Revised version of the paper published in
Communication Methods and Measures 5.2, 1-20, 2011

Agreement and Information in the Reliability of Coding

Klaus Krippendorff
University of Pennsylvania

Coefficients that assess the reliability of data making processes – coding text, transcribing interviews, or categorizing observations into analyzable terms – are mostly conceptualized in terms of the agreement a set of coders, observers, judges, or measuring instruments exhibit. When variation is low, reliability coefficients reveal their dependency on an often neglected phenomenon, the amount of information that reliability data provide about the reliability of the coding process or the data it generates. This paper explores the concept of reliability, simple agreement, four conceptions of chance to correct that agreement, sources of information deficiency, and develops two measures of information about reliability, akin to the power of a statistical test, intended as a companion to traditional reliability coefficients, especially Krippendorff's (2004, pp. 221-250; Hayes & Krippendorff, 2007) *alpha*.

INTRODUCTION

The need of research to be reliable requires no justification. Especially in content analysis and similar research techniques that make use of human coders to generate data from texts or observations, testing the reliability of the coding process is a common methodological requirement. In reliability tests, at least two, ideally many carefully instructed, independently working, and interchangeable coders distinguish among a set of units of analysis by associating them with various values, categories, scale points, or measurements. This duplication of the coding effort gives rise to reliability data that afford assessing agreements or disagreements with the aim of inferring the reliability of the coding instructions and/or of the population of data they generate.

This paper responds to a frequently noted puzzle. In certain situations, chance-corrected agreement measures, which are standard ways to infer the reliability of coding processes, can be small, zero, even negative, while observed agreement seems high. Even agreement observed well above chance may not necessarily lead researchers to trust the reliability of their data. For example, suppose two therapists diagnose two different sets of three out of 100 patients as paranoid. Their agreement – 94% on not diagnosing paranoia – seems overwhelming but does not make much sense as there is no agreement regarding who is paranoid. But suppose the two therapists would identify the same set of three patients as paranoid. Then agreement would be perfect, although one would not know why 97% of the patients are in the mental hospital. The reliability of the diagnoses needs to be questioned on account of the rather few cases on which the therapists agree on a diagnosis. The solution lies in a clearer understanding of the information that researchers require to trust their data. Most recommendations for assessing reliability gloss over this dimension of judgment, which is highlighted here.

WHAT IS RELIABILITY?

In short, reliability is the extent to which different methods, research results, or people arrive at the same interpretations or facts. Inasmuch as data or conclusions obtained by different means can agree with each other but be wrong, reliability is only a prerequisite to validity. It cannot guarantee it.

Measurement Theory

Measurement theory offers one of the more rigorous formulations of reliability. It suggests that a measuring instrument is reliable when it is *not affected by variations in the extraneous circumstances of the measuring process*. Especially when human coders take the place of measuring devices, evaluating, categorizing, scaling, or judging the objects or units of analytical interest they face, numerous extraneous circumstances may affect the outcome of the coding process, for example, lack of understanding the coding task, idiosyncratic habits, or personal interest in the outcome of the research. Obviously, extraneous variations pollute a measuring process, and reduce the ability of researchers to rely on the data it generates.

Measurement theory is likely to confuse reliability and validity when it claims able to separate "true" variation from measurement errors. Truths concern validity, not reliability. Validity can be established by comparing a measuring device with an objective standard. Reliability, by contrast, is obtained by comparing measures of the same phenomena obtained under different circumstances or from different devices assumed to measure the same. It is only in such comparisons that the two variations can be separated: Variations in the phenomenon of interest are manifest in the agreement among these devices, whereas extraneous variations can be attributed to the disagreement observed among them. In content analysis and related social science methods for generating analyzable data, reliability is inferred from the agreement observed among independently working but otherwise interchangeable coders. It is important to notice that both, reliability and validity, are a function of and do require variations.

Interpretation Theory

Interpretation theory provides a more comprehensive and therefore more compelling conception of reliability. Literally, reliability means *the ability to rely on known interpretations by others*. To rely on the data that human coders generate, researchers must be able to reconstruct the distinctions that coders made among what they were describing, transcribing, or recording in analyzable terms. Moreover, to trust research results, users must be able to understand what the researchers were analyzing and how. This is facilitated by being explicit about the analytical process, especially about the categories that the coders were instructed to apply when recording their observations or readings, i.e., generating data. Coding instructions not only delegate researchers' conceptions to coders, making coders their proxies, but also allow the users of research results to reconstruct what coders saw, which requires a shared understanding of these instructions. Here, shared understanding is operationalized as the reproducibility of the distinction that coders were instructed to make among a diversity of phenomena or units of analysis. When coders are carefully chosen to represent the interpretive community of those who need to handle their data, then agreement among coders can be generalized to those who have a stake in the research. Without that agreement, researchers cannot be sure of what they are talking about and the users of research results have no reason to trust researcher's findings.

Although these two approaches differ conceptually, they both rely on assessing agreement in the face of variations in the objects or phenomena among which coders are asked to draw informed distinctions by evaluating, categorizing, scaling, or measuring them. Since perfection is rare, reliability is conceived of as a scale on which perfect reliability and its absence are two crucial reference points.

