematic disagreement are ideal-types. Unreliable data typically contain a blend of them. There is no guarantee that systematic disagreements can be corrected and higher reliabilities achieved thereby, but distinguishing systematic and random disagreement, even in moderately unreliability data can certainly alert researchers of their potential effects on subsequent analyses.

Reliability Data

Let the canonical form of reliability data, also called normal form, be a matrix of *m* independent coders, observers, judges or raters *i* by *r* independent units *u*, containing values *c* or *k* that different coders assign to these units. This paper is concerned with nominal data; hence the values *c* in Figure 2 are nominal values, unordered qualities or categories.

[3] Omitting categories from a variable has its costs, not only in terms of losing potentially valuable information but also making proportions difficult to compare.

Units:		1	2	3	.	.	u	.	.	.	.	.	.	.	r
Coders:	*1*	c_{11}	.	.	.	.	c_{1u}	.	.	.	.	.	.	.	c_{1r}
	:	:					:								:
	i	c_{i1}	.	.	.	.	c_{iu}	.	.	.	.	.	.	.	c_{ir}
		:					:								:
	:	:					:								:
	m	c_{m1}	.	.	.	.	c_{mu}	.	.	.	.	.	.	.	c_{mr}
		m_1	.	.	.	.	m_u	.	.	.	.	.	.	.	m_r

Where: c_{iu} is a category assigned by coder i to unit u.
$m_u \geq 2$ is the number of coders that have categorized unit u –
not all m coders need to categorize all r units.

Figure 2. Canonical form of reliability data

A canonical form of data is the most basic representation from which other forms can be derived, for example by different ways of summing its cell contents. In the following, this paper will utilize the properties of two derived matrices for representing reliability data: coincidence and contingency matrix representations, which must not be confused.

Alpha

The coefficient α (alpha)[4] (Krippendorff, 2004a, pp. 221-256) is a measure of agreement, designed as an index of the reliability of data from which a sample is multiply coded as in Figure 2. Reliability measures safeguard against the possibility that data are due to circumstances that are extraneous to the phenomena of analytical interest to the researchers, by the idiosyncrasies and instabilities of coders, for example. To estimate the degree to which researchers can rely on their data, α pairs all categories within units u that should ideally be the same, aggregates the observed disagreements over all units and expresses this aggregate relative to hypothetical data that could be expected if units were coded blindly or by chance.

One could define the agreement coefficient α in terms of the canonical form of reliability data in Figure 2, but it is easier to state α in terms of coincidence matrix representations of such data as in Figure 3.

[4] Not to be confused with Cronbach's (1951) alpha, which is a test of the consistency of subjects' responses to psychological tests. While called a reliability measure, it is not applicable to situations of multiple coders categorizing data (see Krippendorff, 2004a, pp. 249).

Categories:

	1	.	k	.	.	
1	n_{11}	.	n_{1k}	.	.	$n_{1.}$
.	.	.	.	.	.	.
c	n_{c1}	.	n_{ck}	.	.	$n_{c.}$
.	.	.	.	.	.	.
.	.	.	.	.	.	.
	$n_{.1}$	.	$n_{.k}$	.	.	$n_{..}$ = Number of categories used by all coders

Where: $n_{ck} = \sum_u \frac{number\ of\ c\text{-}k\ \ pairs\ in\ unit\ u}{m_u - 1}$

Figure 3. Coincidence matrix representation of reliability data

Note that a coincidence matrix representation of reliability data is more compact than their canonical form in Figure 2. This is achieved by eliminating references to particular coders, irrelevant to assessments of data reliability, and enumerating the pairable categories rather that listing them. Its cell contents n_{ck} are the frequencies of c-k pairs of categories found in units u, weighted by m_u to assure that each pairable category contributes exactly one to the matrix. Coincidence matrices are symmetrical around their diagonal, $n_{ck}=n_{kc}$, contain perfectly matching categories in their diagonal, n_{cc} and their marginal sums are identical, $n_{c.}=n_{.c}$, enumerating the pairable categories used by all coders. These frequencies sum to $n_{..} \leq mr$, equal to mr when the table in Figure 2 is fully occupied, less then mr when data are missing.

