ASSUMPTIONS BEHIND INTER-CODER RELIABILITY INDICES

Xinshu Zhao, Fudan University and Hong Kong Baptist University

Jun S. Liu and Ke Deng, Harvard University

Forthcoming in *Communication Yearbook 36*, May 2012

Suggested citation

Zhao, X., Liu, J. S., & Deng, K. (in press). Assumptions behind inter-coder reliability indices. *Communication Yearbook 36.*

Authors' Note

Xinshu Zhao is Cheung Kong Chair Professor of Journalism, Fudan University in Shanghai. He is also Chair Professor of Communication, Director of HongComm Survey, and a Co-Director of the Carter Center Initiative at Hong Kong Baptist University. He was Professor and Director of the Center for Research in Journalism and Mass Communication at University of North Carolina at Chapel Hill when the main theses of this article were developed. He has over 100 publications, including a book in Chinese, *Plight of Elections - A Critique of the World's Election Systems and the Constitutional Reforms,* expanded (3rd) edition, Sichuan People's Publishing House, 2008**,** ISBN 978-7-220-07537-7**.**

Jun S. Liu is Professor of Statistics and Biostatistics at Harvard University. Liu is also an elected Fellow of the Institute of Mathematical Statistics and of the American Statistical Association. He won the National Science Foundation Career Award in 1995, received the COPSS (Committee of Presidents of Statistical Societies) Presidents' Award in 2002, and was conferred the Morningside Gold Medal (awarded at the International Congress

of Chinese Mathematicians) in 2010. He served as associate editor and co-editor for the Journal of the American Statistical Association. He has over 150 publications, including a book, *Monte Carlo Strategies in Scientific Computing*, Springer: New York, 2001.

Ke Deng is Research Associate of the Department of Statistics, Harvard University. His research interests include statistical modeling, statistical computation and applications in bioinformatics, text mining and sociology. He has 8 publications.

This study was supported in part by HKBU Faculty Research Grant (2008 & 2009, Zhao PI), HKBU Strategic Development Fund (2009 & 2011, Zhao PI), and grants from Panmedia Institute (2010, Zhao PI) and ENICHD (R24 HD056670, Henderson PI). The authors acknowledge with gratitude the substantial contributions of Guangchao Charles Feng, and thank Jane Brown, Visne K.C. Chan, Timothy F. Hamlett, Sri Kalyanaraman, Juntao Kenneth He, Jing Lucille Li, and Ning Mena Wang for their support and assistance. The authors also thank the editor and reviewers of *Communication Yearbook 36*, whose questions, criticism, and suggestions helped to significantly improve this article.

Correspondence concerning this article should be addressed to Xinshu Zhao at zhao@hkbu.edu.hk, Tel: 852-3411-7492, Fax: 852-3411-7375.

ASSUMPTIONS BEHIND INTER-CODER RELIABILITY INDICES

Abstract

Inter-coder reliability is the most often used quantitative indicator of measurement quality in content studies. Researchers in psychology, sociology, education, medicine, marketing and other disciplines also use reliability to evaluate the quality of diagnosis, tests and other assessments. Many indices of reliability have been recommended for general use. This article analyzes 22, which are organized into 18 chance-adjusted and four non-adjusted indices. The chance-adjusted indices are further organized into three groups, including nine category-based indices, eight distribution-based indices, and one that is double based, on category and distribution.

The main purpose of this work is to examine the assumptions behind each index. Most of the assumptions are unexamined in the literature, and yet these assumptions have implications for assessments of reliability that need to be understood, and that result in paradoxes and abnormalities. This article discusses 13 paradoxes and nine abnormalities to illustrate the 24 assumptions. To facilitate understanding, the analysis focuses on categorical scales with two coders, and further focuses on binary scales where appropriate. The discussion is situated mostly in analysis of communication content. The assumptions and patterns that we will discover will also apply to studies, evaluations and diagnoses in other disciplines with more coders, raters, diagnosticians, or judges using binary or multi-category scales.

We will argue that a new index is needed. Before the new index can be established, we need guidelines for using the existing indices. This article will recommend such guidelines.

Table of Contents

ASSUMPTIONS BEHIND INTER-CODER RELIABILITY INDICES

Content has always been a central concern of communication research. Wilbur Schramm (1973), whom Tankard (1988) called "the father of communication studies," authored *Men, Messages, and Media: a Look at Human Communication*, where "message" meant content. Harold Lasswell (1948), whom Schramm considered one of the "four founding fathers of the field" (Glander, 2000, Ch. 3), defined the discipline as studying "who says what, through which channels, to whom, and with which effect," where "what" is content. With the explosion of "netted" information from increasingly diversified sources, the need for content research has been rising sharply (Neuendorf, 2002).

Modern *content analysis*, a term no more than 70 years old according to Krippendorff (2004a), focuses on "what *is* the content," as supposed to what *should be*. With this empirical emphasis, *validity* and *reliability* have emerged as two methodological pillars. *Validity* addresses whether an instrument measures what it purports to measure. *Reliability* addresses whether the instrument produces consistent results when it is applied repeatedly, i.e. test-retest reliability, or by different people, i.e., inter-coder reliability. While a reliable measure is not necessarily valid, an unreliable measure cannot be valid.

Validity is more difficult to measure numerically. Hence reliability, especially the less costly inter-coder reliability, has been the most popular quantitative indicator of measurement quality in content studies. Researchers in education, psychology, sociology, medicine, marketing and other social science disciplines also use reliability to evaluate the quality of diagnoses, tests and other assessments.

The main purpose of this article is to examine assumptions behind 22 indices of inter-coder reliability, most of which are unexamined in the literature. We will report 24 such assumptions, most of which are rarely met in typical research, meaning that the indices have been often used beyond the boundaries for their legitimate use. As a result, paradoxes and

abnormalities arise. We will discuss 13 paradoxes and nine abnormalities to illustrate the assumptions. We will argue that a new index is needed and, until such a new index is forthcoming, guidelines are needed for using the existing indices.

Our analysis will focus on categorical scales with two coders, and further focus on binary scales where appropriate. The discussion will be mostly situated in analyzing communication content. But the assumptions, patterns and recommendations that we will discuss also apply to studies, evaluations or diagnoses in other disciplines with more coders, raters, diagnosticians, or judges using two or more categories.

The calculations and derivations presented in this article were done by the first author initially by hand and then verified by MS Excel programming. All formulae, calculations, interpretations, and proofs were then independently replicated or verified by the third author under the supervision of the second author. Guangchao Charles Feng, a doctoral candidate at Hong Kong Baptist University, conducted a final round of verifications using R programming (2011, v 2.14). Large portions of this manuscript, especially those related to π, κ and α, were previously presented in two conference papers (Zhao 2011a, 2011b).

I. An Overview of the Inter-Coder Reliability Concept

I.1. Reliability and Related Concepts

Krippendorff (2004b) and Lombard, Snyder-Duch, and Bracken (2002) see *agreement* as the indicator of *reliability,* and consider *association* a separate concept. Tinsley and Weiss (1975, 2000) use *correlation* as the indicator of *reliability* and consider *agreement* as separate. Neuendorf (2002) considers *agreement* and *covariation* as two indicators of *reliability.*

We follow Krippendorff and Lombard et al to use agreement as the indicator of inter-coder or test-retest reliability, and we define *agreement* as proximity between measures. On a

categorical scale, if both coders choose the same category for the same case, that is an agreement. If they choose different categories, that is a disagreement. On a numerical scale, the closer the scores are to each other, the higher the agreement. *Correlation* refers to the *covariation* between measures on numerical scales. For instance, on a 0-10 scale, if Coder 2 chooses 0 whenever Coder 1 chooses 9, and chooses 1 whenever Coder 1 chooses 10, there is a very high correlation but a very low agreement.

Association refers to *covariation* between measures on categorical scales. It is typically used when the concept of "inter-variable agreement" is not appropriate, helpful, or sufficient, while agreement is typically used when the concept of "inter-variable association" is not appropriate, helpful, or sufficient. Suppose, of 200 respondents, all 100 whites are urban, and all 100 non-whites are rural. We say the association between ethnicity and residence is at the highest possible, while it does not help as much to talk about agreement. Suppose the data of 200 cases come from a content analysis, in which Coder 1 reports seeing an urban resident whenever Coder 2 does so, and reports seeing a rural resident whenever Coder 2 does so. This signifies complete agreement. Here it is not as informative to talk about association. Suppose the opposite happens, all 100 whites are rural, while all 100 non-whites are urban. The association is equally high. But if the same data are from the two coders, they would indicate that Coder 1 reports seeing urban residents whenever Coder 2 reports seeing rural residents, and reports seeing rural residents whenever Coder 2 reports seeing urban residents. That would be a complete disagreement.

Association and agreement also differ when distributions are even, e.g., when each ethnic group is half urban and half rural, or when two coders agree with each other half the time. Here association is at the lowest possible, while agreement is 50%, half-way between the lowest and the highest possible. Further, when there is no variation within a variable, e.g., when all respondents are of one ethnicity, or they all live in one locale, association is

undefined. Association is covariation, which is impossible when there is no variation. If the same data come from two coders, which means one or both coders chooses only one category, agreement should and can still be calculated. If both coders agree that all respondents are urban, there is one hundred percent agreement. Later we will show that three popular indices of reliability, i.e., π, κ and α, become un-calculable, hence undefined, when coders agree all cases fall into one category. We will argue that should not have happened if the indices were to measure general agreement.

[INSERT TABLE 1 Reliability and Related Concepts HERE]

Table 1 summarizes the relationship between the key concepts. This article will focus on agreement/reliability indices for categorical scales, and further focus on binary scales where appropriate. We will not deal with association measures such as χ^2, or correlational measures such as Pearson's r or r^2.