Psychometric Theory

Psychometric Theory provides a third conception of reliability, inadequate for assessing the quality of coding but heavily used in psychometric research (Nunnally & Bernstein 1994, p. 213). In this research tradition, reliability is not a function of observed agreement among judges' rating a set of individuals, usually by means of a scale. Rather, "high interrater reliability means that the relation of one rated object to other rated objects is the same across judges, even though the absolute numbers used to express this relationship may differ from judge to judge. Interrater reliability (so the authors conclude) is sensitive only to the relative ordering of the rated objects" (Tinsley & Weiss, 2000, p.98), often measured by Cronbach's (1951) alpha (not to be confused with Krippendorff's *alpha (α)* which will be referred to in the remainder of the paper).

PROPERTIES OF PERCENT AGREEMENT

Percent or simple agreement, i.e., the proportion of the number of units of analysis on which two coders' categorizations, scale values, or measurements match perfectly to the total number of units coded, is easy enough to understand and obtain. However, its obviousness hides the properties that make percent agreement an inadequate measure of reliability. Part of the above mentioned puzzle of observing large agreements while measuring very low, even zero reliabilities stems from not being clear about the properties of percent agreement:

It is affected by the number of values or categories available for coding. The more values among which a coder can choose and the more information the data can provide, the more difficult is it to achieve high percent agreement.

Its zero value does not indicate the absence of reliability but is evidence of maximum disagreement. Zero percent agreement is statistically rare and occurs either when coders cooperate in choosing unlike values, thus violating the requirement of working independently of each other, or follow incompatible coding instructions.

It has no fixed value that could indicate when agreement occurred merely by chance, a condition that is commonly equated with the complete absence of reliability. By chance alone, with two values, one would expect *at least* 50% agreement, with four 25%, with ten 10%, etc. In other words, percent agreement means something altogether different when there are two values as opposed to many. Moreover, "*at least*" means that when values occur with unequal frequency, expected agreement could be much higher than the percentages mentioned, leaving researchers at a loss of how to interpret observed agreement other than 100%.

Worse, as coders are asked to make informed distinctions among diverse phenomena, when variability is lacking and coders do not show evidence of making such distinctions, agreement is 100% but hardly indicative of perfect reliability, as will be discussed below.

Furthermore, percent agreement can be calculated only for two coders (save for averaging percentages among different pairs of codes, which has questionable interpretations). It is limited to nominal or categorical data among which coders' distinctions can be determined as either same or different. When we are puzzled observing that agreement is high while a reliability coefficient turns out to be unacceptably low, then we are giving the inadequate measure of percent agreement more weight than it deserves.

AGREEMENT BY CHANCE

For the above reasons, indices of the reliability of coding processes must express agreement relative to what could happen by chance, the condition at which reliability is absent, and the zero-point of a valid reliability scale. However, what uninformed proponents of particular agreement coefficients largely overlook is that there are several conceptions of chance, some of them inadequate for expressing the reliability of coding. Let me describe four, 1, 2, 3a, and 3b:

1. *Logical independence of categories.* The idea of expressing chance by the logical probability of co-occurring coding categories of a recording instrument goes back to Bennett, Alpert, and Goldstein's (1954) reliability coefficient *S*. The authors argued that the difficulty of achieving agreement increases with the number of categories available for coding, and in the absence of knowing the population proportions of various categories of coded data, the logical probability of categories is the only justifiable baseline to correct the percent of observed agreement. This argument has fuelled at least five reinventions or variations of *S* – Guilford's *G* (Holley & Guilford, 1964), Maxwell's (1970) (random error) coefficient *RE*, Janson and Vegelius's (1979) *C*, Brennan and Prediger's (1981) κ_n, and Perreault and Leigh's (1989) intercoder reliability coefficient I_r. Inasmuch as the logical probability of a category is the inverse of the number of categories available for coding, reliability coefficients based on this conception of chance are biased by the number of categories provided by a coding instrument. In fact, researchers can manipulate the values of reliability coefficients based on this conception of chance by adding unused categories to the calculations. Statistical conceptions seek to overcome this bias. The argument that it is more difficult to achieve agreement the more categories are available for coding is correct, of course, but building this conception of chance into a reliability coefficient confounds the reliability of coding processes and the amount of information about the reliability of coding instruments. We shall treat them separately below.

2. *Statistical independence of coders.* In the tradition of analyzing correlations among variables, Cohen's (1960) *kappa* defines chance or expected agreement as the agreement that would be observed if two coders' behaviors were statistically unrelated, regardless of how often they used the available values. By not counting coders' unequal proclivity for values available for coding as unreliable, *kappa*'s definition fails to treat coders as interchangeable. It numerically rewards them for not agreeing on their use of values and punishes those that do agree (Zwick, 1988). Using the metaphor of randomly drawing balls from an urn, *kappa* randomly draws from *two* urns, one containing the values used by one coder in judging the common set of units and one containing the values used by the other coder, and then computing the probability of matching values. Statistical independence of coders is only marginally related to how

units are coded and data are made and does not yield valid coefficients for assessing the reliability of coding processes or the data they generate.

3. *Statistical independence of the units coded and the values used to distinguish among them.* In this conception of chance, coders are interchangeable and, hence, represented not as individuals but by the values they use to account for the common set of units coded. Without knowledge of the correct valuation of units, this conception takes the distribution of values that all coders collectively use to describe a given set of units as the best estimate of what the population of units is like. To continue with the urn metaphor, it puts all pairable values into *one* urn, draws pairs of values from that urn, and obtains the probability of agreement or disagreement from repeated drawings. Evidently, by putting all pairable values used by all coders into one urn, coders are taken to be interchangeable, as they should be whenever the reliability of a coding process is the issue. However, this conception of chance leads to a further distinction, whether pairs of values are drawn from that urn with or without replacement.