For nominal data and in terms of this coincidence matrix,

$$\alpha_{nominal} = 1 - \frac{D_o}{D_e} = 1 - \frac{\sum_c \sum_{k \neq c} n_{ck}}{\sum_c \sum_{k \neq c} \frac{n_{c.} n_{.k}}{n_{..} - 1}} \qquad (1)$$

Where: D_o is the observed disagreement

D_e is the expected disagreement under conditions of chance

n_{ck}, $n_{c.}$, $n_{.k}$ and $n_{..}$ are frequencies of respectively cells, marginal sums and the total in a coincidence matrix as in Figure 3.

Definition (1) expresses the observed D_o relative to the expected D_e, which is the disagreement that would be observed if coders were categorizing units without examining them, randomly drawing from the distribution of categories $n_{c.}$ *or* $n_{.k}$, which is the best estimate of the distribution of categories in the population of data whose reliability is in question. Thus, and as appropriate for reliability assessments, D_e is the disagreement when *units are statistically independent from the categories* that are to describe them. It is important to recognize that this conception of chance is unlike the *statistical independence of two coders' judgments* (or categorizations) employed by Cohen's (1960) κ, for example. The latter is common in analyzing associations in contingency tables for purposes other than reliability. Adopting this conception of chance not only limits κ to two coders[5] but more importantly, it mistakes systematic disagreements among coders (Krippendorff, 2004b).

[5] There have been several generalizations, said to be of Cohen's (1960) kappa, to multiple coders, most recently by Berry and Mielke (1988). However, their proposal not only merely repackages Krippendorff's (1970, 1980, 2004a) α for interval data and multiple coders, it also crucially deviates from Cohen's (1960) insistence that κ express

These two ways of conceptualizing chance have their origin in the above mentioned difference between measuring the random error in individuals' performance on psychological tests and estimating the reliability of data. Scott's (1955) π (pi) was the first coefficient developed for the latter situation and α turned out to be a generalization of π. Cohen (1960) failed to appreciate that difference – unjustly criticizing Scott for not conforming to psychometric customs – and replaced π's expected agreement by the agreement obtained under condition of statistical independence between two raters' judgments.

Back to (1), when *units are categorized unanimously*, α=1. When *no correlation exists between the units and their categorization*, α=0. Alpha can assume negative values when coders consistently agree to disagree, as in case D of Figure 1, follow different coding instructions or have conflicting understanding of them. However, this should not happen when reliability data are generated appropriately, Under normal conditions, the worst scenario is agreement by chance, as described above, that is $\alpha=0 \pm$ an asymptotically disappearing value as samples increase in size.

While this paper is limited to nominal data, it needs to be understood that α is far more general. It yields separate measures for nominal, ordinal, interval, ratio, and other kinds of data[6] and is applicable to any number of coders, incomplete data, and small sample sizes, including for unitizing data (Krippendorff, 2004a, pp. 211-256). The distinction between systematic and random disagreement for interval data is available since Krippendorff (1970). Software described by Hayes and Krippendorff (2007) also produces confidence limits for α. Confidence limits are important as statistical tests of the null hypotheses, demanded by Berry and Mielke (1988) among others, are quite meaningless as they fail to inform situations in which reliability matters. Reliable data must not merely deviate from chance, α=0, they should only minimally deviate from perfect agreement, α=1.

One interpretation of α, important here, stems from the way it accounts for the distribution of coincidences in Figure 3. All agreement coefficients for nominal data distinguish between matching and non-matching category assignments. However, α explains the reliability data in a coincidence matrix in terms of α times a coincidence matrix containing only perfect agreements plus $(1-\alpha)$ times a coincidence matrix with all categories randomly chosen, see Figure 4 (Krippendorff, 2004a, pp. 226-227).

Figure 4. Alpha's decomposition of coincidence matrices

---

agreement relative to the statistical independence of two coders' judgments. Instead, Berry and Mielke adopt the statistical independence of units and their categorization as a baseline, without acknowledging its origin in Scott's (1955) π and its deviation from Cohen's (1960) κ. Thus, their generalization is not of Cohen's κ but of Scott's π, following Krippendorff 's (1970, 1980, 2004a) α. Siegel and Castellan (1988) also note this confusion of baselines in the literature, without recognizing that the presence or absence of systematic disagreement is the root of their difference. It is not impossible to define the statistical independence among three or more coders, but would be irrelevant for assessing data reliability.

[6] http://www.asc.upenn.edu/usr/krippendorff/dogs.html, accessed 2008. 2.7, includes difference functions for two additional kinds of data.