I.2. Reliability vs. Reliabilities

Popping (1988) identified no less than 39 reliability indices, although some of them are association measures or correlational measures, and some are the same indices under different names. This article will review 22 indices of inter-coder reliability. Many of the 22 are mathematically equivalent, giving us 11 unique indices.

It is assumed that the various indices are indicators of the same concept of inter-coder reliability. Yet the indices produce different—often drastically different—results for the same underlying agreements. As reliability means agreement (Riffe, Lacy, & Fico, 1998), these indices of reliability do not appear reliable themselves.

Under the premise of "various indicators for one reliability," methodologists debate which indicator is the best, whether to use this, that, or several of them in a study, and how to fix or cope when some indices, especially Cohen's κ, behave paradoxically (e.g., Brennan &

Prediger, 1981; Krippendorff, 2004b; Lombard et al., 2002; Zwick, 1988). This review takes a different approach. As the indices produce different results, we suspect there may be multiple reliability concepts, each having one indicator. No more than one index can be the general indicator, while others are for special conditions. Like mediation researchers (e.g. Hayes, 2009; Zhao, Lynch & Chen, 2010) who examined the dominant approach to reveal its hidden premises, this article analyzes each index of inter-coder reliability to uncover its assumptions, which defines the boundaries for its legitimate use, and may explain the paradoxes and abnormalities that arise when it is used beyond the boundaries.

II. A Typology of 22 Indices

The 22 indices we will review fall into two groups. The first group, called non-adjusted indices, includes *percent agreement* (a_o, pre 1901), *Holsti's CR* (1969), *Osgood's* coefficient (1959) and Rogot and Goldberg's A_1. The first three are mathematically equivalent to each other. The four indices assume that all coding behavior is honest, observed agreements contain no random chance coding, hence there is no need to adjust for chance. The second group are known as *chance-adjusted* indices. These 18 indices assume that coders deliberately maximize random chance coding, and limit honest coding to occasions dictated by chance, so the resulted chance agreement must be estimated and removed.

[INSERT TABLE 2 A Typology of 22 Inter-coder Reliability IndicesHERE]

The chance-adjusted group includes three subgroups. The first subgroup of nine indices estimates chance agreement as a function of category in a measurement scale. The second subgroup of eight indices estimates chance agreement as a function of observed distribution. Here "distribution" refers to the pattern by which cases fall into categories. Distribution can be extremely even, e.g., 50% of the advertisements coded have endorsers and 50% do not, or extremely uneven, e.g. 100% have endorsers and 0% do not, or anywhere

between the two extremes. In reliability literature, this important concept has also been referred to as "frequency" (Gwet, 2008), "base rate" (Grove et al., 1981; Kraemer, 1979; Spitznagel & Helzer, 1985), or "prevalence" (Gwet, 2010; Shrout, Spitzer, & Fleiss 1987; Spitznagel & Helzer, 1985). We will follow Cohen (1960), Perreault and Leigh (1989), and Gwet (2010) to call it distribution. The third subgroup has just one index, which uses both category and distribution as the main factors. Table 2 summarizes this typology.

Six indices, namely ρ, I_r, and four non-adjusted indices range from 0 to 1. The maximum of λ_r is also 1, but it can get far below -1, according to one interpretation. The other 15 indices all range from -1 to 1. All 22 indices consider 1 as indicating maximum reliability, 0 as indicating no reliability, and a below-zero score as a random variation from 0. An important question is where the threshold for acceptable reliability is. This article will focus on estimation of reliability, and leave the threshold issue to future research.

III. Non-Adjusted Indices

Our search found four indices that are not adjusted for chance agreement, including *percent agreement*, two equivalents, and Rogot and Goldberg's A_1.

III.1. Percent Agreement and Two Equivalents

The most intuitive indicator of reliability is *percent agreement*, i.e., the number of cases coders agree (A) divided by the total number of targets analyzed (N). Krippendorff (2004b) and Neuendorf (2002) denote this as a_o:

$$a_o = \frac{A}{N} \qquad (1)$$

Scott (1955, p. 322) observed that a_o was "commonly used." Perhaps because it was so common and intuitive, its early users or critics like Benini (1901) did not mention who

invented it. As Osgood (1959) and Holsti (1969) advocated essentially the same index, many researchers referred to it as *Holsti's CR* while a few called it *Osgood's coefficient* (Krippendorff, 2004b). Bennett et al. (1954) pointed out that a_o contains chance agreements from random guessing, and hence inflates reliability. Experts on reliability (e.g., Lombard et al., 2002; Tinsley & Weiss, 1975) often concurred, revealing an important assumption:

Basic Assumption 1 : Zero chance agreement.

Percent agreement (a_o) assumes no chance agreement in any situation, no matter how difficult the task is, or how tired, bored or unprepared the coders are. This assumption leads to an important paradox:

Paradox 1 : Random guessing can be reliable.

Suppose two coders watch television programs to see if they contain subliminal advertisements, which are flashed quickly to avoid conscious perception. Although the coders try to be accurate, the task is so difficult that their coding amounts to nothing but random guessing. Probability theory expects an a_o=50%, which is the midpoint between 0% for no reliability and 100% for perfect reliability.

Because percent agreement fails to take into account chance agreements, it is often considered "the most primitive" (Cohen, 1960, p. 38) and "flawed" (Hayes & Krippendorff, 2007, P. 80) indicator of reliability, leading to decades-long efforts to "account for" and "remove" chance agreements (Krippendorff, 1980, pp. 133-134; Riffe et al., 1998, pp. 129-130).

Critics of a_o argued that "flipping a … coin" or "throwing dice" would have produced some "chance agreements" (Goodman and Kruskal, 1954, p. 757; Krippendorff, 2004a, p. 114, 226; 2004b, p. 413). A coin only has two sides and a die always has six. Drawing marbles may be a closer analogy, because colors and marbles per color can vary like categories and cases per category can vary in typical content studies (Zhao, 2011a & 2011b).

Hereafter we will use "marble" to refer to any physical or virtual element of equal probability, "urn" to refer to a real or conceptual collection of the elements, and "drawing" to refer to a behavioral or mental process of randomly selecting from the elements. Defined as such, marbles, urns, and drawing turn out to be a set of useful analytical tools. They help to expose assumptions and explain paradoxes and abnormalities that otherwise would be more difficult to uncover or understand.

The no-chance-agreement assumption does not necessarily make percent agreement a bad index, but perhaps a special-purpose index. Some authors argued that, for easy cases or "textbook" cases, all agreements could be from a well-developed protocol (Grove et al, 1981, p. 411; Riffe, Lacy, & Fico, 2005, p.151). In such situations, no chance agreement should be expected; hence percent agreement would be an accurate index. Percent agreement cannot be a general-purpose index because all cases are not easy, and all protocols are not well developed.

III.2. Rogot & Goldberg's A_1

Rogot and Goldberg (1966) noted that, when calculating percent agreement on a binary scale, each positive agreement, e.g., two diagnosticians agree a patient has an abnormality, is given an equal weight as a negative agreement, e.g., diagnosticians agree there is no abnormality. Because abnormality is far less frequent than normality, negative agreements as a group are given more weights than positive agreements. To give the two groups equal weights, Rogot and Goldberg (1966) proposed A_1:

$$\boldsymbol{A_1 = \frac{1}{4}\left(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{c+d} + \frac{d}{b+d}\right)} \quad (2)$$

Here a and d are respectively positive and negative agreements, and b and c are two types of disagreements, all in percentages. $A_1=a_o$ when $a=d$ and $b=c$, that is, when two types

of agreements are evenly distributed and the two types of disagreements are also evenly distributed. When $a \neq d$, that is, when agreements are unevenly distributed, A_1 decreases from a_o, and more uneven distributions bring larger decreases. When $b \neq c$, that is, when disagreements are unevenly distributed, A_1 increases from a_o, and more uneven distributions bring larger increases. Because the decreases and the increases are at the equal rate, the average of A_1 should be close to the average of a_o when each is averaged across many studies and data. As A_1 is just a reweighted a_o, they share the same assumption and paradox as discussed above. In general A_1 is not an improvement over percent agreement. Especially, it still does not take into account chance agreement.

IV. An Overview of Chance-Adjusted Indices

To "account for" and "remove" chance agreement (a_c) from percent agreement (a_o), Equation 3 was introduced to calculate reliability index (r_i). The equation was implied in Guttman (1946) and Bennett et al (1954) and made explicit by Scott (1955):

$$\boldsymbol{r_i = \frac{a_o - a_c}{1 - a_c}} \qquad (3)$$

The subtraction in the numerator appears intuitive. Chance agreement (a_c) needs to be removed from the observed agreement (a_o). The subtraction deflates the otherwise inflated index. The subtraction in the denominator, however, is not as intuitive. Reliability index (r_i) is a percentage, of which the denominator serves as the reference. The full length of the reference is 1 for 100%. The subtraction shrinks the reference, making r_i look larger.

There is a behavioral assumption behind the shrinking. To understand the assumption, we may analyze Equation 4, which was implied in Guttman's ρ (1946), Bennett et al.'s S (1954), Scott's π (1955) and Cohen's κ (1960), and made explicit by Cohen (1968, p. 215):

$$\mathbf{1} = \boldsymbol{a_c} + \boldsymbol{d_c} \qquad (4)$$

With a_c representing chance agreement (%) and d_c representing chance disagreement (%), Equation 4 says chance coding constitutes 100% of all coding. Some may argue that "1" here represents "all chance coding." That is true. But all major reliability indices from Guttman's ρ (1946) to Gwet's AC_1 (2008) all state or imply $a_o+d_o=1$, where a_o is observed agreement and d_o is observed disagreement, hence $a_o+d_o = a_c+d_c$, which means "all coding equals all chance coding," or "all coding is chance."

