Manuscript published in
Human Communication Research 30, 3: 411-433, 2004.

Reliability in Content Analysis: Some Common Misconceptions and Recommendations

Klaus Krippendorff
University of Pennsylvania
kkrippendorff@asc.upenn.edu

Abstract

In a recent article published in this journal, Lombard, Snyder-Duch, and Bracken (2002) surveyed 200 content analyses for their reporting of reliability tests; compared the virtues and drawbacks of five popular reliability measures; and proposed guidelines and standards for their use. Their discussion revealed that numerous misconceptions circulate in the content analysis literature regarding how these measures behave and can aid or deceive content analysts in their effort to ensure the reliability of their data. This paper proposes three conditions for statistical measures to serve as indices of the reliability of data and examines the mathematical structure and the behavior of the five coefficients discussed by the authors, plus two others. It compares common beliefs about these coefficients with what they actually do and concludes with alternative recommendations for testing reliability in content analysis and similar data-making efforts.

In a recent paper published in a special issue of *Human Communication Research* devoted to methodological topics (Vol. 28, No. 4), Lombard, Snyder-Duch, and Bracken (2002) presented their findings of how reliability was treated in 200 content analyses indexed in *Communication Abstracts* between 1994 and 1998. In essence, their results showed that only 69% of the articles report reliabilities. This amounts to no significant improvements in reliability concerns over earlier studies (e.g., Pasadeos et al., 1995; Riffe & Freitag, 1996). Lombard et al. attribute the failure of consistent reporting of reliability of content analysis data to a lack of available guidelines, and they end up proposing such guidelines.

Having come to their conclusions by content analytic means, Lombard et al. also report their own reliabilities, using not one, but four, indices for comparison: %-agreement; Scott's (1955) π (pi); Cohen's (1960) κ (kappa); and Krippendorff's (1970, 2004) α (alpha). Faulty software[1] initially led the authors to miscalculations, now corrected (Lombard et al., 2003). However, in their original article, the authors cite several common beliefs about these coefficients and make recommendations that I contend can seriously mislead content analysis researchers, thus prompting my corrective response. To put the discussion of the purpose of these indices into a larger perspective, I will have to go beyond the arguments presented in their article. Readers who might find the technical details tedious are invited to go to the conclusion, which is in the form of four recommendations.

The Conservative/Liberal Continuum

Lombard et al. report "general agreement (in the literature) that indices which do not account for chance agreement (%-agreement and Holsti's [1969] CR – actually Osgood's [1959, p.44] index) are too liberal while those that do (π, κ, and α) are too conservative" (2002, p. 593). For liberal or "more lenient" coefficients, the authors recommend adopting higher critical values for accepting data as reliable than for conservative or "more stringent" ones (p. 600) – as if differences between these coefficients were merely a problem of locating them on a shared scale. Discussing reliability coefficients in terms of a conservative/liberal continuum is not widespread in the technical literature. It

entered the writing on content analysis not so long ago. Neuendorf (2002) used this terminology, but only in passing. Before that, Potter and Lewine-Donnerstein (1999, p. 287) cited Perreault and Leigh's (1989, p. 138) assessment of the chance-corrected κ as being "overly conservative" and "difficult to compare (with) … Cronbach's (1951) alpha," for example – as if the comparison with a correlation coefficient mattered.

I contend that trying to understand diverse agreement coefficients by their numerical results alone, conceptually placing them on a conservative/liberal continuum, is seriously misleading. Statistical coefficients are mathematical functions. They apply to a collection of data (records, values, or numbers) and result in one numerical index intended to inform its users about something – here about whether they can rely on their data. Differences among coefficients are due to responding to (a) different patterns in data and/or (b) the same patterns but in different ways. How these functions respond to which patterns of agreement and how their numerical results relate to the risk of drawing false conclusions from unreliable data – not just the numbers they produce – must be understood before selecting one coefficient over another.

Issues of Scale

Let me start with the ranges of the two broad classes of agreement coefficients, chance-corrected agreement and raw or %-agreement. While both kinds equal 1.000 or 100% when agreement is perfect, and data are considered reliable, %-agreement is zero when absolutely no agreement is observed; when one coder's categories unfailingly differ from the categories used by the other; or disagreement is systematic and extreme. Extreme disagreement is statistically almost as unexpected as perfect agreement. It should not occur, however, when coders apply the same coding instruction to the same set of units of analysis and work independently of each other, as is required when generating data for testing reliability.

Where the reliability of data is an issue, the worst situation is not when one coder looks over the shoulder of another coder and selects a non-matching category, but when coders do not understand what they are asked to interpret, categorize by throwing dice, or examine unlike units of analysis, causing research results that are indistinguishable from chance events. While zero %-agreement has no meaningful reliability interpretation, chance-corrected agreement coefficients, by contrast, become zero when coders' behavior bears no relation to the phenomena to be coded, leaving researchers clueless as to what their data mean. Thus, the scales of chance-corrected agreement coefficients are anchored at two points of meaningful reliability interpretations, zero and one, whereas %-like agreement indices are anchored in only one, 100%, which renders all deviations from 100% uninterpretable, as far as data reliability is concerned. %-agreement has other undesirable properties; for example, it is limited to nominal data; can compare only two coders[2]; and high %-agreement becomes progressively unlikely as more categories are available. I am suggesting that the convenience of calculating %-agreement, which is often cited as its advantage, cannot compensate for its meaninglessness. Let me hasten to add that chance-correction is not a panacea either. Chance-corrected agreement coefficients do not form a uniform class. Benini (1901), Bennett, Alpert, and Goldstein (1954), Cohen (1960), Goodman and Kruskal (1954), Krippendorff (1970, 2004), and Scott (1955) build different corrections into their coefficients, thus measuring reliability on slightly different scales. Chance can mean different things. Discussing these coefficients in terms of being conservative (yielding lower values than expected) or liberal (yielding higher values than expected) glosses over their crucial mathematical differences and privileges an intuitive sense of the kind of magnitudes that are somehow considered acceptable.