Accordingly, α is interpretable as the degree to which the observed reliability data resemble the ideal of perfect agreement as opposed to its absence, the condition of coding by chance as defined above.

The three coincidence matrices have the same marginal sums, $n_{c.}$ and $n_{.k}$, and totals, $n_{..}$, of course, and (1) and Figure 4 suggest that the sums of diagonal and off-diagonal cell contents in the matrix of the observed coincidences satisfy

For *diagonal entries*:
$$\sum_c n_{cc} = \sum_c \left[\alpha n_{c.} + (1-\alpha) \frac{n_{c.}(n_{c.}-1)}{n_{..}-1} \right] \qquad (2a)$$

For *off-diagonal entries*:
$$\sum_c \sum_{k \neq c} n_{ck} = \sum_c \sum_{k \neq c} \left[(1-\alpha) \frac{n_{c.} n_{.k}}{n_{..}-1} \right]. \qquad (2b)$$

Measuring Systematic Disagreement

If α is a measure of agreement, and interpretable as the degree to which data are sufficiently reliable to warrant drawing conclusions from them, its complement, *(1–α)*, the proportion of disagreements D_o/D_e, now needs to be decomposed into systematic and random disagreement. Let

$$\alpha + \sigma + \rho = 1 \qquad (3)$$

Where: α (alpha) measures the agreement in reliability data
σ (sigma) measures the systematic disagreement in these data
ρ (rho) measures the random disagreement in these data

Defining α in terms of coincidence matrices is convenient because they (i) summarize reliability data in terms of the collective use of categories, omitting meaningless references to individual coders, and representing in their margins the best estimate of the distribution of categories in the data; and (ii) depict *all* disagreements that reduce reliability in their off-diagonal cells, without, however, making the separation between systematic and random disagreements transparent. To distinguish the latter, one needs to consult contingency matrix representations of reliability data, as in Figure 5, one for each pair of coders.

One coder's categories:	1	.	k	.	.	
Another coder categories: 1	x_{11}	.	x_{1k}	.	.	$x_{1.}$
.	.	.	.	.	.	.
.	.	.	.	.	.	.
c	x_{c1}	.	x_{ck}	.	.	$x_{c.}$
.	.	.	.	.	.	.
	$x_{.1}$	.	$x_{.k}$	.	.	$x_{..}$ = Number of units categorized by two coders

Where x_{ck} = The number of units categorized as c by the first coder and as k by the second.

Figure 5. Contingency matrix representation of reliability data

To avoid confusion, coincidence matrices cross-tabulate the *categories* used by *m* coders; contingency matrices cross-tabulate the *units* categorized by *two* coders. In coincidence matrices both kinds of disagreements are summed in their off-diagonal cells; in contingency matrices

systematic disagreements are recognizable in their off-diagonal cells. In coincidence matrices marginal sums are equal, in coincidence matrices they are not, $x_{c.} \neq x_{.c}$. To obtain α requires one coincidence matrix for m coders; to obtain systematic disagreements calls for $m(m-1)/2$ contingency matrices. For two coders, the total number $n_{..}$ of categories is twice the number $x_{..}$ of units they coded. When data are missing the proportion $x_{..}/n_{..}$ can vary across coder pairs.

Systematic disagreement becomes evident in deviations of the observed contingencies x_{ck} from what would be expected if disagreements were randomly distributed while preserving α. For two coders, this expectation forms a contingency matrix, as in Figure 5, but containing frequencies e_{ck} according to (4):

$$e_{cc} = \frac{x_{..}}{n_{..}}\left[\alpha n_{c.} + (1-\alpha)\frac{n_{c.}(n_{c.}-1)}{n_{..}-1}\right] \quad \text{and} \quad e_{ck} = \frac{x_{..}}{n_{..}}\left[(1-\alpha)\frac{n_{c.}n_{.k}}{n_{..}-1}\right]. \tag{4}$$

Apparently, the expressions in the angular parentheses in (4) are identical to those in (2). They define the two kinds of cell contents – diagonal and off-diagonal respectively – in a hypothetical coincidence matrix (for m coders), being composed of α times the cell contents in a matrix with perfect matches plus $(1-\alpha)$ times the cell contents in a matrix in which units and their categorizations are statistically independent – as in Figure 4. The proportion $x_{..}/n_{..}$ corrects for two phenomena, (i) the fact that coincidence matrices, enumerating the categories used by m coders, contain larger totals than contingency matrices, counting pairwise judged units, and (ii) the possibility of missing data, i.e., that coders categorized unequal but overlapping subsets of the r units, causing $x_{..}$ to vary across coder pairs. Equation (4) assures the baseline for each pair of coders to equal α for m coders.