But *chance coding allows and includes honest coding* in a two-stage process, according to Equations 3 & 4. In the first stage, coders code all cases completely randomly by drawing marbles. If they draw a certain pattern, e.g., the same color, they report findings according to a pre-determined color-category matching scheme. For example, if the marbles are white they would say that an advertisement has an endorser while if the marbles are black they would say there is no endorser, without looking at the advertisement under coding. If and only if the coders draw another pattern, e.g. different colors, they would go to the second stage, during which they would code honestly. Hence honest coding (h) equals chance disagreement (d_c):

$$\boldsymbol{d_c} = \boldsymbol{h} \qquad (5)$$

Here honest coding (h) is defined as percent of cases that coders code by actually examining the objects and categorizing objectively following the instructions during training. Chance coding thus precedes, permits, confines and constrains honest coding. Since honest coding is limited within chance disagreement, it is the chance disagreement, but not all coding, that should be the baseline for percentage calculation. This is why the denominator in Equation 3 should be shrunk from 1 to1-a_c= d_c=h.

Replacing a_o with1-d_o and replacing a_c with 1-d_c in Equation 3, we obtain an alternative expression of r_i (Krippendorff, 1980, p.138; 2004a, p. 417):

$$\boldsymbol{r_i = 1 - \frac{d_o}{d_c}} \qquad (\ 6\ )$$

Most of the chance-adjusted indices share Equations 3 through 6 as they are. Benini's β (1901) and Perreault and Leigh's I_r (1989) modify the two equations, which we will discuss later.

The marble drawing scenario was implicit in Guttman's ρ (1946), Bennett et al's S (1954), Scott's π (1955) and all other chance-adjusted indices that followed. Goodman and Kruskal (1954) discussed flipping a coin, and Krippendorff (1980) discussed throwing dice, making the scenario explicit. Zhao (2011a, b) rephrased it as drawing marble to allow more accurate analysis of various indices. This *Guttman-Goodman Scenario* has been widely accepted because it was told as hypothetical stories. Few believe that coders regularly maximize chance coding in actual research. Yet few realize that, by applying Equations 3 through 6, which are key components of S and all other chance-adjusted indices, we are treating maximum randomness as real occurrences. Riffe, Lacy, & Fico (2005, p.151) did realize this, pointing out "that agreement can take place by chance does not mean it does…All agreements could easily be the result of a well-developed protocol." Grove and colleagues (1981, p. 411) had the same view: "chance agreement means the agreement would be observed if two raters assigned diagnoses to cases at random. Now this is not what diagnosticians do. They assign the easy cases, or 'textbook' cases, to diagnoses with little or no error, they may guess or diagnose randomly on the others."

If we accept this *Grove-Riffe Scenario*, we may argue that Equations 3 and 6 are inappropriate, as they are based on a behavior that should never happen and probably never did. Even if deliberate and systematic random coding does happen, the data should be thrown out and no reliability should be calculated. Deliberate random coding would be a type

of cheating. A simpler cheating would be that two coders always agree with each other, without looking at any cases, throwing any dice or drawing any marbles. They would have gotten 100% agreement. The fabricated agreements cannot and need not be removed from the data. The data should be thrown away, not analyzed.

So we need to lay bare the assumptions behind Equations 3~6, which are shared by all chance-adjusted indices reviewed in this article:

Basic Assumption 2 : Maximum random.

By removing chance agreement using Equation 3 or 6, these reliability indices assume that deliberate and systematic chance coding is not hypothetical, but real — no empirical research should "remove" or "correct for" anything that's not real.

Basic Assumption 3 : Limited honesty.

By estimating reliability using Equation 3 or 6, theses indices assume that honest coding is confined to a portion of the cases defined and confined by random chance.

Assumption 4 : Specified random.

There is an infinite number of ways to be random. Coders may flip a fair coin, throw a biased die, or draw marbles of various numbers of various colors without replacement. Each method produces a different estimate of chance agreements. Because maximum randomness is hypothetical, there is no empirical justification to pick one method over another. Each index picks one way, analogous to a man picking a favorite tie from a large selection. Scott's π assumes drawing from a shared urn with replacement. Cohen's κ assumes drawing from separate urns with replacement. Krippendorff's α assumes drawing from a shared urn without replacement. And so on. Each index treats its way as the only way of being random.

This assumption is not as fundamental as the previous ones. We will not attach the word "basic" to such assumptions so as to draw more attention to the more important ones.

These assumptions entail that the chance-adjusted indices operate under a Guttman-Goodman Scenario, yet each index has been recommended for typical coding, which follows a Grove-Riffe Scenario. The mismatch between the assumption and the reality creates paradoxes:

Paradox 2 : Nothing but chance.

Equation 4, which says $1=a_c+d_c$, represents a critical assumption in all chance-adjusted reliability indices reviewed in this article: chance coding, which includes chance agreements (a_c) and chance disagreements (d_c), covers 100% of the cases coded.

We found this paradoxical because we believed, under the Grove-Riffe Scenario, coders code objectively at least sometimes, *before and beyond* random chance. Assumptions 2 & 3, under the Guttman-Goodman Scenario, stipulate that coders maximize random coding, and code honestly only when marbles' colors mismatch. "Nothing but chance in the first stage" is an operating boundary for these indices, beyond which paradoxes arise. If coder behavior follows the Grove-Riffe rather than Guttman-Goodman Scenario, Equation 4 is incorrect, and therefore these indices are all incorrect.

Paradox 3 : Apples compared with oranges.

In Equation 3, the numerator represents "honest agreements," while the denominator represents "chance disagreements." The division compares the numerator as a part with the denominator as the whole to produce a percentage figure. But why compare *honest agreements* with *chance disagreements*? Are we comparing apples with oranges? Why not compare some apples with all apples, e.g. *honest agreements* with *all coding*? We found this paradoxical because we did not realize chance disagreement *is* honest coding under Assumptions 2 & 3 — coders code honestly when and *only* when marbles disagree.

Under the Grove-Riffe Scenairo, all coding can be honest, not just those confined to chance disagreement (Riffe et al., 2005). We should replace *maximum-randomness* and *limited-honesty* assumptions with *variable-randomness* and *complete-honesty* assumptions.

Paradox 4 : Humans are subgroup of men.

When we mathematically divide men by humans, we are asking "what percent of humans are men?" assuming men are a subgroup of humans. When we divide d_o by d_c (Equation 6), we are asking "what percent of *chance* disagreements is *observed* disagreements?" assuming observed disagreements are a subgroup of chance disagreements. But should not chance disagreements and honest disagreements be two subgroups of observed disagreements? If so, we should divide chance disagreements by observed disagreements, not vice versa. Dividing d_o by d_c is analogous to saying "humans are a subgroup of men."

Paradox 5 : Pandas are a subgroup of men.

Equations 3~6 imply a_o-a_c+d_o=d_c, which implies that *honest agreements* (a_o-a_c) and *observed disagreements* (d_o) are two subgroups of *chance disagreements* (d_c), which is analogous to saying that "pandas and humans are two subgroups of men."

This appears paradoxical because we thought, under a Grove-Riffe Scenario, chance disagreement is a subgroup of observed disagreement. Nevertheless, under the Guttman-Goodman Scenario and especially Assumption 2, coders disagree (observed disagreement) when and *only* when marbles disagree (chance disagreement). Therefore observed disagreement should be a subgroup of chance disagreement.

Unfortunately, the major chance-adjusted indices all share Equations 3, 4, and 6 under the Guttman-Goodman Scenario, which we will call *maximum-randomness* equations. No index has been built under the Grove-Riffe Scenario.

V. Category-Based Indices

Accurately estimating chance agreement may be as important as properly removing it (Equations 3, 4, and 6). How to estimate chance agreement is where major indices differ.

Guttman (1946), a pioneer in social psychology and social science methodology, calculated chance agreement (a_c) as the inverse of the number of categories (K) available to the coders:

$$\boldsymbol{a_c = \frac{1}{K}} \qquad (\ 7\ )$$

Equation 7 assumes maximum randomness just as Equations 3 and 6 do. But this is a particular type of randomness: drawing randomly from marbles equally distributed among K colors, which correspond to K categories, each coder has $1/K$ probability of choosing one particular category; two coders have $(1/K)*(1/K)$ probability of agreeing on the category. Multiplying this product by K categories, we see a probability of $(1/K)*(1/K)*K=1/K$ that the two coders would agree by chance. This equation and the rationale are the foundation of the category-based indices discussed below.

V.1. Bennett et al's *S* and Six Equivalents

Bennett et al (1954) recommended a reliability index, S:

$$\boldsymbol{S = \frac{K}{K-1}\ (a_o - \frac{1}{K})} \qquad (\ 8\ )$$

Equation 8 can be derived by inserting the right side of Equation 7 into Equation 3. In other words, S implies directly Equations 1, 3, and 7, and indirectly Equations 4 through 6. So, S assumes maximum randomness not only when chance agreement is removed (Equations 3 and 6), but also when chance agreement is calculated (Equations 7 and 8).

By removing chance agreement, S aims to avoid Assumption 1 and Paradox 1. Nevertheless, by using Equations 3 and 6 to execute the removal, S adopts maximum-randomness and limited-honesty assumptions. Adding Equations 7 and 8, S assumes the following *Bennett Scenario* for two coders:

1. The coders place K sets of marbles into an urn, where K equals the number of coding categories. Each set has an equal number of marbles and has its own color.

The coders agree on which color represents which category. Again, in this article "marble" refers to any physical or virtual element of equal probability, and "urn" refers to any real or conceptual collection of the elements.

2. They take a target to be coded. Here *target* is anything under coding, such as an advertisement, a news story, a patient, etc.
3. One coder draws a marble randomly from the urn, notes the marble's color, and puts it back. The other coder does the same.
4. If both draw the same color, each reports that the target belongs to the corresponding category according to the predetermined color-category pairings, without looking at the target. Only if they draw different colors would they code objectively, at which point they may honestly agree or disagree, and report accordingly.
5. The coders repeat Step 2 and the subsequent steps, and end the coding session when they have thus "coded" all targets.