If it were the issue of striking a balance between conservative and liberal coefficients, it would be easy to follow statistical practices and modify larger coefficients by squaring them and smaller coefficients by applying the square root to them. However, neither transformation would alter what these mathematical functions actually measure; only the sizes of the intervals between 0 and ±1.

Lombard et al., by contrast, attempt to resolve their dilemma by recommending that content analysts use several reliability measures. In their own report, they use α, "an index …known to be conservative," but when α measures below .700, they revert to %-agreement, "a liberal index," and accept data as reliable as long as the latter is above .900 (2002, p. 596). They give no empirical justification for their choice. I shall illustrate below the kind of data that would pass their criterion.

Relation Between Agreement and Reliability

To be clear, agreement is what we measure; reliability is what we wish to infer from it. In content analysis, reproducibility is arguably the most important interpretation of reliability (Krippendorff, 2004, p.215). I am suggesting that an agreement coefficient can become an index of reliability only when

(1) It is applied to proper reliability data. Such data result from duplicating the process of describing, categorizing, or measuring a sample of data obtained from the population of data whose reliability is in question. Typically, but not exclusively, duplications are achieved by employing two or more widely available coders or observers who, working independent of each other, apply the same coding instructions or recording devices to the same set of units of analysis.

(2) It treats units of analysis as separately describable or categorizable, without, however, presuming any knowledge about the correctness of their descriptions or categories. What matters, therefore, is not truths, correlations, subjectivity, or the predictability of one particular coder's use of categories from that by another coder, but agreements or disagreements among multiple descriptions generated by a coding procedure, regardless of who enacts that procedure. Reproducibility is about data making, not about coders. A coefficient for assessing the reliability of data must treat coders as interchangeable and count observable coder idiosyncrasies as disagreement.

(3) Its values correlate with the conditions under which one is willing to rely on imperfect data. The correlation between a measure of agreement and the rely-ability on data involves two kinds of inferences. Estimating the (dis)agreement in a population of data from the (dis)agreements observed and measured in a subsample of these data is an inductive step and a function of the number of coders involved and the proportion of units in the recoded data. Inferring the (un)reliability of data from the estimated (dis)agreements is an abductive step and justifiable mainly in terms of the (economical, social, or scientific) consequences of using imperfect data. An index of the degree of reliability must have at least two designated values, one to know when reliability is perfect, and the other to know when the conclusions drawn from imperfect data are valid by mere chance.

Note that (1) defines a precondition for measuring reliability. No single coefficient can determine whether coders are widely available, use the same instructions, work independently, and code identical units of analysis. Researchers must ensure their peers or critics that the reliability data they generate satisfy these conditions. Many methodological problems in testing reliability stem from violating the requirement for coders to be truly independent, being given coding instructions they cannot follow, or applying them to data that they fail understand.

The two methodological problems considered here result from choosing inadequate measures of agreement – calling something a reliability coefficient does not make it so – and applying indefensible decision criteria on their results. Since Lombard et al. discuss the relative merits of the above mentioned measures, correctly citing widely published but disputable claims, I feel compelled to provide mathematical demonstrations of how these coefficients actually differ and whether they satisfy (2) and (3) above. Let me discuss several better-known candidates.

A Comparison of Seven Agreement Coefficients

To begin, Lombard et al. are correct in discouraging the use of association, correlation, and consistency coefficients, including Cronbach's (1951) alpha, as indices of reliability in content

analysis. Association measures respond to any deviation from chance contingencies between variables, correlations moreover from linearity, whereas (2) stipulates that reliability must be indicated by measures of agreement among multiple descriptions. Although the authors do not report on how often content analysis researchers fail to realize this crucial difference and use inappropriate indices (I could cite numerous examples of such uses and even name explicit proponents of such practices), one cannot strongly enough warn against the use of correlation statistics in reliability tests. I agree with the authors' assessment of the inappropriateness of such coefficients, and therefore need not consider them here. However, I take issue with their presentation of the differences among chance-corrected agreement coefficients. A crucial point is whether and how the population of data whose reliability is in question enters the mathematical form of a coefficient, whether not only (2) but also (3) is satisfied. To illustrate the issues involved, I shall compare the five coefficients that the Lombard et al. found to be most commonly used, plus Benini's (1901) β (beta), and Bennett, Alpert, and Goldstein's (1954) S in their most elementary forms: for dichotomous data generated by two coders. In such a severely restricted but mathematically exceptionally transparent situation, reliability data can be represented by means of the familiar proportions a, b, c, and d of a 2-by-2 contingency table, shown in Figure 1. In this figure, a+d is the observed %-agreement, A_o; b+c is the observed %-disagreement; and its marginal sums show the proportions p of 0s and q=1−p of 1s as used by the two coders A and B, respectively.

Figure 1
Generic 2-by-2 Contingency Table

		Coder A			
Values:		**0**	**1**		**Population Estimates**
Coder B	**0**	a	b	p_B	$\bar{p} = (p_A + p_B)/2$ = proportion of 0s in data
	1	c	d	q_B	$\bar{q} = (q_A + q_B)/2 = 1 - \bar{p}$ = proportion of 1s in data
		p_A	q_A	1	

Figure 2 states the above-mentioned agreement coefficients in terms of Figure 1 and in α's economical form:

$$\text{Agreement} = 1 - \frac{D_o}{D_e} = 1 - \frac{\text{Observed Disagreement}}{\text{Expected Disagreement}}$$

where, when the observed disagreement D_o=0, agreement =1; and when the two disagreements are equal, D_o=D_e, agreement =0. So, D_o expresses the lack of agreement; whereas D_e defines the zero point of the measure.