An observed disagreement is systematic whenever $x_{ck} \neq e_{ck}$. Quantitatively and for one pair of coders, systematic disagreement is evident in above zero χ^2-values

$$\chi^2 = \sum_c \sum_k \frac{(x_{ck} - e_{ck})^2}{e_{ck}}. \tag{5}$$

Since α enumerates pairwise disagreements, obtaining the systematic disagreement σ for m coders requires consideration of $m(m-1)/2$ χ^2-values. Thus, the systematic disagreement for m coders is defined by the proportion:

$$\sigma = (1-\alpha)\sqrt{\frac{\sum_{all\ pairs\ of\ coders}\chi^2}{\sum_{all\ pairs\ of\ coders}\chi^2_{max}}}, \tag{6}$$

and, according to (3), the proportion of random disagreements becomes the remainder:

$$\rho = 1 - \alpha - \sigma. \tag{7}$$

The square root of $\sum\chi^2/\sum\chi^2_{max}$ resembles Cramér's V, a popular statistics for measuring associations in nominal data. Unfortunately, Cramér's (1946, p. 282) approximation to χ^2_{max} is based on assumptions that do not apply here. Therefore, using the correct value of χ^2_{max} is preferable. It is found in a contingency matrix with frequencies w_{ck} that exhibit the largest possible systematic disagreement under the constraints of the row and column sums $x_{c.}$ and $x_{.k}$ of the original reliability data. In simple situations, w_{ck}'s are easily obtained by hand. However, constructing such matrices can become tedious when they are large and pairs of coders

are many. For these situations the following algorithm for obtaining maximum systematic disagreements will prove useful.

Algorithm for obtaining maximum systematic disagreements:

Given a contingency matrix of reliability data containing cell frequencies w_{ck}. Initially, $w_{ck} = x_{ck}$. The algorithm changes its cell contents while preserving its marginal sums $w_{c.} = x_{c.}$ and $w_{.k} = x_{.k}$.

(1) For any two diagonal cells $w_{cc} > 0$ and $w_{kk} > 0$:
Reset $w'_{cc} = w_{cc} - min(w_{cc}, w_{kk})$
$w'_{kk} = w_{kk} - min(w_{cc}, w_{kk})$
$w'_{ck} = w_{ck} + min(w_{cc}, w_{kk})$
$w'_{kc} = w_{kc} + min(w_{cc}, w_{kk})$

(2) Proceed with (1) until its condition for resetting is no longer applicable.

(3) If all cells $w_{cc} = 0$: Go to (6).

(4) For the remaining diagonal cell $w_{cc} > 0$, select any off-diagonal cell $w_{ab} > 0$; $a \neq c$ and $b \neq c$:
Reset $w'_{cc} = w_{cc} - min(w_{ab}, w_{cc})$
$w'_{ab} = w_{ab} - min(w_{ab}, w_{cc})$
$w'_{cb} = w_{cb} + min(w_{ab}, w_{cc})$
$w'_{ac} = w_{ac} + min(w_{ab}, w_{cc})$

(5) Proceed with (4) until w_{cc} is no longer reducible.

(6) For any two off-diagonal cells $w_{ab} > 0$ and $w_{cd} > 0$; $a \neq c$ and $b \neq d$ and NOT($w_{ad} = w_{cb} = 0$):
Reset $w'_{ab} = w_{ab} - min(w_{ab}, w_{cd})$
$w'_{cd} = w_{cd} - min(w_{ab}, w_{cd})$
$w'_{ad} = w_{ad} + min(w_{ab}, w_{cd})$
$w'_{cb} = w_{cb} + min(w_{ab}, w_{cd})$

(7) Proceed with (6) until its condition for resetting is no longer applicable or the resetting becomes cyclical. When cyclical, stop at the configuration that yields the largest χ^2:
$$\frac{(w_{ab} - e_{ab})^2}{e_{ab}} + \frac{(w_{cd} - e_{cd})^2}{e_{cd}} + \frac{(w_{ad} - e_{ad})^2}{e_{ad}} + \frac{(w_{cb} - e_{cb})^2}{e_{cb}}.$$

The resulting contingency matrix contains a distribution of frequencies w_{ck} of maximum systematic disagreement.[7]

Once this contingency matrix is constructed, the largest χ^2-value becomes

$$\chi^2_{max} = \sum_c \sum_k \frac{(w_{ck} - e_{ck})^2}{e_{ck}}. \qquad (8)$$

with which σ can be obtained from (6) and ρ from (7).