3a. *With replacement.* Here, one value is drawn from the urn, noted and returned. A second value is drawn from that urn, its agreement or disagreement with the former is recorded and the value is returned. When repeated, this procedure leads to an expected agreement that *includes*, however, the possibility of pairing values with themselves. Scott's (1955) *pi* as well as Pearson et al.'s (1901) *intraclass correlation* coefficient – entering all observed pairs of values twice, as *a-b* and *b-a* into his product-moment correlation – follow this procedure implicitly and are, hence, appropriate only when the sample sizes of values are infinite or at least very large in which case the probability of self-matching becomes small.

3b. *Without replacement.* Here, two values are drawn from the same urn. Their agreement or disagreement is recorded and both values are returned. When repeated, this procedure leads to an expected agreement or disagreement that *excludes* the possibility of pairing values with themselves. In effect, this expected agreement or disagreement responds to the size of the sample of values and asymptotically approximates the agreement expected when sample sizes are infinite. Hence, the expected agreement obtained by drawing pairs of values without replacement is more general and consistent with the way observed agreement is established. Calculating expected agreement or disagreement without replacement is implicit in Krippendorff's (2004, pp. 221-250) *alpha*, which renders that reliability coefficient, sensitive to the size of the sample of values used by all coders.

It follows that *kappa*'s expected agreement is entirely inadequate for assessing the reliability of coding. It puts the zero-point of its reliability scale on a coding-unrelated location, as noted elsewhere and expressed in various terms (Brennan & Prediger, 1981; Krippendorff, 1978, 2008; Zwick, 1988). Besides Cohen's (1960) and Cronbach's (1951) coefficients, there are others that correct percent agreement by baselines that do not model coding processes, for example, Pearson's product-moment coefficient r_{ij}, Goodman and Kruskal's (1954) $lambda_r$ and Lin's (1989) coefficient of concordance r_c, to name only a few. The expected agreements or disagreement of Krippendorff's *alpha*, Scott's *pi*, and Person's *intraclass correlation* do reflect the absence of a statistical relation between a set of units and how they are coded but they differ in whether chance is obtained with or without replacements, which is to assume infinite or finite sample sizes.

Alpha is the most general coefficient among them, not only because responding to unequal sample sizes includes *pi* and the *intraclass correlation* as special cases, but also for being applicable to any number of coders, not just two, missing values, and different metrics (levels of measurement) – nominal, ordinal, interval, ratio, polar, and circular (Krippendorff, 1993) – and subsuming and extending other familiar statistics (Krippendorff, 2004, pp. 244-250).

SOURCES OF INADEQUATE INFORMATION ABOUT RELIABILITY

The above mentioned puzzle suggests at least three ways information about reliability could be lacking in the data from which reliabilities are obtained: the use of default categories, the lack of variability, and inadequate sample sizes.

The Use of Default Categories

Largely from the tradition of measurement theory, demanding that *all units of analysis* in a sample be measured by the same instrument, without omission, stems the requirement that the values of variables need to be not only mutually exclusive in definition but also exhaustively applicable to all units to be coded. Mutual exclusivity is to assure the researcher that the available values adequately distinguish among the units of analysis. Agreement measures verify the extent of mutual exclusivity empirically. If values were ambiguous, one would not know for sure how a particular unit was perceived or read by coders. Information is always linked to drawing distinctions.

Exhaustiveness is to assure that no unit is omitted from a study on account of the coding instrument having no place for it. Unless a variable is logically tight, the exhaustiveness of the values of a variable tends to be achieved by adding a default, blank, catch-everything-else, or "not applicable" category to the variable. In survey research, interviewees may not respond to a question and this fact must be recorded, for example as "no answer," or when a question does not apply, "not applicable." In content analysis, the category "other" or "none of the above" is common. For example, recording the gender of fictional characters on TV as male or female seems logically tight, but leaves open the possibility of actors whose gender is unknown, whether because they appear only briefly on the screen, are non-gendered (e.g., being referred to as children, neighbors, or voters), have changed their gender, or have none, as is typical of robots, comic characters or animals appearing on the screen. Introducing the category "all other," "unknown," "not applicable," "no response," "absent," "missing," or "999" admits a category to a variable that contains no useful information for the subsequent analysis of these data.

An anecdote may suffice: In preparing for a study of television characters, we ran reliabilities for numerous personality variables and found them generally wanting. Trying to salvage the data we tested all distinctions that coders made and found that the set of values of interest where too unreliable to consider. The only reliable distinction occurred between "cannot code" and the set of values of interest lumped into one. This gave us two 'reliable' categories "cannot code" and "can code," which says something about the appropriateness of the coding instrument but nothing about the nature of the television characters to be studied. Whenever default categories are added to a coding scheme, they not only attest to the logical incompleteness of the coding scheme relative to the data to be coded, but also invite an easy way out of making difficult

distinctions – an important source of unreliability. In either case, they do not distinguish among units coded, and provide no information about them.

Because it is the analytically relevant values that have the potential of contributing to research findings, default categories should not be included in reliability tests concerning the data subsequently analyzed. For example, instead of computing the reliability from reliability data in Figure 1,

45 pairable entries in the reliability data

Coder A:	*	*	*	*	*	3	4	1	2	1	1	3	3	*	4
Coder B:	1	*	2	1	3	3	4	4	*	*	*	*	*	*	*
Coder C:	*	*	2	1	3	4	4	*	2	1	1	3	3	*	1

FIGURE 1: Reliability Data Including Default Values *

in which "*" stands for the non-informative default category "other than 1 through 4," it is advisable to eliminate all *s and obtain reliabilities from data in Figure 2, containing only the values that matter.