Figure 2
The Dichotomous Forms of Seven Agreement Coefficients

	Agreement	=	**1**	–	**Observed**	/	**Expected Disagreements**	
%-agreement	$\mathbf{A_o}$	=	1	–	(b + c)			
Osgood (1959); Holsti's	**CR**	=	$[1-(b+c)]$				$\frac{2 \cdot N_{A \cap B}}{N_A + N_B}$	
Bennett et al. (1954)	**S**	=	1	–	(b + c)	/	$2 \cdot ½ \cdot ½$	where ½ is the logical probability of 0 and of 1
Scott (1955)	**π**	=	1	–	(b + c)	/	$2\bar{p}\bar{q}$	where $\bar{p} = \frac{p_A + p_B}{2}$ and $\bar{q} = 1 - \bar{p}$
Krippendorff (1970)	**α**	=	1	–	(b + c)	/	$\frac{n}{n-1} 2\bar{p}\bar{q}$	where n = the number of 0s and 1s used jointly
Cohen (1960)	**κ**	=	1	–	(b + c)	/	$p_A q_B + p_B q_A$	

Benini (1901) $\beta = 1 - [(b+c)-|b-c|] / [p_A q_B + p_B q_A - |b-c|]$

Evidently, all coefficients in Figure 2 contain the same observed disagreement, the proportion of mismatches (b+c), which satisfies part of (2). The %-agreement measure, A_o, stops there, making no allowance for disagreements that are expected by chance, assuming nothing about the properties of the data in question, and depriving researchers, as already stated, of a meaningful second anchor for their reliability scale. A_o cannot indicate the absence of reliability, as called for in (3).

Osgood's (1959, p.44) coefficient, named CR by Holsti (1969, p.140), amounts to the product of two proportions, the %-agreement, A_o, equivalent to 1–(b+c), and the proportion of the number $N_{A\cap B}$ of units coded jointly to the average number of those coded individually, N_A and N_B. Unlike the other coefficients reviewed here, Osgood's responds not only to disagreements in coding but also to disagreements in the numbers of identified units, however, without reference to what would amount to the absence of reliability: chance. Thus, Osgood's coefficient suffers from the same problems that %-agreement does.

Bennett et al. (1954) were probably the first to realize that %-agreement becomes more difficult to achieve as the number of available categories increases. Their coefficient, S, corrects for this effect. For just two categories, S calculates the disagreement in the two cells, b and c, that can be expected by chance as 2·½·½ or 50%. Here, ½ is the logical probability of the distinction between category 0 and 1. It is not very flattering to the literature on content analysis that this coefficient has been reinvented with minor variations at least five times since its original proposal: as Guilford's G (Holley & Guilford, 1964); as the R.E. (random error) coefficient (Maxwell, 1970); as C (Janson & Vegelius, 1979); as κ_n (Brennan & Prediger, 1981), and as the intercoder reliability coefficient $I_{r.}$ (Perreault & Leigh, 1989). The authors of the last two derivations at least knew of S. The justifications given in the literature for using this coefficient range from fairness to each category and appropriateness to the discipline of its advocates,[3] to the absence of hard knowledge about the true distribution of categories in the population from which reliability data are sampled. By treating all categories as equally likely, S is insensitive to unequal (non-uniform) distributions of categories in the population of data, fails to respond to disagreements among coders regarding their frequencies of using these categories, becomes inflated by unused categories, and not satisfying (3), it cannot indicate the reliability in the population of data.

Regarding the absence of knowledge about the true distribution of categories in the population of data, it would make good sense, indeed, to calculate expected disagreements from the proportions of categories in that population. After all, it is the nature of the data – not of the coders' proclivity for particular categories or systematic coding habits; not the categorical structure of a coding instrument – that empirical inquiries are ultimately concerned with and on which researchers hope all coders would agree. As stated in (2), content analysts have to accept the epistemological fact that data are knowable only through their descriptions and the true proportions of categories in the population of data remain unknown until the whole population of units of analysis is reliably observed, transcribed, categorized, or coded. Without *a priori* knowledge of the data, according to (3), these proportions must be estimated from the reliability sample, using as many coders as possible (at least two), and assuming that individual differences among them wash out with large numbers. It is standard statistical practice to take the mean of multiple coder judgments on a sample as estimates of the otherwise unknowable population proportions. With only two coders and two categories, as in Figure 1, $\bar{p} = (p_A + p_B)/2$ is the best estimate of the proportion of 0s in the population of data and its complement, $\bar{q} = (q_A + q_B)/2 = (1 - \bar{p})$, is the best estimate of the proportion of 1s in the same data.

In Figure 2, π and α can be seen to be alike in calculating their expected disagreements in cells b and c as $2\bar{p}\bar{q}$, relying on precisely this population estimate, thus satisfying the inductive step of (3) and the interchangeability of coders mentioned in (2). Evidently, nowhere does π and α "assume" that

"coders have distributed their values across the categories identically" as Lombard et al. (2002, p.591) claim. π and α merely estimate the population proportions and calculate their expected disagreements in these terms. Confusing the computation of expected disagreements from population estimates with the assumption that coders have used their categories with identical frequencies and that disagreements between them are ignored is rooted in the failure to recognize that coder interchangeability is necessary to get to the population estimates. In Figure 2, π and α can be seen to refer to the population of data whereas the other coefficients do not.

π and α differ in one respect, in the factor $n/(n-1)$, which is recognizable in α but not in π. n is the total number of categories used to describe all units by all coders. This factor corrects α for the effects of small sample sizes and few coders. Numerically, α exceeds π by $(1-\pi)/n$. But as sample sizes increase, the factor $n/(n-1)$ converges to 1, the difference $(1-\pi)/n$ converges to 0, and π and α become asymptotically indistinguishable.

Turning now to κ, its expected disagreement differs from π's and α's. The sum of $p_A q_B$ and $p_B q_A$ compute the proportions in cells b and c that can be expected under the condition that coders A and B are statistically independent. κ does this, much like the familiar χ^2 statistic does. The latter is used to test null-hypotheses regarding associations, not agreement. Thus, and in violation of the second part of (2), κ's expected disagreement is a function of the individual coder preferences for the two categories 0 and 1, not of the estimated proportions $\bar{p}$ of 0s and $\bar{q}$ of 1s in the population of data. This expected disagreement renders κ zero when the two coders' use of categories are statistically independent. As κ deviates from perfect agreement, it becomes increasingly determined by coder preferences and says less about the data it is to evaluate.