This concludes the development of the quantitative distinction between systematic and random disagreements for nominal data.

[7] In the absence of formal proof, this claim is limited to conditions explored empirically.

Three Numerical Examples

Figure 6 goes through three numerical examples of two coders categorizing of 60 units each with disagreements conforming to cases A, C, and D from Figure 1.

A **C** **D**

Coder-by-coder **contingency** matrices x_{ck}:

6	6	6	6	24	11				11				15	15
	6	6	6	18		11			11			15		15
		6	6	12	8	8	3	8	27		15			15
			6	6				11	11	15			.	15
6	12	18	24	60	19	19	3	19	60	15	15	15	15	60

Category-by-category **coincidence** matrices n_{ck}:

12	6	6	6	30	22		8		30				30	30
6	12	6	6	30		22	8		30			30		30
6	6	12	6	30	8	8	6	8	30		30			30
6	6	6	12	30			8	22	30	30				30
30	30	30	30	120	30	30	30	30	120	30	30	30	30	120

α=0.207 **α=0.471** **α= –0.322**

Coder-by-coder **contingency** matrices e_{ck}**, expected under conditions of α**:

6	3	3	3	15	9	2	2	2	15		5	5	5	15
3	6	3	3	15	2	9	2	2	15	5		5	5	15
3	3	6	3	15	2	2	9	2	15	5	5		5	15
3	3	3	6	15	2	2	2	9	15	5	5	5		15
15	15	15	15	60	15	15	15	15	60	15	15	15	15	60

χ^2**=36** χ^2**=77.33** χ^2**=120**

Coder-by-coder **contingency** matrices with **largest systematic disagreements** w_{ck}:

			24	24		11			11			15		15
		18		18				11	11				15	15
	12			12	19			8	27	15				15
6				6		8	3		11		15			15
6	12	18	24	60	19	19	3	19	60	15	15	15	15	60

χ^2_{max}**=300** χ^2_{max}**=310** χ^2_{max}**=120**

The results:

Agreement	**α = 0.207**	**α = 0.471**	**α = –0.322**
Systematic disagreement	**σ = 0.275**	**σ = 0.254**	**σ = 1.322**
Random disagreement	**ρ = 0.518**	**ρ = 0.275**	**ρ = 0.000**

Figure 6. Numerical comparison of three reliability data with typical systematic disagreements

In example A, systematic disagreement is apparent by occurring in only one triangle of the contingency matrix containing x_{ck}, the other being empty. Random disagreement is evident in the uniform distribution of frequencies in that occupied triangle. There is some agreement due to the fact that the diagonal entries are slightly larger than chance. One may appreciate this fact in the contingency matrix containing e_{ck}, which omits all systematic disagreement. Its diagonal frequencies are twice as large as its off-diagonal frequencies. The resulting agreement measure α and the amount of systematic disagreements σ would have to be rather small when compared with the random disagreements ρ. Their numerical values confirm this intuition.

As suggested above, Figure 6 demonstrates that contingency matrices containing frequencies x_{ck} are capable of representing systematic disagreement whereas coincidence matrices containing frequencies n_{ck} are designed not to. Since α is defined in terms of the latter, this graphically demonstrates that α does not differentiate between the two kinds of disagreement.

In example C, disagreements are limited to one coder who almost randomly assigns the available categories to a category that the other coder unambiguously distinguishes. Quantitatively, the systematic disagreement σ (limited to one row) and random disagreement ρ (categories are distributes almost randomly in that row) are about the same. With three categories matching perfectly and one matching less than by chance, α is larger than in example A. As already suggested, if the intended analysis of these data allows the confusing category to be omitted from the data, the remaining three categories would have perfect reliability.

Also, as mentioned when discussing cases A and C in Figure 1, Cohen's κ adds a measure of systematic disagreement to the agreement instead of subtracting it from that measure, therefore exceeding Krippendorff's α whenever it this disagreement is evident. In numerical example A, κ=0.250 **>** α=0.207 and in C, κ=0.502 **>** α=0.471. If sample sizes were large, to which κ is limited, this difference would be more pronounced: in A κ=0.250 **>** α=0.200 and in C, κ=0.502 **>** α=0.467. This demonstrates numerically, κ's failure to adequately account for systematic disagreements of these kinds, and κ's insensitivity to small sample sizes.