26 pairable values that matter

Coder A:						3	4	1	2	1	1	3	3		4
Coder B:	1		2	1	3	3	4	4							
Coder C:			2	1	3	4	4		2	1	1	3	3		1

FIGURE 2: Reliability Data Excluding Default Values

Another reason for excluding default categories is that they often are of a logically different kind. For once, they are ambiguous regarding the values that matter. The agreement on * in the 2nd and 14th units of Figure 1 reveals nothing about that unit, just as not diagnosing 94% of patients as paranoid suggests nothing about why they are admitted to the mental institution. The difference between a default category and a legitimate value is difficult to imagine and articulate. If values 1 through 4 in Figure1 were points on an interval scale, * had no place on that scale. Pairing default categories with other default categories and with values intended to provide the researcher with information of interest to them makes no sense in a reliability test and should be avoided where possible.

Removing the 18 *s from the data in Figure 1 to obtain the reliability data in Figure 2 in effect removes three units, two for total lack of information, and one for containing only one value, which cannot be paired with another value in that unit. Because the statistical independence of units and the pairable values used to describe them does not depend on the identity of coders but on the values they provide, the expected agreement or disagreement is obtained from the remaining total of 26 pairable values, which manifest the very distinctions and only these that researchers had built into the values of a variable, presumably because these values are relevant to their intended analyses.

For those curios, assuming * to be a legitimate value, and treating the reliability data in Figure 1 as nominal data, yields α=0.239. Omitting that default category as uninformative, the 26 pairable values in Figure 2, tabulated as example 1 in Figure 4, yield α=0.698. (In

coincidence matrices, *alpha*'s observed disagreements, here D_o=6, are found in the off-diagonal cells and its expected disagreements, here D_e=19.84, are computed from the margins of this matrix for the same off-diagonal cells). It should be noted, however, that this the higher reliability computed for data in Figure 2 has nothing to do with * providing no information about the units coded, is not the result of reducing the 45 pairable values to 26 and the 15 units to 12 due to omitting the default category, nor is it affected by some units being coded by two coders and others by more. *Alpha* is not influenced by such variations. Although the omission of default categories from reliability data increases reliability in this particular case, this need not be so.

There are two recommendations regarding the use of non-informative categories.

(1) Where possible: *replace default categories by well defined values that, while perhaps less important to a particular research project, allow coders to unambiguously distinguish among mutually exclusive qualities of the units they encounter*.

(2) Realizing that it is not always possible to anticipate all qualities of units that coders may face, *offer coders rules that minimize their use of default categories for mere convenience*. Without such rules, the use of default categories tends to become a function of coders' unequal attention to details that do matter and their unequal willingness to spent time and effort in figuring out which value appropriately describes a given unit.

Lack of Variability

Suppose reliability data turn out as in Figure 3 with a=b=1. In this extreme case, percent agreement is perfect or 100%, albeit uninformative. However, randomly drawing pairs of values from an urn containing only one kind of values would yield an expected agreement identical to the observed, hence, α=0. This finding is the extreme version of the puzzle mentioned in the beginning. But why should reliability then be absent?

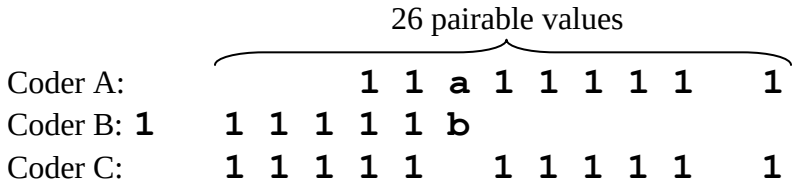

FIGURE 3: Reliability Data with Low Variability

As suggested above, measurement theory equates reliability with the extent to which variation in the measures can be explained by variation in the nature of the units or phenomena measured. In the absence of such variation, researchers would not know whether their measuring instrument *can* respond to differences among units should they occur – after all, even a defective thermometer may still show a numerical value each time one looks at it. One would not know whether it indicates anything unless one can make it change that value in response to a changed situation. Even when the variability of an instrument can be assured by other means, whenever a variable is fixed at a single value, the data in that variable could not be correlated with anything, can explain nothing, and might as well be discarded as useless. This is the condition at which reliability data cannot provide any information about the reliability of a population of data – as seen in example 2 of Figure 4.

Data in Figure 2

Values:	1	2	3	4	
1:	6			2	8
2:		4			4
3:			7	1	8
4:	2		1	3	6
	8	4	8	6	26

Figure 3, a=b=1

	1	2	3	4	
	26				26
					0
					0
					0
	26	0	0	0	26

Figure 3, a=1, b=4

	1	2	3	4	
	24			1	25
					0
					0
	1				1
	25	0	0	1	26

Figure 3, a=b=4

	1	2	3	4	
	24				24
					0
					0
				2	2
	24	0	0	2	26

Reliabilities:

$$\alpha_{nominal} = 1 - \frac{D_o}{D_e} = 1 - \frac{6}{19.84} = 0.698 \qquad \alpha = 1 - \frac{0}{0} = 0.000 \qquad \alpha = 1 - \frac{2}{2} = 0.000 \qquad \alpha = 1 - \frac{0}{3.84} = 1.000$$