I suggested elsewhere (Krippendorff, 1978) that κ is a hybrid coefficient. It enters the observed disagreement just as all agreement measures do but corrects this by a conception of chance that derives its logic from association measures. This inconsistency explains why κ behaves so oddly in the numerical examples in Figure 3 below. But, faced with this characterization of κ, Fleiss (1978, p. 144), a major proponent of κ, conceded that when coders are interchangeable, π (and α) would be the correct measure of reliability. The use of κ, he wrote, should be restricted to reliability studies in which one pair of coders judge all units of analysis and unequal coder preferences are not problematic. Thus, κ fails to recognize that the two coders' unequal uses of categories could be a reliability problem. Notwithstanding κ's popularity, the amount of research devoted to this coefficient, and the interpretations that Lombard et al. cite from the literature, the mathematical structure of Cohen's κ is simply incommensurate with the logic of the situation that content analysts are facing when the reliability of their data is in question. κ cannot be recommended as one of several alternative indices, as Lombard et al. are suggesting.

As seen in Figure 2, Benini's (1901) β[4] differs from κ only in its subtracting the absolute difference $|b-c|$ from both, κ's observed and expected disagreements. This adjustment to κ preserves its reliance on the statistical independence of the two coders and therefore disqualifies it from being interpretable as an index of the reliability of data. The importance of this seemingly small adjustment is that β, unlike κ, carries its dependence on the coders' unequal use of categories to its logical conclusion, measuring 1.000 when agreement is the largest one possible, given these coders' marginal distributions. This might not be so easily recognizable in the mathematical form of Figure 2, but the behavior that follows from it might become clear in Figure 3 below.

Common Misinterpretations of κ and π and Their Behavior

In comparing Scott's π with Cohen's κ, Lombard et al. cite Craig (1981), Hughes and Garnett (1980), and Neuendorf (2002) to which one could add several others (notably Fleiss, 1981, p. 218), claiming that π "does not account for differences in how the individual coders distribute their values

across the coding categories, a potential source of systematic bias" and that "it (π) assumes the coders have distributed their values across the categories identically, and if this is not the case, the formula (for π) fails to account for the reduced agreement" (2002, p. 591). Comparing π with κ, the authors maintain that the way κ multiplies the marginal proportions "has the effect of accounting for differences in the distribution of values across categories for different coders" (p. 592). However, just the opposite is correct. It is κ, not π, that fails to count the observed disagreements among coders regarding their individual preferences for particular categories as errors. A simple demonstration will suffice. Compare the first two contingency tables of frequencies in Figure 3.

Figure 3
Three Contingency Tables with Equal/Unequal Margins and Largest Agreement

Categories:		Coder A a	b	c		Coder A a	b	c		Coder A a	b	c	
	a	12	9	9	30	12	18	18	48	12		36	48
Coder B	**b**	9	14	9	32	0	14	18	32		32		32
	c	9	9	20	38	0	0	20	20			20	20
		30	32	38	100	12	32	56	100	12	32	56	100
		A_o = .460				A_o = .460				A_o = .640			
		π = .186				π = .186				π = .457			
		κ = .186				κ = .258				κ = .506			
		β = .186				β = .511				β =1.000			

These two tables are identical in the %-agreement they exhibit but differ in how their mismatching categories are distributed in these tables. In the left table, coder A and B agree on their marginal frequencies; in the right table, they do not. When they do agree, π, κ and β are equal, as they should be. But when coders disagree on these frequencies, when they show unequal proclivities for the available categories, as is apparent in the margins of the table in the middle, κ exceeds π. κ does not ignore the disagreements between the coders' use of categories, but adds it to the measure as an agreement! This highly undesirable property benefits coders who disagree on these margins over those who agree and it clearly contradicts what its proponents (Cohen, 1960; Fleiss, 1975) argued and what Lombard et al. (2002) have found to be the dominant opinion in the literature. Evidently, there are still 46 out of 100 units with matching categories in the diagonal cells. What accounts for this difference is that the 54 mismatches, occupying the cells of both off-diagonal triangles in the left table of Figure 3, have now migrated to one off-diagonal triangle in the center table. It makes for an uneven distribution of the mismatching categories, increasing not agreement, but the predictability of the mismatching pairs of categories. Unlike κ, π is evidently not affected by where the mismatching categories occur, satisfying (2) by not distinguishing who contributed which disagreements and, when data are nominal, which categories are confused. Predictability has nothing to do with reliability.

Figure 3 demonstrates another peculiarity of κ. Not only does κ counter intuitively exceed π when disagreements in marginal frequencies are present, unlike β, κ cannot reach 1.000 when such disagreements exist. This already had been observed by Cohen (1960), noted as a drawback by Brennan and Prediger (1981) and others, and may also be seen in the right table of Figure 3. This table has the same marginal frequencies as the one in the center but exhibits the largest possible agreement, given the marginal constraints. Under these conditions, κ cannot exceed .505, its largest possible value for these marginal frequencies. By contrast, β registers this very condition by measuring 1.000.[5]

I am less concerned with this additional peculiarity of κ, except to note that β is always equal to or larger than κ and κ is always equally equal to or larger than π. Having shown the reasons for these inequalities, both in mathematical terms and by numerical examples, characterizing these coefficients in terms of the aforementioned conservative/liberal dimension would be besides the point of this demonstration. When the reliability of data is the issue, κ is simply wrong in what it does. Its behavior clearly invalidates widely held beliefs about κ, which are uncritically reproduced in the literature.

I have to say that the above misinterpretation of κ goes back to its inception. To justify his unfortunate modification of Scott's (1955) π, Cohen incorrectly criticized π for ignoring "one source of disagreement between a pair of judges, … their proclivity to distribute their judgments differently over the categories" (1960, p. 41). Figure 3 showed that κ behaves contrary to what Cohen had intended. Instead of including this error as disagreement, κ credits this error towards agreement. Brennan and Prediger (1981) observed this highly undesirable property of κ as well, pointing out that "two judges who independently, and without prior knowledge, produce similar marginal distributions must obtain a much higher agreement rate to obtain a given value of kappa, than two judges who produce radically different marginals. … [The former judges] are in a sense penalized" (p. 692) for agreeing on marginal frequencies. Zwick (1988) has considered this statistical artifact. Her advice to users of κ is to test for unequal margins before applying κ. Its violating (2) and (3) renders κ just about worthless as a reliability index in content analysis. The same can be said about β, although I have not heard anyone claiming as much.