Note that the Bennett Scenario is a special case of the broader Guttman-Goodman Scenario discussed earlier. The Guttman Scenario reveals several additional assumptions of *S*:

Basic Assumption 5 : Categories equal marble colors.

There is an infinite number of ways to be random. The coders could use any number of urns, any number of marbles, any number of marble colors, and choose any distribution pattern of the colors; they could draw with or without replacement; and they could decide on different color-category matching. Each of these parameters may affect chance agreement. To estimate *the* chance agreement, *S* made several assumptions, one of which is that coders set the number of marble colors equal to the number of categories in the coding scheme.

Assumption 6 : Equal number per color.

Coders put in the urn an equal number of marbles per color.

Assumption 7 : Drawing with replacement.

While maximizing random coding, coders draw marbles with replacement. All other chance-adjusted indices assume the same, except Krippendorff's α (1970, 1980), which assumes drawing without replacement.

Assumption 8 : Color mismatch equals honesty.

Coders code honestly when marbles' colors mismatch. Most of the chance-adjusted indices assume the same, except Gwet's AC_1 (2008) and Goodman and Kruskal's λ_r (1954), which we will discuss later.

Basic Assumption 9 : Categories reduce chance agreements.

Equation 7 assumes that category is the *only* parameter affecting chance agreement. Nothing else, including the distribution pattern of the cases coded, affects chance agreement. More categories mean less chance agreement. Two categories imply 50% chance agreement, while 10 categories imply 10% chance agreement. As categories approach infinity, chance agreement approaches 0%. Accordingly, we say the indices sharing Equation 7 are *category based*.

Bennett et al. (1954) compared S with a_o and a_c. They appeared to be aware that their chance agreements (Equation 7) were only hypothetical, so they used S only as convenient references complementing other information, including a_o, a_c, and K. Between the lines of Bennett et al (1954), we do not sense that S is the only or better indicator of reliability, but instead one more piece of information added to the overall picture. This nuanced understanding is not often seen in the writings of some later authors of various indices of inter-coder reliability.

Since 1954, S has been independently reinvented at least six times. Some of the reinventions have minor variations or more restricted applications. They are usually based on different reasoning and always bear different labels: Guilford's G (Guilford, 1961; Holley and Guilford, 1964), Maxwell's RE (1977), Jason and Vegelius' C (1979), Brennan and

Prediger's k_n (1981), Byrt, Bishop, and Carlin's *PABAK* (1993) and Potter and Levine-Donnerstein's *redefined Pi* (1999).

V.2. Guttman's ρ

About eight years before Bennett et al. (1954), Guttman (1946) proposed the same Equation 8 and implied the same Equation 7. But Guttman calculates a_o in a unique way:

$$\boldsymbol{a_o = \frac{1}{2}\left(\frac{N_{l1}}{N} + \frac{N_{l2}}{N}\right)} \qquad (\ 9\ )$$

N_{l1} and N_{l2} are, respectively, the mode frequency reported by each coder. Suppose on a binary scale Coder 1 reports 85 cases in Category 1 and 15 cases in Category 2, while Coder 2 reports 55 cases in Category 1 and 45 cases in Category 2, N_{l1}=85 and N_{l2}=55. When the right sides of Equations 7 and 9 replace respectively a_c and a_o in Equation 3, R_i is Guttman's ρ. By contrast, all other indices reviewed in this article use Equation 1 to calculate a_o. Except for the calculation of a_o, ρ is identical to *S*. So ρ shares all assumptions of *S* that have been discussed, and one paradox that will be discussed below.

Guttman's overriding concern appears to be keeping reliability scores between 0 and 1. Equation 9 achieves that objective, making Guttman's ρ one of the few chance-adjusted indices that never fall below zero. A side effect is that Guttman's a_o only crudely approximates percent agreement, leading to the following assumption:

Assumption 10 : Percent agreement needs to be approximated but not calculated.

Mode is not percent agreement. But the two are correlated. Generally, when distributions are skewed at the same direction, e.g., both coders report 90% positive, the more skewed is the distribution, the closer Guttman's a_o is toward percent agreement; when distributions are skewed at the opposite directions, the more skewed is the distribution, the farther away Guttman's a_o is from percent agreement. At one extreme, if both coders report that 100% cases fall into the same category, percent agreement and Guttman's a_o are both 1.

At the other extreme, when one coder reports 100% positive and another 100% negative, percent agreement is 0% while Guttman's a_o is 1. If both coders report 50 & 50% distributions, Guttman's a_o is 0.5 while percent agreement can be anywhere between 0 and 100%. As distributions are far more likely to skew at the same direction than opposite directions in actual coding, Guttman's a_o may be seen as a crude approximation of percent agreement.

But it is so crude that we hesitate to call Guttman's a_o an estimation of agreement. This may look more detrimental today as we now define reliability in terms of agreement. So we are not surprised that ρ has rarely been used. Bennett et al (1954) copied Equation 7 entirely from Guttman (1946) without mentioning ρ, and introduced *S* by changing only one thing, the calculation of a_o. Scott (1955) cited *S* but not ρ while developing π. And we know it was π that served as an inspiration for Cohen's κ (1960) and Krippendorff's α (1970).

We also would not recommend ρ, as ρ has all the defects but not all the benefits of *S*. Guttman (1946) was however the first we know to introduce Equation 7, which implies Equations 3~6 that contain the basic concepts and premises for reliability calculation in the past six decades. Today, when researchers calculate chance-adjusted reliability, few calculate ρ, yet almost all use Equations 3~6, thereby adopt the assumptions behind.

V.3. Perreault and Leigh's I_r

Hayes and Krippendorff (2007, p. 80) and Krippendorff (2004b, p. 417) considered Perreault and Leigh's I_r (1989) a simple modification of *S*. The modification was to take the square root of *S* when *S* is zero or above, otherwise define reliability as zero:

$$\boldsymbol{I_r = \sqrt{S} \quad (S \geq 0)} \qquad (10)$$

$$I_r = 0 \quad (S < 0) \qquad (10)$$

At two key spots, I_r=1 when S=1, and I_r=0 when S=0. Everywhere else, I_r is larger than S, with the largest difference at S=-1 and I_r=0, and the largest above-zero difference at S=0.5 and I_r≈0.71. So I_r is an elevated version of S, implying an interesting assumption:

Assumption 11 : Reliability index needs to be elevated across scale.

Perreault and Leigh's I_r (1989) assumes that a reliability index needs to be elevated numerically across the scale, after adjusting for chance using Equation 3 or 6. The only other index that makes the same assumption is Benini's β. Taking the square root of a 0~1 variable produces little change in the pattern of its behavior other than elevating it numerically. Consequently, I_r adopts all assumptions and paradoxes of S, one of which we discuss below.

V.4. A Paradox Shared by Nine Category-Based Indices

Users treat ρ, I_r, S and its six equivalents as general indicators for typical studies. As typical studies do not follow Assumptions 2~9, paradoxes arise. The shared equations (3~6) lead to shared Paradoxes 6~7, while Equation 7 leads to another classic paradox:

Paradox 6 : Empty Categories Increase Reliability.

Scott (1955) observed "given a two-category sex dimension and a P_o (our a_o) of 60 per cent, the S …would be 0.20. But a whimsical researcher might add two more categories, 'hermaphrodite' and 'indeterminate,' thereby increasing S to 0.47, though the two additional categories are not used at all." The same paradox can be replicated for Guttman's ρ with identical numbers (a_o=.6, ρ=.2 increased to ρ=.47), assuming each coder reports 60% for one gender and 40% for another. The same paradox also shows for Perreault and Leigh's I_r, if we take the square roots of 0.2 and 0.47, which would approximate 0.45 and 0.69 respectively. Now that we know the assumptions behind S, ρ and I_r, there are two ways to interpret Paradox 6:

1) The coding followed a Guttman Scenario in accordance to Assumptions 2~9. Assumption 5, which equates categories with marble colors, requires the coding in

Paradox 6 be separated into two sessions. In the first session the coders draw from two colors, while in the second they draw from four colors. With four colors, there are more chances of color mismatch, therefore more chances of honest coding, therefore higher reliability. There is no paradox if coders indeed coded this way.

2) The coding followed a Grove-Riffe Scenario in accordance with the variable-random and complete-honesty assumptions. Coders did not use any urns or marbles. Assumptions 2~9 have been violated; therefore S, ρ, or I_r should not have been calculated. Paradox 6 is not a real paradox. It is only the symptom of special-purpose indices applied beyond their boundaries.

Scott's (1955, pp.321-322) interpretation was: "The index (S) is based on the assumption that all categories in the dimension have equal probability of use $1/K$ by both coders. This is an unwarranted assumption for most behavioral and attitudinal research. Even though k categories may be available to the observers, the phenomena being coded are likely to be distributed unevenly, and in many cases will cluster heavily in only two or three of them … S would appear to be an unsatisfactory measure of coding reliability."

Scott was right to reject one assumption of S that "categories … have equal probability of use" which is implied in the categories-equal-colors and equal-number-per-color assumptions. Scott however accepted, possibly unknowingly, the more detrimental assumptions of S, namely maximum-randomness and limited-honesty. Consequently, while Scott's π eliminates one symptom of S, it causes other symptoms that are arguably more problematic, which we will discuss below.

VI. Distribution-Based Indices

The eight indices reviewed in this section all assume that distribution is the most important factor affecting chance agreement. They differ with each other in other details.

VI.1. Scott's π and Two Equivalents, *Revised K* and *BAK*

Of the chance-adjusted inter-coder reliability indices, Scott's π is second only to Cohen's κ (1960) in popularity. In Communication and Mass Media Complete (CMMC), citations for "Scott's Pi" rose from 11 in 1994 to 61 in 2009, totaling 597 for the period. It has been also recommended later under two different names, Siegel and Castellan's *Revised K* (1988) and Byrt et al's *BAK* (1993). Because they are mathematically equivalent to each other, our discussions and findings hereafter about π also apply to *Revised K* and *BAK*.