Information:

$i_{data} = {}^{26}/_{202} = 0.129$, $i_{coding} = {}^{26}/_{138.5} = 0.188$

$i_{data} = {}^{0}/_{\infty} = 0.000 = 0.188$

$i_{data} = {}^{26}/_{720} = 0.036 = 0.188$

$i_{data} = {}^{26}/_{373} = 0.070 = 0.188$

FIGURE 4: Coincidence Matrices, *Alpha*-Reliabilities, and Information Measures for Data in Figures 2 and 3 and for the Coding Instrument that Generated these Data

Adherents of the interpretation theory of reliability may start differently but come to the same conclusion. They equate reliability with the ability of researchers and their stakeholders to comprehend and where needed reproduce the distinctions coders made among the units of analytical interest. By assigning the same value to all units, there is no evidence that distinctions were made and reproducing them is impossible to verify. Coders may have become so habituated to using the most common value so that unexpected kinds are overlooked or they may have eased their coding task by agreeing among themselves to use the most plausibly value throughout – a sure way to get 100% agreement. It may also be that the researcher provided the coders with units that showed no diversity at all. Unless there is evidence for coders to have exercised their ability to distinguish among units, the data they generate are meaningless.

Suppose variation is minimal, as when a=1 and b=4 in Figure 3, also tabulated as example 3 in Figure 4. Here too, expected and observed agreements are necessarily equal and α=0. Again, this makes good sense. Regarding the only unit that seems different, at least for one coder, the other coders disagree as to whether and how it differs from the majority. But now, suppose the two coders agree on the uncommon value for that unit, a=b=4. There is perfect agreement throughout, as can be seen in example 4 of Figure 4, and α=1. This radical jump of *alpha* from zero to one on account of just one coder changing the valuation of just one unit should raise questions. Evidently, chance-corrected agreement coefficients are more sensitive to rare cases than to frequent ones. Therefore, unusual qualities require special attention by coders. The lesson drawn from these examples is that chance-corrected agreement coefficients are perfectly fine indicators of the reliability of coding processes when units display ample diversity but seem oblivious to the very amount of information that a computed reliability coefficient provides about the reliability of coding.

Moreover, lack of variability always means that information about some values is either missing – in the previous version of Figure 3 about the values 2 and 3 – or inadequate – as for 4. When some values remain unused by coders, a practical suggestion for coping with the lack of

variation (see also Tinsley and Weiss, 2000, p. 109) is: *select units for reliability tests that show enough diversity to cover all or most values available for coding*, for example by stratified, not random sampling, and not necessarily from the population of data being investigated. This recommendation does not aim at improving reliability but the information about it. This recommendation will be qualified below.

Inadequate Sample Sizes

The third source of inadequate information about the reliability of coding does not need much discussion. Obviously, the reliability data in Figure 2, recording the valuation of only 12 units with a total of 26 pairable values assigned by three coders, is insufficient for interpreting a reliability coefficient as a measure of a very much larger population of data not yet coded. This inadequacy is a traditional sampling issue. The power of reliability tests depend on the frequency with which informative values are used (Krippendorff, 2004, p. 240) and the only remedy of this inadequacy is to *increase the sample size of the reliability data,* units, coders or both.

INFORMATION ABOUT RELIABILITY

I contend that the above mentioned puzzles can be solved by considering a measure of the extent a computed reliability can be trusted to infer the reliability in question. The ability of a high reliability coefficient to lead to the conclusion that the data of interest are sufficiently reliable when in fact they are, is often called the power of a statistical test – the probability that a test will not falsely reject a null hypothesis that is false, or the probability of not making a Type II error. Because null-hypotheses are of little interest in reliability assessments where very large deviations from chance agreement need to be achieved, statistical power can serve here only as a mere analogy.

For *a priory* determination of the sample size of sufficiently powerful reliability data, following Bloch and Kraemer (1989), Krippendorff (2004, p. 239) provides an approximation to the required reliability sample size of *units* in a binary distinction between one value *c* and its complement, *not-c*. It requires, though, an estimate of the probability of rare values. This paper proposes two *post hoc* measures of the likelihood that a computed reliability coefficient speaks to the reliability of interest.

It should be noted that the likelihood of reliability data to speak about the reliability of interest is affected not only by the number of units coded but also by the number of coders employed in distinguishing among them and assigning values to them, one value per unit. The more coders are employed and the more pairable values they collectively generate, the more one can trust the computed reliability. This likelihood, the extent to which a reliability data are informative about what the test claims to measure will be called the *amount of information i* and expressed as the proportion:

$$i = min\left(1,\ \frac{Number\ of\ \ pairable\ values\ observed}{Number\ of\ \ pairable\ values\ required}\right)$$

The proportion *i* is positive. Its maximum is capped at 1.00 at which the reliability data are informationally adequate or exceed what is required.

Because the computation of the required sample sizes of values (often: units × coders) can be complicated, Table 1 provides the needed numbers of *values*, $T(P_c, \alpha_{min}, p)$, as a function of the probability P_c of any one value of a variable, the smallest tolerable *alpha* α_{min} relative to the research problem at hand, and the desired significance level p. When P_c falls between two probabilities, listed in Table 1, researchers may need to interpolate the required number. Interpolation invites inaccuracies, of course. However, one should realize that the mathematical form adopted here, provides conservative estimates of the required sample sizes. Other approximations are conceivable but not pursued here.