Numerical Comparisons

Following their own recommendation to compute several agreement coefficients and to find a balance between conservative and liberal coefficients, Lombard et al. calculated the values of the four aforementioned indices, %-agreement, π, κ, and α, for 36 of their variables. Their corrected table (2003, pp. 470-471) provides good empirical examples for discussing what their numerical differences mean. However, since all content analysts work hard to achieve reliable data, such a table cannot possibly reveal the full ranges of these coefficients. Therefore, let me state them generally:

$$0 \leq \%\text{-agreement} \leq 1$$

$$-1 \leq \pi, \alpha, \text{ and } \kappa \leq +1$$

For nominal variables, which account for the majority of the authors' data, their inequalities are:

$$\%\text{-agreement} \geq\geq \kappa \text{ and } {}_{\text{nominal}}\alpha \geq \pi$$

Careful readers of Lombard et al.'s corrected table will notice the small differences among the three chance-corrected agreement coefficients and might come to the seriously mistaken conclusion that the choice among these coefficients would not matter much. However, even small differences mean rather different things, starting with their zero values:

%-agreement = 0: one coder describes all units of analysis in terms not chosen by the other

π = 0: multiple descriptions are chance events, assuming large numbers of units of analysis

α = 0: multiple descriptions are chance events, adjusted for variable numbers of units and coders

κ = 0: coders are statistically independent of each other, assuming large numbers of units of analysis

As already stated, when the sample size is large, theoretically infinite, ${}_{\text{nominal}}\alpha = \pi$. Otherwise, ${}_{\text{nominal}}\alpha$ exceeds π by $(1-\pi)/n$, which corrects ${}_{\text{nominal}}\alpha$ for small reliability sample sizes. With the authors' sample size of n=256 (2 coders × 128 units), that difference is noticeable only in the third digits. Smaller samples would result in larger differences.

As above demonstrated, when coders agree on their use of categories, on their marginal distributions, κ = π. When coders disagree regarding these distributions, κ exceeds π, responding to the increased predictability of one coder's categories from those of the other. Predictability has nothing to do with reliability measures and must not contaminate them. In the authors' table, the values of κ and π

turn out to barely differ, suggesting that the two coders exhibit only small marginal differences. However, Figure 3 shows that such differences could be much larger.

Lombard et al. also report the reliabilities for ordered data. If agreement concerns ordered reliability data – ranks, intervals, and proportions – an agreement coefficient that is appropriate to these data utilizes this information and can be expected to exceed nominal coefficients, which ignore that information. α is applicable to metrics other than nominal; %-agreement, π, and κ are not. In the authors' table, the names of variables with ratio metrics are superscripted "b." %-agreement, π, and κ are inappropriate for these variables. However, since the authors happen to calculate these coefficients, comparing them with the values of the $_{ratio}\alpha$ coefficient may show the reader how much %-agreement, π, and κ respectively omit.

Consequences of Lombard et al.'s Reliability Standards

As already mentioned, Lombard et al. (2002) applied the following criterion for accepting content analysis findings as sufficiently reliable: $\alpha \geq .70$, otherwise %-agreement $\geq .90$ (p. 596). They take α as a conservative index and %-agreement as a liberal one, presumably convinced that the truth lies somewhere between these two. Their findings, listed in terms of absolute and relative frequencies (percentages) for the above-mentioned 36 variables in their corrected table (2003, pp.470-471), are the result of applying this criterion to their own data. The authors do not report findings for variables that lacked reliability, according to their criterion, which is fair and a common practice. To see what passes their criterion, let me examine the data on two of these variables, starting with the 25th, mentioning "'simple agreement' only." The reliability data for this variable are tabulated in Figure 4.[6]

Figure 4
Reliability Data on the Agreement Coefficient Used: "'Simple Agreement' Only"

		Coder C			
Categories:		**0**	**1**		
Coder J	**0**	83	1	84	+ 3 without a match by C } $N_J = 89$
	1	2	0	2	
		85	1	86	$= N_{C \cap J}$
		+ 1 without a match by J			
		$N_C = 87$			

$A_o = .965$
$CR = .943$
$S = .930$
$\alpha = -.012$
$\pi = -.016$
$\kappa = -.018$
$\beta = -.024$

Figure 4 also lists the value of Osgood's coefficient (Holsti's CR), which Lombard et al. discuss but do not report[7] and of Benett et al.'s S, and Benini's β for comparison.

The 0-0 cell in this table shows the two coders agreeing that this category was absent in 83 articles. Its 0-1 and 1-0 cells indicate a total of three cases of one coder identifying this category while the other did not. And in four cases, one coder noted the absence of this category while the other abstained from coding the article. The four chance-corrected agreement coefficients for these data are

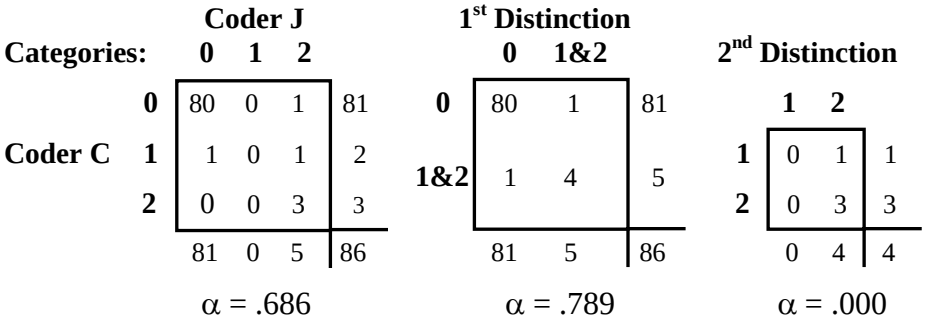
near zero, suggesting the virtual absence of reliability. Yet, the authors' decision criterion suggest otherwise. Unable to accept the data on account of α=–.012, which measures significantly less than .700, the criterion relies on the fact that the %-agreement of 96.5% is well above the 90% that Lombard et al. require and so, the authors feel justified in accepting this variable as reliable and report that 1% (or 2/137) of the articles they examined mention "simple agreement" only (2003, p.471)[8].