Like other major chance-adjusted indices, Scott's π shares the same chance-removing procedure (Equations 3, 4, and 6) while adopting its own chance-estimating procedure. For a binary scale, Scott (1955) estimates chance agreement (a_c) using the average of two coders' positive answers (N_p) and the average of their negative answers (N_n):

$$\boldsymbol{a_c = \left(\frac{N_p}{N}\right)\left(\frac{N_p}{N}\right) + \left(\frac{N_n}{N}\right)\left(\frac{N_n}{N}\right)} \qquad (\ 11\ )$$

Here N_p is from the two coders' (*1* and *2*) positive decisions (N_{p1} & N_{p2}):

$$\boldsymbol{N_p = \frac{N_{p1} + N_{p2}}{2}} \qquad (\ 12\ )$$

And N_n is from the coders' negative decisions (N_{n1} & N_{n2}):

$$\boldsymbol{N_n = \frac{N_{n1} + N_{n2}}{2}} \qquad (\ 13\ )$$

When the right side of Equation 11 is inserted into Equation 3, r_i is Scott's π. Like *S*, π assumes maximum randomness. Two coders draw with replacement from the same urn of *N* marbles, N_p black and N_n white. The probability of one coder getting black is N_p/N, both

getting black is $(N_p/N)^*$ (N_p/N), both getting white is $(N_n/N)^*$ (N_n/N). The probability of their agreeing through marble drawing is the sum of the two products, hence Equation 11.

Although Scott's π accepts the categories-equal-colors assumption, it rejects the equal-number-per-color assumption, allowing the number of marbles for each color to vary between 0 and *N*. Hence it succeeds in excluding category (*K*) per se as a parameter and avoids the categories-increase-reliability paradox. By sharing Equations 3, 4 and, 6, however, π shares maximum-randomness and limited-honesty assumptions. Further, π adopts average distribution as a parameter (Equation 11), hence adopts more consequential assumption:

Basic Assumption 12 : Conspired quota.

To calculate chance agreement under the maximum randomness assumption, we need to know the marble distribution. *S* assumes even distribution across all colors, making category a parameter. Scott's π rejects this assumption. So what is the distribution? No one knows, because marble drawing is only hypothetical. Even if marble drawing had happened, marble distribution can be anywhere between 0% & 100% and 100% & 0%. Scott's π assumes that average of the "observed distributions" reported by the coders is also the marble distribution. That means that π mathematically equates marble distribution with observed target distribution.

But there is no natural linkage between the two. Coders may draw from an urn of 40% & 60% marbles while coding a pile of 90% & 10% commercials. If the research is done reasonably well, its observed distribution should be related to the targeted commercials but normally unrelated to the marbles.

Under a Guttman-Goodman scenario, marble distribution must be set before drawing, which has to take place before the coding that produces the observed distributions. There is only one way marble distributions could equal observed distributions regularly and precisely— if someone sets a quota that is accurately executed. While ordinary marble drawing contains sampling errors, Equation 11 leaves no room for error, implying that π

assumes a strict quota — the two coders execute the quota so faithfully that the average distribution they report is identical to the marble distribution in the urn.

Equation 11 uses the average of two coders' observed distributions, implying that the two coders set one quota, share one urn, and work together to deliver the quota, hence "conspired quota," or "collectively strict quota."

To justify using observed distribution, it is often argued that the observed distribution is a reasonable estimate of the population distribution (Cohen, 1960, p. 40; Krippendorff, 2004b, p. 418; Scott, 1955, p. 324). This reasoning mixed two populations, *target population* under study, such as news and ads, and *marble population* in the urn, from which coders hypothetically draw. Observed distribution can be a reasonable estimate of target distribution, but normally not a legitimate estimate of marble distribution.

Equation 11 needs a marble distribution, and employs observed distribution as a surrogate. The equation does not need the distribution of the target population. The sample-population linkage does not justify Equation 11 or π, while a conspired quota does. This implies another assumption behind Scott's π, which was later also adopted by κ, α, and AC_1:

Assumption 13 : Trinity distribution.

This is a group of three assumptions. 1) Observed sample distribution equals target population distribution; 2) observed sample distribution equals marble distribution; hence 3) observed sample distribution equals marble distribution. The first assumption is consistent with probability theory assuming a probability sample. The latter two are inventions implied in π, which cannot be justified by probability theory or empirical evidences.

Gwet (2010, p. 40) commented: "Scott's π is … very sensitive to trait prevalence." This is because distribution (prevalence) is a main factor in π, even though the index is supposed to measure agreement but not prevalence. We will discuss later that distribution also affects Gwet's AC_1, although inversely.

By sharing maximum-randomness equations (3, 4, and 6), π also shares the underlying assumptions of *maximum-randomness* and *limited-honesty* (2 & 3). By adopting Equation 11, π also shares *replacement-drawing* and *mismatch-equals-honesty* assumptions (7 & 8), and three additional assumptions below:

Assumption 14 : Constrained task

A study is not to investigate how many targets are in what category, which has been pre-decided by the quotas, but to place targets into appropriate categories under the quotas.

Assumption 15 : Predetermined distribution.

Executing a quota implies that distribution is determined before coding. Therefore the observed distribution must remain unchanged within a study when the coders improve their work, as their "work" is not to assess distribution between categories.

[INSERT TABLE 3 Scott's Chance Agreement (a_c) as a Function of Two Distributions* HERE]

Assumption 16 : Quota & distribution affect chance agreements.

Chance agreement a_c is a function of marble distributions, which is predetermined by the quotas. This assumption is implied in the maximum-randomness and conspired-quota assumptions. If all marbles in the urn are of one color, the coders have no chance to code honestly; they have to agree all the time, by chance. If the marbles are 50% black and 50% white, the coders have a 50% chance of agreeing randomly and 50% chance coding honestly.

As quota determines both observed distribution and chance agreement, the latter two also correlate with each other. Table 3 displays Scott's chance agreement as a function of observed distributions. According to Equations 3 and 6, chance agreement a_c is a bar that percent agreement must pass to produce a positive index, and pass by margins to produce a good-looking index. Higher a_c means a higher bar and lower looking reliability. An important pattern is that the more skewed is the observed distribution, the higher the bar, the lower the π.

[INSERT TABLE 4 Assumptions of 22 Inter-coder Reliability Indices HERE]

These assumptions, as summarized in Table 4, portray the following *Scott Scenario* for a binary scale, which is another case of the broader Guttman-Goodman Scenario:

1. Two coders set a quota for the black and white marbles, and fill the urn accordingly. They also agree on which color represents positive and which negative. We will assume black-positive & white-negative pairings hereafter.
2. They take a target to be coded.
3. One coder draws a marble randomly from the urn, notes the marble's color, and puts it back. The other coder does the same.
4. If both draw black, each reports positive; if both draw white, each reports negative; in either case they do not look at the target being coded. Only if one draws a black and the other draws a white would they code objectively, at which point they may honestly agree or disagree, and report accordingly.
5. The two coders calculate the average of positive cases and the average of negative cases they have reported. If one average reaches the quota, they stop drawing, report the remaining targets according to the quota, then end the coding session. If neither average reaches the quota, they repeat Step 2 and the subsequent steps.

The *Scott Assumptions* (2~4,7,8,12~16), as illustrated in the *Scott Scenario*, constitute the boundaries beyond which Scott's π should normally not be used. Scott's π, however, has been used as a general indicator of reliability for typical coding. As typical coding is closer to a Grove-Riffe Scenario than a Scott Scenario, paradoxes and abnormalities arise, which we will discuss after analyzing two closely related indices, κ and α.

VI.2. Cohen's κ and an Equivalent, A_2

Cohen's κ (1960) has been the most often used chance-adjusted index of reliability. In Social Sciences Citation Index (SSCI), Cohen (1960) was cited 203 times in 1994 and 306 times in 2010, totaling 3,624 during the period. Rogot and Goldberg (1966) proposed A_2, which Fleiss (1975) pointed out is equivalent to κ. So all our discussion about κ also applies to A_2.

Cohen (1960, pp. 40-41) disagreed with Scott's estimation of chance agreement, a_c arguing: "(Scott) assumes…the distribution … is … equal for the judges … (which) may be questioned," because (p. 38) "the judges operate independently." So he replaced two coders' *average* positive (N_p) and negative answers (N_n) in Equation 11 with each coder's (*1* and *2*) *individual* positive (N_{p1} & N_{p2}) and negative (N_{n1} & N_{n2}) answers:

$$\boldsymbol{a_c = \left(\frac{N_{p1}}{N}\right)\left(\frac{N_{p2}}{N}\right) + \left(\frac{N_{n1}}{N}\right)\left(\frac{N_{n2}}{N}\right)} \qquad (\ 14\ )$$

When the right side of Equation 14 is inserted into Equation 3, r_i is Cohen's κ. Cohen (1960) agreed with Scott (1955) on one important point: "the distribution of proportions over the categories for the *population* is known." Here, like Scott (1955), Cohen (1960) conceptually mixed the target population with the marble population, treating the two as one. He injected into κ the observed distribution as if it was the marble distribution, but justified the injection in terms of the target distribution. In other words, κ shares the *trinity distribution* assumption, making distribution a major parameter like π does. Consequently, κ adopts a quota assumption similar to π's, and behaves quite similarly to π. By adopting maximum-randomness equations (3, 4, and 6), κ also shares maximum-randomness and limited-honesty assumptions with *S* and π. The only difference among them is how to estimate chance agreement a_c, and the only difference between π and κ is how to set and execute the quota. While π assumes that two coders set one quota, and work together to execute it, κ assumes differently:

Basic Assumption 17 : Individual quotas.