Smallest acceptable α_{min} =	**.667**				**.800**				**.900**			
Level of significance p =	**.100**	**.050**	**.010**	**.005**	**.100**	**.050**	**.010**	**.005**	**.100**	**.050**	**.010**	**.005**
P_c = .5 or V = 2 values	36	60	119	146	62	103	206	252	128	211	422	518
= .25 or = 4	49	81	161	198	84	139	277	340	172	283	566	694
= .1 or = 10	104	172	344	421	178	293	587	719	361	595	1190	1459
= .05 or = 20	200	329	657	806	340	560	1119	1372	657	1131	2263	2775
= .025 or = 40	412	679	1358	1665	664	1095	2189	2684	1344	2214	4430	5431
= .01 or =100	966	1591	3182	3901	1640	2701	5403	6624	3307	5447	10896	13359

Note: $T(P_c, \alpha_{min}, p) = 2z_p^2\left(\frac{(1+\alpha_{min})(3-\alpha_{min})}{4(1-\alpha_{min})P_c(1-P_c)} - \alpha_{min}\right)$ = the minimum number of values required for the *c*|*not c* distinction

Where: P_c = the probability of value *c*
α_{min} = the smallest alpha for coding to be accepted as reliable
z_p = the standardized *z*-statistics at the level of significance p

TABLE 1: The Minimum Number $T(P_c, \alpha_{min}, p)$ of Values Required for the Distinction Between any one Value *c* and Values *non-c* to be Informationally Adequate

Accepting the general form of *i*, one needs to distinguish between two kinds of information which pertain to different inferences to which reliability data may be employed and different sampling methods that are appropriate for improving them: The amount of information about the data used for drawing conclusions about a research question and the amount of information about the coding instrument employed in that process.

Information about the Data in Question i_{data}

To trust the reliability of the *data used for drawing conclusions about the research questions posed*, reliability data should be representative of the population of units to be coded and, hence, be obtained as a probability sample of units from that population. Probability samples assure that *each kind of unit in the population to be studied has an equal chance to be included in the sample of reliability data*, with the additional provision that *each occurs with adequate frequency*. The latter qualification is achievable by sampling not to a minimum number of units, but until the least frequent kind is informationally adequate. The *proportion of information about the data* is:

$$i_{data} = min\left(1,\ \frac{n}{T_{data}}\right)$$

Where: n = the observed number of pairable values in the reliability data
$T_{data} = T(P_{min}, \alpha_{min}, p)$ = the minimally required number of values, computed or obtained from Table 1
P_{min} = the smallest probability of any one analytically important value c of a variable, used at least once

For example, suppose the value used least often occurs 12 times out of a total of 446, then $P_{min} = {}^{12}/_{446} = 0.027$ or, to be conservative, about 0.025, which is found listed in Table 1. Suppose one aims for $\alpha_{min} = 0.800$ at $p = 0.050$, the table shows T_{data}=1095 as the required number of values. Dividing the number of values used in coding the data by those required yields $i_{data} = {}^{446}/_{1095} = 0.407$.

Informational adequacy could be achieved by generating additional values:

$$N_{values\,to\,be\,added} = (1 - i_{data}) \times T_{data}$$

In the above example, $(1 - 0.407) \times 1095 = 649$ values are found lacking. Suppose the 446 values resulted from two coders recording 223 observations. The required increase in values can be achieved either by adding 325 units to the reliability data and continuing to employ two coders, by coding the same 223 units but by five instead of two coders, or by any combination of increases in number of units or coders.

Information about a Variable in the Coding Instrument i_{coding}

By contrast, to trust *the reproducibility of a coding or measuring instrument* applied elsewhere, with different coders, under different circumstances and for various distributions of data, units sampled need to have sufficient diversity for reliability data to be adequately informative about the reliability of all values of a variable. In this situation, the units sampled need not represent a particular population of current interest to the researcher; rather, they need to involve all coding decisions that define the variable in question. To obtain adequate data for coding a variable with a particular instrument, sampling must be such that *each distinction among units, defined by the values available for coding, occurs with adequate frequency*. That is, the units in an informationally adequate sample must have sufficient diversity to cover all values of a variable – ideally being uniformly distributed – are sufficiently frequent, and generated by an ideally large number of coders whose use of the coding instructions is representative of how the community of researchers understand them.

The *proportion of information $i_{coding\,(c)}$ in any one value c* has a form similar to i_{data}:

$$i_{coding(c)} = min\left(1,\ \frac{Vn_c}{T_{coding}}\right)$$

Where: n_c = the number of pairable values c in the reliability data
V = the number of values c available for coding the variable in question
$T_{coding} = T({}^{1}/_{V}, \alpha_{min}, p)$ = the minimally required number of values, computed or obtained from Table 1, wherein ${}^{1}/_{V}$ takes the place of P_c for a uniform distribution of values,

The *proportion of information i_{coding} about all values c in a variable* then becomes the simple average of $i_{coding\ (c)}$:

$$i_{coding} = \frac{1}{V}\sum_{c=1}^{V} i_{coding\ (c)} = \sum_{c=1}^{V} min\left(\frac{1}{V}, \frac{n_c}{T_{coding}}\right)$$

The number of pairable values that a researcher would have to add to the existing data is, just as for i_{data}, but computed with T_{coding}:

$$N_{values\,to\,be\,added} = \left(1 - i_{coding}\right) \times T_{coding}$$

Informational adequacy for the reliability of coding processes is achievable either by coding an additional number of units by the same number of coders, but unlike for i_{data}, selecting values of *the missing kinds*, by adding coders to provide *the missing values*, or measures of both, in either case to achieve an ideally uniform distribution of values.