Note that in Figure 4, all 96.5% coincidences pertain to absences, the 0s. Regarding the 1s of the variable mentioning "'simple agreement' only," which the authors report as their findings, the two coders do not agree at all, not even once! The 1-1 cell in Figure 4 is completely empty. And in the three cases in which one coder identifies "'simple agreement' only," the other does not. If the %-agreement measure would be allowed to go down to 90%, the number of mismatches could triple without shaking the authors' confidence in the reliability of the reported finding. Eighty-six out of 137 units of analysis is a decent reliability sample, but could one trust a claim that the 137 articles in the data contained two mentions of this category when coders cannot agree on even one? In the calculation of reliability, large numbers of absences should not overwhelm the small number of occurrences that authors care to report.[9] Without a single concurrence and three mismatches, the report of finding 2 out of 137 cases is about as close to chance as one can get – and this is born out by the near zero values of all the chance-corrected agreement coefficients.

For Lombard et al., this case was not an oversight. In their Table 1 (2002, p. 592), they reproduce Perrault and Leigh's hypothetical 2-by-2 data (1989, p. 139) with very uneven marginal frequencies that yield κ=.000 while showing 82% agreement – just to argue for the conservative nature of κ, and by extension, of all chance-corrected agreement measures. The marginal frequencies in the table of Figure 4 are even more uneven. Yet most striking and often mystifying those who hold on to the %-agreement conception is the case in which all coders use one and the same category for all units of analysis, yielding 100% agreement. Such data can be obtained by broken instruments or coders who fell asleep or agreed in advance of the coding effort to make their task easy. As suggested in (3), appropriate indices of reliability cannot stop at measuring agreement but must infer the reproducibility of a population of data; and one cannot talk about reproducibility without evidence that that it could be otherwise. When all coders use only one category, there is no variation and, hence, no evidence of reliability. In the case of the slightly less extreme data in Figure 4, Lombard et al.'s criterion for accepting data as reliable clearly fails to warn researchers about significant unreliabilities in data and induces a false sense of certainty about the conclusions drawn from these data when they actually are indistinguishable from chance events. Their criterion for accepting data as reliable does not separate the wheat from the chaff. The use of %-agreement should be actively discouraged, especially as a fallback criterion. Instead, I recommend that only chance-corrected agreement coefficients that satisfy (2) and (3) be used for inferring the reliability of data.

Because agreement coefficients are averages over the categories in a variable, which allows unreliable categories to hide behind reliable ones, I am suggesting that reliabilities be obtained for all distinctions that matter. To state proportions of frequencies, the distinctions between these categories and their complements need to be reliable. If differences in frequencies of two categories are to be reported, the two categories must be reliably distinguishable. Overall agreement measures applied to a multi-category variable do not provide such assurances. For a simple numerical example, consider one of Lombard et al.'s variables, the 20th, recording whether articles report reliability figures (2003, p. 470). It recorded data in three categories: 0=No; 1=Yes together with findings; and 2=Yes separately[10] and measures α=.686. This borderline measure should signal doubt. The data for this variable are reproduced in the left table of Figure 5.

Figure 5
Reliability Data on Whether Article Reports Reliability Figures and Two Distinctions



If its categories were equally unreliable, one could let the overall reliability of the variable stand. When this is not the case and when all categories are equally important to a research effort, one has to find the least reliable category. This can be done by computing the reliabilities for all distinctions in a variable, here, between any one category and the remaining categories lumped into one, also called individual category reliability. Lumping categories 1 and 2 and evaluating the distinction between 0 and 1&2, as shown in the center table of Figure 5, yields $\alpha(0|1\&2)=.789$, which is significantly larger than the overall $\alpha=.686$, suggesting also that this variable contains other categories that are less reliable than category 0. With $\alpha(1|0\&2)= -.006$ and $\alpha(2|0\&1)=.739$, category 1 turns out to be the unquestionable culprit. If the three categories were not equally important, if one could restrict the findings to the distinction between 0=absent and 1&2=present, the correct reliability would be $\alpha=.789$, not .686. The subordinate distinction between 1 and 2, whose data are shown in the right table of Figure 5, is a perfect chance event, $\alpha=.000$ exactly. Notwithstanding the low frequencies in the latter distinction, this analysis would render it a mistake to report on any difference in the frequencies of categories 1 and 2.[11] Should this distinction be important, the variable must be rejected for not exceeding chance. When it is ignored, the overall measure would be inaccurate. In other words, reliability should assess all relevant details and not be contaminated by including irrelevant distinctions, which can overstate or understate the reliability of what matters.

Multiple Coders, Multiple Coding Sets, Multiple Metrics

Amplifying Neuendorf (2002, p. 163) who merely quotes a concern expressed elsewhere about the appropriateness of using different coders for coding different but overlapping sets of units, Lombard et al. (2002) make it a point of recommending against this attractive possibility (p. 602) – without justification, however. I can imagine three: (a) Potter and Levine-Donnerstein (1999) argue that the overlap needs to be large enough, which is correct. (b) Fleiss (1974) advised that "the use of κ … [be] restricted to reliability studies involving the same pair of judges." This restriction applies only to the use of κ. It is not methodologically motivated and not generalizable to other coefficients. Finally, (c), if reliability data are generated by multiple coders of different but overlapping sets of data, the practice of averaging agreement measures among pairs of coders would actually average unrelated reliability data. The software PRAM[12], on which Lombard et al. relied for parts of their calculations, does just this. So, the injunction would make sense for this makeshift approach to calculating multi-coder agreements. However, I cannot see any methodological justification for the authors' proposed injunction.