Cohen's κ uses observed *individual* distributions, implying that each coder sets his own quota, places marbles accordingly into his own urn, and works individually to assure that the distribution he reports meets his own quota, hence "individual quotas."

Cohen (1960, Table 1) adapted "agreement matrix of proportions" from the χ^2 procedure to justify and explain κ. While χ^2 multiplies margins of an association matrix to calculate the probabilities expected under the no-association hypothesis, Cohen's κ (1960, p. 38) multiplies margins of an agreement matrix to calculate a_c.

There is, however, a crucial difference between the two matrices, as we alluded to in Section I. The variables of an association matrix, such as race and locale, may be independent of each other, while the variables of an agreement matrix are coders' observations of the *same* targets, and hence normally cannot be independent of each other. By multiplying the distributions of race and locale, χ^2 assumes that each is independent. Likewise, by multiplying individual distributions of the coder observations, κ assumes that each is independent. If each is independent, they cannot come from objective observations of the same targets. We have to find another source to justify the presumed independence, which we found in two independently predetermined quotas. This analysis does not apply to π, α or AC_1, each of which uses average rather than individual distributions, hence assumes a conspired rather than individual quota.

Table 5 for Cohen's a_c is to be compared with Table 3 for Scott's a_c. The comparison reveals that Cohen's a_c is usually lower and never higher than Scott's a_c, which means that κ is usually higher and never lower than π. The most striking difference occurs when the two observed distributions are skewed in the opposite directions, where Cohen's a_c approaches 0%, while Scott's a_c approaches 50%.

[INSERT TABLE 5 Cohen's Chance Agreement (a_c) as a Function of Two Distributions* HERE]

Feinstein and Cicchetti (1990, p. 548) observed "The reasoning (of κ) makes the assumption that each observer has a relatively fixed probability of making positive or negative responses. The assumption does not seem appropriate, however for most clinical observers. If unbiased, the observers will usually respond to whatever is presented in each particular instance of challenge." "Fixed probability" is quota. Feinstein and Cicchetti (1990) recognized κ's individual quota assumption more than 20 years ago without naming it so. As discussed earlier it is a strict quota, not "relative."

The *Cohen Assumptions* (2~4, 7, 8, 14~17), which are also summarized in Table 4, portray the following *Cohen Scenario*, which is another special case of the Guttman-Goodman Scenario:

1. Each coder sets a quota for the black and white marbles, and fills his or her urn accordingly.
2. They take a target to be coded.
3. One coder draws a marble randomly from his urn, notes the marble's color, and puts it back. The other coder does the same from her urn.
4. If both draw black, each reports positive; if both draw white, each reports negative; in either case they do not look at the target being coded. Only if one draws a black and another draws a white would they code objectively, at which point they may honestly agree or disagree, and report accordingly.
5. Each coder calculates the positive and negative cases that he or she has reported. If either reaches the quota, he or she stops drawing, reports the remaining targets according to the quota, then ends the coding. If neither reaches the quota, he or she repeats Step 2 and the subsequent steps.

If a study conforms to the Cohen Scenario and Cohen Assumptions, κ would be an appropriate index of inter-coder reliability, otherwise κ would be inappropriate. When κ is

applied in violation of the Scenario and the assumptions, paradoxes arise, which κ shares with π and Krippendorff's α. We will discuss these paradoxes after analyzing α.

VI.3. Krippendorff's α

Krippendorff's α (1970, 1980) may not be as often cited as Scott's π or Cohen's κ. But it is among the most often recommended (Hayes & Krippendorff, 2007; Krippendorff, 2004b). Like Scott (1955) and Cohen (1960), Krippendorff (1980) also adopted Equations 3, 4, and 6. But Krippendorff believed that Cohen made a mistake by using individual distributions, and Scott made a mistake by assuming marble drawing with replacement, which fails to correct for sample size (cf. Krippendorff, 2004b). So Krippendorff's estimation for chance agreement retains Scott's average distributions but assumes no replacement:

$$\boldsymbol{a_c = \left(\frac{2N_p}{2N}\right)\left(\frac{2N_p - 1}{2N - 1}\right) + \left(\frac{2N_n}{2N}\right)\left(\frac{2N_n - 1}{2N - 1}\right)} \qquad (\ 15\ )$$

In Equation 11, Scott gave the first and second drawing the same probability, assuming replacement. In Equation 15, Krippendorff subtracted one for the second drawing, assuming no replacement. With two coders, this is the only mathematical difference between α and π, which has important consequences. When the sample gets larger, the relative impact of subtracting one gets smaller, Krippendorff's a_c approaches Scott's a_c, and α approaches π. This can be seen by comparing Table 6 with Table 3. When the sample is smaller than 50, however, Krippendorff's a_c can be noticeably smaller than Scott's. Table 7 shows Krippendorff's a_c as a function of target sample.

[INSERT TABLE 6 Krippendorff's Chance Agreement (a_c) as a Function of Two Distributions (N=100)* AND
TABLE 7 Krippendorff's Chance Agreement Rate (a_c) as a Function of Coded Targets (N) and Average Distribution (N_p/N) HERE]

When the right side of Equation 15 is inserted into Equation 3, r_i is Krippendorff's α.

By adopting the *maximum-randomness equations* (3, 4, and 6), Krippendorff's α adopts the maximum-randomness and limited-honesty assumptions (2 & 3) and other related assumptions summarized in Table 4. By retaining average distribution (Equation 15), α also adopts Scott's assumptions of conspired quota (12) and trinity distributions (13). To reject the replacement assumption (7), however, α adds several unique assumptions.

Basic Assumption 18 : Drawing without replacement.

All other chance-adjusted indices assume drawing with replacement. Krippendorff's α (1970, 1980) is the only one that assumes no replacement, which implies other unique assumptions explained below.

Assumption 19 : Trinity size.

When drawing without replacement, the size of the marble population, N_m, becomes important. Assuming half black and half white, if two coders draw from an urn containing only two marbles (N_m=2), the probability of getting the same color is zero; if N_m rises to four, the probability rises to nearly 0.167; if N_m rises further, the probability rises further; if N_m approaches infinity, the probability approaches 0.5. We need N_m to calculate Krippendorff's a_c and α. But N_m is usually not known. Under a Grove-Riffe Scenario, coders don't draw marbles to determine which cases to be coded randomly or honestly. Even if they do, N_m could be anything above zero. Krippendorff's α assumes each coder puts one marble in the urn for each target; so, with two coders, N_m is twice the target sample, N:

$$\boldsymbol{N_m = 2N} \qquad (16)$$

Krippendorff's α also assumes all marbles in the urn are drawn, so marble population equals marble sample. Therefore a *trinity-size* assumption: *marble sample* and *marble population* equal each other, and each doubles the *target sample*.

Krippendorff (1970, 1980, 2004a) argues that the non-replacement assumption "corrects for" sample sizes. But which sample -- target or marble? Krippendorff's non-

replacement argument would make sense if he means targets, that is, coders do not put every news story or advertisement back for recoding. Krippendorff's calculation in Equation 15 would make sense if he means marbles, that is, if coders indeed draw marbles without replacement, the subtraction by one would be necessary. But normally the argument does not justify the calculation because normally the targets and marbles are not linked. Coders may code targets with no replacement while drawing marbles with replacement; under a Grove-Riffe Scenario, coders code targets without first drawing marbles. To justify the calculation, α needs a special link between marble size and target size. Trinity-size assumption provides that link, by requiring that coders set the number of marbles according to the size of the target sample.

Also, mathematically Equation 15 needs marble *population*, not target *sample* that the equation actually uses, or marble (die) *sample* that Krippendorff could be referring to. The trinity-size assumption also closes this gap, by making the three essentially one.

The trinity-distribution assumption (13) also links marbles to targets. But the trinity-distribution assumption is shared by α, π, κ, and AC_1, while the trinity-size assumption is unique to α. AC_1, π or κ makes no assumption about the size of a population or sample, of marbles or targets, as their replacement assumption makes size irrelevant.

Assumption 20 : Predetermined target size.

Krippendorff's α assumes that the sizes of marble population, marble sample, and target sample are decided before a study and remain unchanged within the study. To test and improve their protocol, content researchers sometimes expand target samples in the middle of a study. For example, a researcher may test her protocol on a sample of 20 targets, calculate reliability, and then apply the protocol to 80 additional targets and calculate the reliability for the 100 targets combined. Krippendorff's α assumes such adjustment of sample size can never happen within a study. Instead, α assumes the coders treat the 20 cases and the 100 cases as two separate studies, meaning (a) the coders draw from 40 marbles to code the 20

cases, and (b) the coders draw from 200 marbles to code the 100 cases, including re-coding the 20 cases already coded. When α is applied to situations where coders expand their sample without drawing marbles, abnormalities arise, which we will show below.

Other indices like *S,* π, κ and AC_1, all assume replacement, so they do not assume a fixed N_m or N within a study. If two coders draw from an equal number of black and white marbles with replacement, the probability of getting the same color is 50% regardless of N_m or N.

Assumption 21 : Larger samples increase chance agreements.

It is often said that α is superior to π and all other indices in part because "α … is corrected for small sample sizes" (Krippendorff, 2004a, p. 250). This is appealing, as we are accustomed to statistical indicators that reward larger samples. For example, everything else being equal, statistical significance is more likely with a larger sample of respondents, and Cronbach's alpha is larger with a larger sample of measures.

Krippendorff's "correction," however, does the opposite. It systematically rewards smaller samples. As shown in Table 7, everything else being equal, a smaller sample produces a smaller a_c, hence a higher α. This is a consequence of the trinity-size and non-replacement assumptions (18, 19): a smaller target sample means a smaller marble population, which means lower a_c and higher α.

In typical studies under a Grove-Riffe Scenario, such a correction is not needed for marble sample or target sample, because marbles were actually not drawn to determine when to code randomly or honestly, and targets were not drawn for deliberate random coding. As the correction is not needed, α is not needed. When α is applied in such a study, sample-size related paradoxes arise, which we will discuss shortly.