Corrective Actions Available

The importance of distinguishing i_{data} and i_{coding} may be illustrated by an example. Suppose a content analysis requires identifying countries covered in TV news, coders reliably identified the 20 countries mentioned, and i_{data} is sufficiently large, then one may well proceed to an analysis of these data, refraining from drawing any conclusions about the countries not mentioned, except for the possibility of noticing their absence. But given that there are about 195 countries in the world plus emerging countries, unrecognized countries, contested countries, and countries that form unions or have changed their names, generalizing the reliability from coding only 20 out of 195 or about 10% of all countries to coding the remaining ones would not be justifiable. Thus, reliability data obtained form a particular population may be informative about the reliability of coding that population, but not about the reproducibility of a variable in which some or many values remain unused – regardless of the number of coders involved.

An important difference between the two information measures concerns the actions they suggest to improve or overcome the informational deficiencies they respectively assess. As already suggested, the two measures are linked to radically different sampling methods. For example:

When $i_{data}<1$, lack of information about the reliability of the data, affords two actions:

(1) *Probability sampling* continues until the least frequent value is informationally adequate. Obligated to assure that the distribution of reliability data remains representative of the population of analytical interest, researchers can hardly act otherwise.

(2) However, should the value with the lowest frequency be analytically dispensable, one may decide to refrain from drawing any inferences from that value and select the value with the next lowest frequency of analytically indispensible significance and proceed as in (1).

When $i_{coding}<1$, lack of information about the reliability of the coding process, requires *stratified sampling* until all available values occur with sufficient frequency, the uniform distribution being the most information-efficient distribution in this case. Thus, in testing the reliability of coding processes, researchers must sample enough from each kind of units.

Rectifying informational inadequacies for the reliability of a coding instrument reveals an epistemological dilemma. How can one stratify not-yet coded *units* in advance of coding when

being of one kind or another is determined only after the coders assigned *values* to them? This dilemma can be dissolved in several (increasingly undesirable) ways. In practice, this may be less of a problem.

(1) Researchers usually know the kind of units they are analyzing and while only reliable coding can determine their kind, researchers may well define strata within which samples are drawn, awaiting post factum verification for their intuitions.
(2) Researchers may employ quota sampling, admitting units selectively until all values have numerically adequate frequencies. Both actions are possible only when all relevant qualities of units occur in the population of units.
(3) When some qualities associated with particular values do not occur in that population, then further sampling is hopeless. Under these conditions, it is permissible to draw units from other sources.
(4) And failing their availability as well, one may need to make up examples of each kind to fill that void.
(5) Finally, if no examples can be found, one might question the logic by which the variable is constructed. After all, the word "variable" means "able to vary," distinguishing among alternatives. Should alternatives not exist, a variable that claims them is flawed conceptually.

Discussion of the Properties of the Two Information Measures

Both information measures are defined as proportions of the number of values observed to the number of values required for the computed reliability to be adequately informative about what it claims to measure. One could conceptualize this proportion as a probability, the probability of avoiding Type II errors. However, as above mentioned, the null-hypothesis of chance agreement is not relevant in reliability assessments where the issue is to deviate minimally from perfect agreement. Hence, this proportion should be interpreted as the probability of being able to conclude that the computed reliability equals the reliability of a population of coded data or of a coding instrument, respectively. When reliability data are informationally adequate, this probability is one. Following are some observations of note:

i_{data} and i_{coding} are not correlated with the agreement coefficient *alpha,* except when $i_{data}=0$ then also $\alpha=0$, but not vice versa.

When $i_{data}=i_{coding}=1$, all available values do occur in adequate numbers and the computed *alpha* is sufficiently informative about the reliability of the data and of the coding instrument that generated then.

When $i_{data}=1$, i_{coding} equals the proportion of used to unused values. To improve i_{coding}, this would suggest sampling the unused values only.

When coders fail to distinguish among the units in the sample, assign all units to the same value, as when a=b=1 in Figure 3, tabulated as example 2 of Figure 4, $i_{data}={}^{26}/_{\infty}=0$ and information about the data is absent. If all units in the population are of the same kind, there is no point in continuing to sample from that population. Researchers have to develop new coding instructions that do distinguish among the units of interest.

When all available values of a variable occur in the data at least once, differences between i_{data} and i_{coding} may have several explanations. In Figure 2 and example 1 of Figure 4, there are

n_1=8, n_2=4, n_3=8, and n_4=6 pairable values, summing to n =26. Aiming for α_{min}=0.800 at a p=0.050 level of significance, P_{min}=$^4/_{26}$=0.154 (located between 0.1 and 0.25 in Table 1), T_{data}= T(0.154, 0.8, 0.05) is a number between 139 and 293 and in fact computed as 184. Compared with 26 values found in the data, the amount of information i_{data}= $^{26}/_{184}$ = 0.141. To achieve adequate information about the data, this calculation suggests that (1–i_{data})184=158 pairable values are missing, calling for an increase in the sample size by a factor of $^1/i_{data}$=$^1/_{0.141}$=7.08, and aiming at a distribution of values n'_1=57, n'_2=28, n'_3=57, and n'_4=42.