It should be noted that α is designed for the very situation that the authors seek to rule out (i.e., for variable numbers of interchangeable coders, including when coding different but overlapping sets of units, causing data to be missing). The authors acknowledge the ability of α to accommodate multiple coders and all common metrics or scales of measurement, not just the situation of two coders

and nominal data to which comparisons of $_{nominal}\alpha$ with the other nominal coefficients is limited. However, more important is to realize that α is a large family of agreement coefficients with identical assumptions about reliability, yielding measures that are comparable across a diversity of data – not to be confused with comparing the numerical results of coefficients with incompatible assumptions along a continuum. Although α reaches far beyond measures known in the literature, it embraces several known coefficients. For two coders and large sample sizes, α reduces not only Scott's π, which is limited to nominal data, as Figure 2 demonstrated, but also to Spearman's ρ (rho) without ties, which is defined for rank orderings, and to Pearson's intraclass correlation, R_I, appropriate for interval data – not his product moment correlation, r_{ij}. Thus, as a family, α can be compared across different metrics, and enables content analysts to apply identical decision criteria to them. Researchers may use α conservatively or liberally, as they please; α, in and of itself, is neither.

Recommendations

Let me conclude with four recommendations for establishing the reliability of given data, measured by the degree to which a coding process is reproducible with different coders, elsewhere, and under conditions that should not affect the results:[13]

(i) *Reliability data*, the sample of data from which the trustworthiness of a population of data is to be inferred, have to be generated by coders that are widely available, follow explicit and communicable instructions (a data language), and work independently of each other. Reliability data must be representative of the data whose reliability is in question (not of the population of ultimate research interest); and the more coders participate in the process and the more common they are, the more likely can they ensure the reliability of data. Coders must be interchangeable, may code different subsamples of data, provided there is enough duplication or overlap.

(ii) *A decisive agreement coefficient* should measure agreements within multiple descriptions, regardless of numbers and kinds of coders. Its values should be indicative of the likelihood that conclusions drawn from imperfect data are valid beyond chance. For two coders, large sample sizes, and nominal data, π is such a coefficient. When data are ordered, it is advantageous to select a coefficient that responds to the information in their metric (scale characteristic or level of measurement) but assumes not more than warranted by the data in hand. α can handle multiple coders, nominal, ordinal, interval, ratio, and other metrics, missing data, and small sample sizes. Content analyses that assess reliability in terms of any association coefficient, Pearson's r, for example, Benini's (1901) β, Cohen's (1960) κ, Cronbach's (1951) alpha, Goodman and Kruscal's (1954) λ_r (lambda r), and %-agreement should be rejected as these measures are incompatible with reliability concerns in content analysis. For any other measure and when in doubt, the mathematical structures of proposed indices should be examined for their ability to shed light on the reproducibility of the data making process. Unsubstantiated claims should be questioned.

(iii) *An acceptable level of agreement* below which data are to be rejected as too unreliable must be chosen depending on the costs of drawing invalid conclusions from these data. When human lives hang on the results of a content analysis, whether they inform a legal decision or tip the scale from peace to war, decision criteria have to be set far higher than when a content analysis is intended to merely support scholarly arguments. In case of the latter, to be sure that the data under consideration are at least similarly interpretable by other scholars (as represented by different coders), I suggested elsewhere to require $\alpha \geq .800$, and where tentative conclusions are still acceptable, $\alpha \geq .667$ (Krippendorff, 2004, p. 241).[14] Except for perfect agreement, there are no magical numbers, however. The ones suggested here should be verified by suitable experiments. To ensure that the measured agreement is representative of the data in question, confidence intervals should be consulted. Testing the null-hypothesis that observed agreement deviates from chance has no bearing on reliability, which concerns deviations from perfect agreement or 1.000.

(iv) *All distinctions that matter should be tested for their reliability***.** Where a system of several variables is intended to support a conclusion (e.g., as in an index, a regression equation, or any multi-variate analysis), the reliability of each variable should be measured and the smallest among them should be taken as the reliability of the whole system. Averaging the agreement measures of several variables, especially when they include easily coded clerical ones, can easily mislead researchers about the reliability of variables that matter. This logic applies to individual categories as well. Where differences in frequencies of the categories of a variable influence the conclusions of a research effort (e.g., in reports on differences, changes, or proportions – as exemplified in Lombard et al. [2003]), the reliability of each distinction should be tested and the smallest one should be taken as the reliability of the whole variable. This may not be required when a subsequent analysis concerns variances (e.g., in tests concerning correlations or associations), which are averages, just as measures of agreement of multi-category variables are. After data have been generated, reliability may be improved by discarding unreliable distinctions, recoding or lumping categories or dropping variables that do not meet the criterion adopted in (iii). Resolving disagreements by majority among three or more coders may make researchers feel better about their data, but does not affect the measured reliability (Krippendorff, 2004, p. 219)**.**

I commend Lombard, et al. (2002, 2003) for bringing the sad state of reliability testing to the attention of content analysts. The above criticism is directed less to the authors then to the literary practices in communication research. As a critical scholar, I defend the principle of encouraging multiple voices to speak through a text. However, when it comes to discussing mathematical objects, such as agreement measures and their use as indices of the reliability of data, mathematical proofs and demonstrations should speak louder than majority opinions, even when published in respectable journals. Quoting from the work of other scholars does not absolve our responsibility for investigating and judging what we are reproducing.

References

Benini, R. (1901). *Principii di Demographia*. Firenze: G. Barbera. No. 29 of *Manuali Barbera di Scienze Giuridiche Sociali e Politiche*.

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

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. *Psychometrica, 16*, 297-334.

Craig, R. T. (1981). Generalization of Scott's index of intercoder agreement. *Public Opinion Quarterly, 45*, 260-264.

Fleiss, J. L. (1975). Measuring agreement between two judges on the presence and absence of a trait. *Biometrics 31*, 651-659.