Equations 3, 4, 6, and 15 constitute Krippendorff's α for binary scale with two coders. With multiple coders and multiple categories, Krippendorff's α takes more complex forms

(Hayes & Krippendorff, 2007; Krippendorff 2004a, 2004b). While this review focuses on a binary scale with two coders, these boundaries also apply to more categories and more coders.

The thirteen *Krippendorff Assumptions* (2~4, 8, 12~16, 18~21), again summarized in Table 4, portray the following *Krippendorff Scenario*, which is another case of the broader Guttman-Goodman Scenario:

1. Two coders set a quota for the black and white marbles. They also set the number of marbles to be twice the target sample. They fill the urn accordingly.
2. They take a target to be coded.
3. One coder draws a marble randomly from the urn, notes marble's color, and puts it aside without placing it back into the urn. The other coder does the same from the same urn.
4. If both draw black, each reports positive; if both draw white, each reports negative; in either case they do not look at the target being coded. Only if one draws a black and the other draws a white would they code the target objectively, at which point they may honestly agree or disagree, and report accordingly.
5. The two coders calculate the average of positive cases and the average of negative cases they've reported. If one of the two averages reaches the predetermined quota, they report the remaining targets according to the quota, and end the coding session. If neither average reaches the quota, they repeat Step 2 and the subsequent steps.

When α is applied beyond the boundaries defined by the assumptions and illustrated in the Scenario, it creates abnormalities and paradoxes. Here we discuss three that are unique for α:

Paradox 7 : Punishing larger sample and replicability.

Suppose two coders code 40 online news stories to see if they were commentaries in disguise. With N=40, they generate 20 positive agreements, 10 negative agreements, and 10

disagreements. This means a 62.5% & 37.5% distribution, a_o=75%, and Krippendorff's α=0.4733, which may appear improvable given the relatively small *N*. Suppose the researcher expands the target sample 10 fold by coding 360 more stories. For the 400 targets combined, the coders produce 200 positive agreements, 100 negative agreements, and 100 disagreements, replicating the 62.5% & 37.5% distribution and 75% a_o. The only difference is Krippendorff's α, which is decreased to 0.4673. It's not a huge decrease. But for 10 times as much work of the same quality and the same agreement rate, we would not have expected *any* decrease.

This unexpected phenomenon will appear more dramatic if *N* is smaller. Suppose the coders take four stories out of their original 40, including two positive agreements, one negative agreement, and one disagreement. With the same distribution and agreement rate but a dramatically smaller *N*, one would not expect any improvement in the reliability score. Instead, Krippendorff's α improves to 0.5333, which is a 12.68% increase for one tenth of the work of the same quality. While calculating reliability on four items is not a good practice, α rewards it with a higher reliability score.

When the decrease in α caused by an increased *N* is large enough, it could offset or even overcome an increase in a_o, producing a "larger sample, higher agreement, but lower α." Suppose the researcher expands *N* from 4 to 1,000, producing 501 positive agreements, 251 negative agreements, and 248 disagreements. This would produce a much larger *N* and a slightly improved a_o (from 75.0% to 75.2%) while the distribution remains unchanged. Yet α still decreases, from 0.5333 to 0.4712. This phenomenon is limited to situations when the increase in a_o is small relative to the larger increase in sample size, and the resulted drop in α is usually not large. It however adds another dimension to the paradox.

Reliability is often understood as replicability. But in these cases α punishes replicability. The same phenomena do not exist for π, κ or other major indices, none of

which is affected by *N*. In the three examples of *N*=4, 40, or 400, the other indices all remain the same. They report larger reliability in the example of *N*=1,000, because a_o is higher.

Two examples from Krippendorff (1980, pp. 133-135; 2007, pp. 2-3) can be adapted to illustrate the same phenomenon. Both have *N*=10, distribution 70% & 30%, a_o=0.6 and α=0.09524. If *N* increases to 100 while distribution and a_o remain the same, one might expect α to improve or at least remain the same. Instead, α drops to 0.05238.

We found this abnormal because we assumed normal studies in which researchers pretest 10 cases, calculate reliability, add 90, and test reliability again, all in full honesty. In this Grove-Riffe Scenario, more of the same quality deserves no punishments, and more of the better quality deserves rewards. Krippendorff's α, however, assumes that coders maximize random coding by drawing marbles *without replacement*. They don't simply "add cases." They instead draw from 10 marbles each to code the 10 messages, then draw from 100 marbles each to code the 100 messages, including redrawing to recode the 10. More coding means more marbles, which mean more chance agreements, which have to be punished.

These phenomena are not isolated. They are a part of the paradoxical pattern of Krippendorff's a_c. Table 7 shows that Krippendorff's a_c is positively correlated with *N*: larger *N* leads to higher a_c, at any level of distribution! Higher a_c means lower reliability. Under a Grove-Riffe Scenario, larger *N* means more cases coded hence higher replicability, which Krippendorff's α punishes systematically. When we see a larger *N,* we see more honest coding, for which the bar should *not* be raised. But when α sees a larger *N*, it sees more marbles drawn, hence more chance agreements, hence a raised bar.

Paradox 8 : Purely random guessing can be somewhat reliable.

Suppose two coders coded four television stories to see if they contain subliminal advertisements. The task was so difficult that the coders end up guessing randomly. As probability theory would predict, each of them reported two positives, two negatives, with a

50% agreement rate (a_o=.5, N=4), as if they had flipped four coins each. As one might expect, most of the reliability indicators, including Scott's π and Cohen's κ, are exactly 0.00. Krippendorff's α, however, stands out at 0.125. It's a tiny sample and it is not a spectacular α. But why is it not zero?

In Krippendorff's α, only "drawing with quota and without replacement" qualifies as random (Assumptions 4, 12 and 18). Random guessing or flipping coins does not qualify, because neither allows quota and both have replacement. Guessing with coins generated more agreement than drawing with quota without replacement. We attribute the difference to "another kind of randomness," and do not believe it deserves a higher reliability score. Krippendorff's α attributes the difference to honest coding, and rewards it with a higher α.

Paradox 9 : Random guessing may be more reliable than honest coding.

Extending the above example, this α=0.125, from a_o=.5, N=4, from totally random guessing, is better than α=0.095 from two Krippendorff examples, each having a_o=0.6, N=10, from totally honest coding (Krippendorff, 1980, pp. 133-135; 2007, pp. 2-3). So, according to α, more agreement from an objective process can be less reliable than less agreement from a random process. There are two reasons for this phenomenon. First, α assumes some of our random guessing is honest coding. Second, Krippendorff's examples have a larger N (10) than our coin flipping (4), and α assumes that larger N generates more chance agreements, which have to be "corrected for," meaning punished.

Paradoxes 7~9 offer some evidences that Krippendorff's α should not be used beyond the highly restrictive boundaries defined by the Krippendorff Scenario and the Krippendorff Assumptions.

VI.4. Paradoxes and Abnormalities Shared by π, κ, α and Equivalents

Paradoxes are unexpected qualitative features of an index that seem to defy logic or intuition. There are also unexpected numerical outcomes of an index when it is used in

typical research, which we will call abnormalities. As paradoxes and abnormalities are closely linked, we will number them consecutively. The purpose of the discussion is to further illustrate the assumptions.

In addition to its unique sample-size-related paradoxes, α shares paradoxes 2~5 with all other chance-adjusted indices. Further, π, κ and α also share a few of their own paradoxes and abnormalities, which we discuss below.

We will first discuss three abnormalities that have been better known for κ. We will show that π and α suffer from the same abnormalities. We will then discuss other abnormalities and paradoxes not yet in the literature. As noted earlier, all findings about π also apply to Siegel and Castellan's *Revised K* (1988) and Byrt's et al's *BAK* (1993), and findings about κ also apply to Rogot and Goldberg's A_2.

Abnormality 10 : High agreement, low reliability.

Feinstein and Cicchetti (1990) called this a paradox for Cohen's κ (1960). Lombard et al. (2002) and Krippendorff (2004b, p. 426) debated over the same phenomenon for κ and π. Here is a more dramatic example. Suppose two coders code 1,000 magazine advertisements for cigarettes in the United States, to see whether the Surgeon General's warning has been inserted. Suppose each coder finds 999 "yes" and one "no," with 998 positive agreements and two disagreements, generating a 99.8% agreement rate. But π, κ and α are all below zero (-.001 or -.0005). Zero indicates a totally unreliable instrument. Given the near-perfect agreement, it's difficult to understand why the instrument is that bad.

Some authors found this paradoxical because they assumed the coders code honestly. The three indices, however, assume that all observed agreement (a_o=99.8%) is due to chance because each coder draws from 999 or 998 black marbles and one or two white marbles. The marbles show different colors only twice, which are the only opportunities for honest coding (Assumption 8). The coders disagrees both times, hence the low π, κ and α.

Abnormality 11 : Undefined reliability.

When two coders agree that the distribution of one category is 100% and another is 0%, π, κ or α are undefined. 0% & 100% and 100% & 0% are the two ends of all possible distributions, like the two ends of a ruler that define its length and scope. If a ruler is completely broken at both ends, it is probably not accurate in between.

Many found this paradoxical because we expected perfect agreement to be credited with a decent reliability score, and because we believed some agreements must be honest, no matter how skewed a distribution is. But π, κ and α assume that a 0% & 100% target distribution means that all marbles are of one color, hence there is no chance for color mismatch or honest coding, hence π, κ or α should not be calculated. In defense of the undefined π and α, Krippendorff (2004b, p.425) explained,

> Such data can be obtained by broken instruments or by coders who fell asleep or agreed in advance of the coding effort to make their task easy. … appropriate indices of reliability cannot stop at measuring agreement but must infer the reproducibility of a population of data; one cannot talk about reproducibility without evidence that it could be otherwise. When all coders use only one category, there is no variation and hence no evidence of reliability.