By contrast, assuming an interest in the reliability of the coding instrument, the above reliability data, having V=4 values available for coding, require T_{coding}=T($^1/_4$, 0.8, 0.05)=139 additional values. The information about the coding process is i_{coding}=$^8/_{139}$+$^4/_{139}$+$^8/_{139}$+$^6/_{139}$= 0.187. Since all four kinds of values evidently do occur in the population of data, it would be possible to sample from that population by stratified or quota sampling. Here, the number of pairable values that are missing is 113, which has to be selected unequally from the four values to yield $^{139}/_4$=34.75 or about 35 of each kind.

As evident in the two accounts of data in Figure 2, if all values occur in the reliability data at least once, because unequal distributions of values are less informationally efficient than uniform distributions, i_{data} typically provides less information and calls for larger sample sizes than i_{coding}.

When coders make very few distinctions among units, as in Figure 3 with a=1 and b=4, tabulated in example 3 of Figure 4, and aiming for α_{min}=0.800 at p=0.050, and P_{min}= $^1/_{26}$ being small, T_{data}=T($^1/_{26}$, 0.8, 0.05) is large, in fact 1023, and i_{data}=$^{26}/_{1023}$=0.025. Here, the sample size of pairable values would have to be increased by a factor of $^1/i_{data}$=39.35, i.e., from 26 to a staggering total of 1023 values for probability sampling to yield informationally adequate reliability data.

Continuing with Figure 3, had coder A and B agreed on the 10th unit, a=b=4, this unit being the only unit different from the majority of agreements on value 1, the calculated reliability would have jumped from α=0 to α=1, an apparent oddity already discussed. However, the amount of information about the data would have increased only to i_{data}=$^{26}/_{373}$=0.070, still far too small for the computed reliability to give us any confidence that the data of interest can be relied upon. Incidentally, since the 26 pairable values are well below the informationally required uniform distribution, i_{coding}=0.188 remains unaffected by how these values are distributed, in the four coincidence matrices of Figure 4.

By giving a reason for why 100% agreement, even a reliability coefficient of α=1 may not be sufficient to declare data or a coding instrument reliable, the puzzle described in the beginning of this paper is solved.

SUMMARY AND CONCLUSION

This paper attempts to explain the frequently encountered and seemingly counterintuitive situation of observing high percent agreement while calculating reliabilities that are discouragingly low, even near zero. More important is the opposite phenomenon of having reasonable doubt in the ability to infer the reliability of data or of a coding process from high, even perfect agreement coefficients. These puzzles are not solvable by merely correcting observed agreements by chance, i.e., by what can be expected when values are randomly paired

and assigned to the sampled units, but by introducing a measure of how much information given reliability data provide about the reliability of a population of data or of the variable that generates them. Two simple information measures are proposed. Both are proportions of the observed to the required number of values, akin to the probability of avoiding Type II errors in statistical power calculations. These information measures are meant as companions to *alpha* and other chance-corrected agreement coefficients, and to aid practical decisions regarding whether one can trust the data and/or the coding instrument that generated them.

It should be noted that the sample sizes and distributions of values in the reliability data to which the information measures respond, also affect the confidence intervals of the calculated reliability coefficients (for *alpha*, see Krippendorff, 2004, pp. 237-238 and Hayes & Krippendorff, 2007), providing access to Type I errors. If sample sizes are small, confidence intervals are large. While confidence intervals are important qualifiers of computed reliability measures, they do not inform researchers about when computed reliabilities can be trusted and how reliability data might be made more informative about the reliability they are meant to assess.

AKNOWLEDGEMENTS

I am grateful to Ron Artstein for valuable suggestions on an earlier draft of this paper, and to Andrew Hayes for encouraging me to simplify access to the information measures.

REFERENCES

Bennett, E. M., Alpert, R. & Goldstein, A. C. (1954). Communications through limited response questioning. *Public Opinion Quarterly, 18*, 303-308.

Bloch, D. A. & Kraemer, H. C. (1989). 2×2 kappa coefficients: Measures of agreement or association. *Biometrics, 45*, 269-287.

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.

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

Goodman, L. A. & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49*, 732-764.

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

Holley, W. & Guilford, J. P. (1964). A note on the G-index of agreement. *Educational and Psychological Measurement, 24*, 749–754.

Janson, S. & Vegelius, J. (1979). On generalizations of the G index and the phi coefficient to nominal scales. *Multivariate Behavioral Research, 14*, 255–269.

Krippendorff, K. (2008). Systematic and random disagreement and the reliability of nominal data. *Communication Methods and Measures, 2*, 323-338.

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

Krippendorff, K. (1993). Computing Krippendorff's Alpha Reliability. http://repository.upenn.edu/asc_papers/43/ accessed 2009.8.5.

Krippendorff, K. (1978). Reliability of binary attribute data. *Biometrics, 34*, 142-144.

Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics, 45*, 255-268.

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.

Pearson, K, et al. (1901). Mathematical Contributions To The Theory of Evolution. IX: On The Principle of Homotyposis and Its Relation To Heredity, To Variability of The Individual, and To That of Race. Part I: Homotyposis in The Vegetable Kingdom. *Philosophical Transactions of The Royal Society* (London), *Series A, 197*, 285-379.

Perreault, W. D. & Leigh, L. E. (1989). Reliability of nominal data based on qualitative judgments. *Journal of Marketing Research, 26*, 135–148.

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

Tinsley, H. E. A. & Weiss, D. J. (2000). Interrater reliability and agreement. In Tinsley, H. E. A. & Brown, S. D. (Eds.). *Handbook of Applied Multivariate Statistics and Mathematical Modeling* (pp. 94-124). New York: Academic Press.

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