Fleiss, J. L. (1978). Reply to Klaus Krippendorff's "Reliability of binary attribute data." *Biometrics, 34*, 144.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. New York: Wiley & Sons.

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

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

Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Reading, MA: Addison-Wesley.

Hughes, M. A., & Garrett, D. E. (1990). Intercoder reliability estimation – Approaches in marketing: A generalizability theory framework for quantitative data. *Journal of Marketing Research, 27*, 185-195.

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. (1970). Bivariate agreement coefficients for reliability data. In E. R. Borgatta & G. W. Bohrnstedt (Eds.), *Sociological methodology 1970* (pp. 139-150). San Francisco, CA: Jossey Bass.

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

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology. Second Edition*. Thousand Oaks, CA: Sage.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication research: An assessment and reporting of intercoder reliability. *Human Communication Research, 28*, 587-604.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2003). Correction. *Human Communication Research, 29*, 469-472.

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

Neuendorf, K. A. (2002). *The content analysis guidebook.* Thousand Oaks, CA: Sage.

Osgood, C. E. (1959). The representational model and relevant research. In I. de Sola Pool (Ed.), *Trends in content analysis* (pp. 33-88). Urbana: University of Illinois Press.

Pasadeos, Y., Huhman, B., Standley, T., & Wilson, G. (1995, May). *Applications of content analysis in news research: A critical examination*. Paper presented to the annual convention of the Association for Education in Journalism and Mass Communication, Washington, DC. Cited in M. Lombard, J. Snyder-Duch, & C. C. Bracken (2002). Content analysis in mass communication research: An assessment and reporting of intercoder reliability. *Human Communication Research, 28*, 587-604.

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

Potter, W. J., & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research, 27*, 258-284.

Riffe, D., & Freitag, A. (1996, August). *Twenty-five years of content analyses in Journalism & Mass Communication Quarterly*. Paper presented to the annual convention of the Association for Education in Journalism and Mass Communication, Anaheim, CA, cited in D. Riffe, S. Lacy, & F. G. Fico (1998). *Analyzing media messages.* Mahwah, NJ: Erlbaum.

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

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

Endnotes

[1] The authors used a beta version of the software package PRAM, an acronym of "Program for Reliability Assessment with Multiple-coders" (Skymeg Software, 2002), also described by Neuendorf (2002, pp. 241-242), to calculate %-agreement, π, and κ, and a separate unpublished software to calculate α (Lombard et al., 2002, p. 596).

[2] Lombard et al. (2002, p.590) claim that %-agreement can also be computed for any number of coders without explaining how this could be accomplished. The aforementioned software PRAM includes a feature to average pairwise %-agreements. This average of averages cannot express patterns of disagreement that inevitably arise when multiple coders are involved and becomes of dubious validity when coders code different sets and numbers of units.

[3] For example, Perreault and Leigh argue that "in most marketing research studies (and in many other areas of applied research), there is no *a priori* knowledge of the likely distribution or responses" (1989, p. 139), much as I said in (2), but they then proceed to define expected disagreement as in S, in terms of the number of available categories, which is equivalent to assuming categories to be uniformly distributed.

[4] Since β is less familiar than the other coefficients, I offer this definition: $\beta = \dfrac{\sum_i p_{ii} - \sum_i p_{Ai}p_{Bi}}{\sum_i \min(p_{Ai}, p_{Bi}) - \sum_i p_{Ai}p_{Bi}}$, where, in a contingency table, i is a generic category, p_{ii} is the proportion of pairs of matching categories i, and p_{Ai} and p_{Bi} are the marginal sums for category i used by coder A and B respectively.

[5] It might be noted that Cohen (1960), probably unfamiliar with Benini's β, discussed a ratio κ/κ_{max} (p. 43), which equals β. I have not seen it used, however.

[6] On December 12, 2002, Matthew Lombard kindly made the authors' data available to me and in return received the recalculations of α.

[7] In a footnote to the original table, Lombard et al. write, "Holsti's method is not reported because it is identical to Scott's pi in the case of two coders evaluating the same units" (2002, p. 598). In the revised table, the authors replaced "Scott's pi" by "percent agreement" (2003, p. 471), which makes this statement a mathematical possibility, but one that is not born out by their data. Figure 4 shows reliability data in which one coder categorized N_C=87 articles, another categorized N_J=89 articles, and both categorized $N_{C\cap J}$=86 articles, rendering Osgood's coefficient (Holsti's CR) $= A_o \dfrac{2N_{C\cap J}}{N_C + N_J} = .965\dfrac{2\cdot 86}{87+89} = .943$.

[8] The original table reports 2% (Lombard, et al. 2002, p.579). I do not know what prompted this revision.

[9] Arguably, 99% is a large proportion and 1% is a small one. Considering small errors, say ±1%, 99±1% still defines a large proportion with a relatively small error, but 1±1% refers to a small proportion with a relatively large error. Thus, a range between 0% and 2% seems more severe than a range between 98% and 100%.

[10] http://astro.temple.edu/~lombard/carman.htm, accessed in January 2003.

[11] Lombard et al. are not explicit about the 6% (8 articles) they report as containing reliability information (2003, p.470). I presume, however, it refers to categories 1&2 lumped together, in which case the proper reliability should have been computed with data on the 1st distinction, not for the whole variable, and reported as .739, not as .686 – widespread practice notwithstanding.

[12] PRAM, op. cit.

[13] These recommendations do not agree with Lombard et al.'s (2002) guidelines 2, 4, parts of 8 and 9, the common practice of calculating average reliabilities for multi-category variables of which the frequencies and proportions (%) of individual categories are reported, but particularly not with criterion they have adopted in accepting their own findings as reliable (p. 596; pp. 600-602).

[14] These standards were suggested for α, and the experiments that led to them concerned α only. Other coefficients may require different standards. Setting standards for all coefficients alike, even discussing them as if that made sense, glosses over their mathematical differences and the assumptions that go into their construction. This would apply also to conceptualizing agreement coefficients on a conservative/liberal continuum according to the numerical results they produce, discussed above.