Example D represents a situation in which two coders' category assignments do not match even once. This exemplifies the condition of maximal disagreement. With the two coders' categorizations being perfectly predictable from each other, showing no uncertainty and no randomness, random disagreement is absent and ρ measures zero indeed. It is important to remember that agreement below chance is indicated by negative α-values. In this example, agreement is not only below chance, it is totally absent, which causes α to assume its most extreme negative value, –1.322 (for four categories and n= 120), and because all disagreements are perfectly systematic, systematic disagreement σ measures +1.322, fully accounting for all disagreements in the data. Note that for example D there are six possible distributions of frequencies w_{ck}, including the original frequencies x_{ck}. They do not affect, however, the χ^2_{max}-value and the systematic disagreement σ. As discussed with the help of Figure 1, researchers can take advantage of this systematic disagreement by either instructing the mistaken coder or transforming that coder's data after they were generated and obtain perfectly reliable data.

As already suggested, not all systematic disagreements lend themselves to such dramatic improvements of data reliability. In example A, for example, researchers may have a hard time finding instructions that would help coders to correct their habit of systematic disagreeing with each other. In that example, random disagreement ρ is also large, 0.518, and improvements would not be worth the trouble. Even under the most favorable conditions reliability could not exceed $\alpha+\sigma$=0.482. Case E in Figure 1 shows the juxtaposition of two patterns, frequencies in four diagonal cells and frequencies in four off-diagonal cells. The former counts toward agreement, the latter toward

systematic and random disagreement. Both are measurable but unlikely correctable. Generally, while the systematic disagreement measure σ cannot guarantee that data reliability is improvable, it unquestionably measures spurious structures in the data and can therefore warn researchers that the use of these data could cause Type I errors in their findings.

Reliability should always be assessed with data on all *m* coders' categorizations. The same is advisable when systematic disagreement σ and random disagreement ρ are obtained to see what prevents coding from reaching perfect reliability. But for diagnostic purposes one may want to obtain all three measures separately for each coder pair and for each category and examine the contingency matrices when disagreements are high. Since all three measures express average agreements or disagreements, they could hide bad coders or categories in a majority good ones.

Concluding Comments

This paper distinguishes between systematic and random disagreements for the nominal version of Krippendorff's α-agreement. Distinguishing these disagreements can turn the dread of unreliable data into (i) indications about the extent to which their reliability might be improved by either creating better coding instructions or transforming the data such that the systematic disagreement is reduced and (ii) warnings concerning the possibility two kinds of statistical errors that users of unreliable data could face when drawing conclusions from them.

Distinguishing these disagreements seems computationally complex compared to the simplicity of obtaining α by (1) from coincidence matrices as in Figure 3. This appearance is due in part to the effort of presenting the numerical ideas in this paper graphically. An efficient procedure for computing the two disagreements can bypass the construction of contingency matrices used here. This paper is an invitation to software developers to add the quantities of σ and ρ to computations of the reliability statistics α for nominal data. Their usefulness, I hope, has been demonstrated.

References

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

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

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

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton NJ: Princeton University Press.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, pp. 297-334.

Fleiss, J. L., & Everitt, B. S. (1971). Comparing the marginal totals of square contingency tables. *British Journal of Mathematical and Statistical Psychology, 24*, 117-123.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*, 77-89.

Krippendorff, K (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement, 30*, 61-70.

Krippendorff, K. (2004a). *Content Analysis: An Introduction to Its Methodology. 2nd Edition*. Thousand Oaks CA: Sage.

Krippendorff, K. (2004b). Reliability in content analysis: some common misconceptions and recommendations. *Human Communication Research, 30*, 411-433.

Maxwell, A. E. (1970). Comparing the classification of subjects by two independent judges. *British Journal of Psychiatry, 116*, 651-655.

Nunnally, J. C. & Bernstein, I. H (1994). *Psychometric Theory, 3rd ed*. New York: McGraw-Hill.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly, 19*, 321-325.

Stuart, A. (1955). A test of homogeneity of marginal distributions in a two-way classification. *Biometrika, 42*, 412-416.

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin, 103*, 347-387.