To those who assume coders intend to be honest, the explanation is still puzzling. Suppose 100% of the target population of magazine ads under study had the Surgeon General's warning. Suppose coders agreed that 100% of the target sample had the warning. Suppose there was no broken instrument, no falling asleep or agreeing in advance, but only honest and diligent coding, as evidenced in the perfect agreements between the coders, and between the sample and the population. Why is this not an "evidence" that reliability is good, or at least calculable?

Now that we know π, κ or α is to be used only under assumptions of strict quota, maximum randomness, and trinity distribution within the Guttman-Goodman Scenario, Krippendorff's (2004b) explanation could be sensible, if we think of his "population" as "marble population." Under strict-quota and trinity-distribution assumptions, zero variation in

the observed targets is evidence for zero variation in the marbles. Coders *are assumed* to "agree in advance" to make the marbles all one color, and to code honestly *only* when the marbles mismatch. There is no chance for color mismatch, hence no chance for honest coding, hence no "evidence that it (the observation) could be otherwise. … hence no evidence of reliability." Krippendorff's defense in effect provides support for our observation that π, κ and α assume maximum randomness, strict quota, and trinity distribution.

Abnormality 12 : Zero change in a_o causing radical drop in reliability.

These indices are supposed to measure agreement. Feinstein and Cicchetti (1990) argued that Cohen's κ should rise and fall with agreement rate, a_o. So should all other reliability indices. Kraemer (1979) pointed out that, with no change in a_o, κ changes with "base rate," which we call "distribution." Uneven distribution generates lower κ than even distribution. Grove et al. (1981) and Spitznagel and Helzer (1985) called it the "base rate problem" for κ. Feinstein and Cicchetti (1990) called it a paradox for κ. It's not as widely known that π and α can produce the same abnormality.

Here is a stronger example for all three indices. Revising Abnormality 10, suppose two coders initially agree on 998 "yes" and one "no," plus one disagreement, producing a_o=99.9%, π=.6662, and κ=.6662, α=.6663. Suppose both coders flip an erroneous negative decision, resulting in 999 agreed positives and one disagreement, and increasing the average of the positives from 99.85% to 99.95%. While a_o remains 99.9%, π, κ and α each drops from .666 to .0000 or -.0005, which covers two thirds of the distance between "perfectly reliable" and "totally unreliable."

This happens because the coders code honestly without quota, violating π, κ & α's strict quota assumption. Distributions changed as the coders improved their work, violating the predetermined-distribution assumption. The violation of the same two assumptions also causes the next four abnormalities (13-16).

Abnormality 13 : Eliminating disagreements doesn't improve reliability.

Extending the example in Abnomality 12: Suppose one coder finds his only negative finding erroneous and flips, reducing disagreements by half, and increasing agreements to 99.9%. One might expect π, κ and α to improve half way toward 1, to be around 0.5. Instead, κ and α barely move, to be 0, and π remains negative, at -.0005. Suppose the other coder also flips his only negative finding, improving agreement to 100%. One might expect π, κ and α jump to 1. Instead, none of the three can be calculated, repeating Abnormality 11.

Abnormality 14 : Tiny rise in a_o causing radical rise in reliability.

With 998 agreements on "yes," suppose one coder flips his positive decision in one of the two disagreements. Now disagreements decrease to one and agreements increase to 999. a_o improves slightly from 99.8% to 99.9%. Given what we have seen in Abnormality 13, one might expect the three indices to change little. Instead, π and κ jump from -.001 to .6662, while α jumps from -.0005 to .6663, each covering two-thirds of the distance between "totally unreliable" and "perfectly reliable."

Abnormality 15 : Rise in a_o causing radical drop in reliability.

Suppose two coders initially had two disagreements and 998 agreements, with 997 positive and one negative, producing an a_o=99.8%, π=.499, κ=.4993, and α=.4992. Suppose one coder finds all his three negative decisions erroneous, and flips each, resulted in 999 positive agreements and one disagreement. While a_o increases to 99.9%, κ and α drop drastically to 0, and π drops even more, to -.0005.

Abnormality 16 : Honest work as bad as coin flipping.

Suppose we show at normal speed 60 television segments, 50 of which contain subliminal advertisements barely recognizable. One coder finds the ads in all 60 segments, making 10 false alarms, while the other recognizes only 40, calling 10 false negatives. The 40 positive agreements and 20 disagreements produce a 66.667% a_o and an 83.333% average

distribution, which matches the target distribution. While the instrument may seem adequate, especially considering the difficult task, π= -.2, κ=0 and α= -.2.

Now suppose we ask the coders to flip coins *without looking at any television segments*. Their a_o is expectedly 50%, 16.667% lower than honest coding. Their average distribution is also expected to be around 50%, 33.333% lower than the target distribution. This, however, produces π=0, κ=0 and α=0.0083. So, honest coding that produces more accuracy and more agreement is no better or even worse than dishonest coding that produces less accuracy and agreement, according to π, κ or α.

This appeared puzzling because we assumed all of the 67% agreements were honest under a Grove-Riffe Scenario. But π and α presume the coders draw from 50 black and 10 white marbles. Without a single glance at the targets, they should generate 72% agreement, much higher than the 67% they actually report, leading to justifiable π= -.2 and α= -.2.

Under the Cohen Scenario, κ presumes one coder draws from 40 black and 20 white while the other from 60 black and no white. Without a glance at the TV, they should obtain 67% agreements, implying that they have not produced any honest agreement. So κ should be zero.

Paradox 17 : Punishing Improved Coding.

Abnormality 15 is a case of improved coding causing a drastic drop in π, κ and α, from half-way reliable (0.5) to not at all reliable (0)! Of all the symptoms of π, κ and α, this one may be among the most troublesome. Abnormality 12 is another example. After the errors are corrected, π, κ and α drop even more drastically.

Paradox 18 : Punishing agreement.

The three a_c not only move significantly, they also move to punish the good and reward the bad. Table 3 shows that, when one coder's distribution N_{p2}/N is 100%, Scott's a_c is positively linked to the other coder's distribution N_{p1}/N; an increase in N_{p1}/N brings it closer to N_{p2}/N, producing a higher agreement a_o and a higher a_c, which means a higher bar.

The same pattern exists when N_{p1}/N =100%, N_{p2}/N =0%, or N_{p1}/N =0%. The maximum agreement at the lower left and upper right corners of Table 3 makes a_c=100%, which is impossible to pass. As agreement rate decreases from either corner along any of the four sides, a_c decreases at an averaged half rate, until maximum disagreement at the upper left or lower right corner where a_c=50%, which is the lowest possible bar in Scott's π.

Tables 6 and 7 show that Krippendorff's a_c behaves almost exactly the same as Scott's a_c when the sample is large enough. Cohen's a_c behaves in the same pattern, except the paradox is twice as dramatic: as a_o decreases from either corner along any of the four sides of Table 5, a_c decreases at the same (rather than half) rate, until it reaches maximum disagreement at the upper left or lower right corner where a_c=0% (rather than 50%). Again, higher agreement brings a higher bar, and the lower agreement brings a lower bar.

While the paradoxical pattern is strongest in the four sides encompassing Tables 3, 5, and 6, it also manifests itself inside although in less dramatic rates. The three indices are advertised as general indices of reliability, which is defined as agreement. Why do they *systematically* punish agreement and reward disagreement?

We found this paradoxical because we compared across different distributions, violating the quota and predetermined-distribution assumptions. Each of the three indices would reward higher agreement, but only within a predetermined distribution decided by the quota(s). If the distribution changes, a different study including a different round of marble drawing is assumed. More skewed distribution in a different marble population produces higher chance agreement, hence less honest coding, which π, κ and α punish according to Assumption 16.

Paradox 19 : Radically and erratically moving bar.

To highlight the dramatic paradoxes and abnormalities, the above examples used extremely uneven distributions, such as 99.8% & 0.2%. More even distribution such as 60% & 40% would produce the same pattern, although less dramatic symptoms. Scott's, Cohen's

and Krippendorff's chance agreements (a_c) are all functions of distribution. Uneven distribution produces higher a_c, which is *the* bar that a_o must pass in order to produce an above-zero index. Both a_c and a_o have100% as the maximum. The closer is a_c to100%, the less room above it, the less chance for a high index. When distribution reaches 0% or 100%, a_c reaches 100%, leaving *no* chance for a_o to pass a_c. That's the technical reason π, κ and α are all undefined there.

Tables 3, 5, and 6 show how a_c changes with two distributions. Chance agreement a_c can reach as high as 100%, but it moves gradually with no gap or abrupt jump, starting from 0% (Cohen), 49.7% (Krippendorff when N=100), or 50% (Scott). This demonstrates that the undefined π, κ, and α are not isolated incidents under extreme circumstances. They are symptoms of intrinsic defects of the three supposedly general indicators. The moving bars also explain why π, κ and α change with distribution.

We found the phenomenon paradoxical because we didn't think the bar, as a part of the general indicator for typical studies, should move with distribution. But π, κ and α are not general indicators. Each is to be used only when all of its assumptions are met. Under these assumptions, especially those derived from Assumptions 15 & 16, the bar should move.

Paradox 20 : Circular logic.

The three indices are functions of coder's observation of distribution, whose quality depends on the quality of the coding instrument. But that is the very instrument that the indices evaluate. The three indices depend on an instrument's reliability to assess the instrument's reliability! We found this circular because we thought the reported distributions embedded in π, κ or α came from coders' observations. We were wrong. The distributions came from pre-determined quotas independent of the observations, according to Assumptions 12,14,15, and 17. The logic would not be circular if coders behave under a Scott, Cohen, or Krippendorff Scenario.

These paradoxes and abnormalities show that π, κ or α cannot be general indicators of reliability. They might be useful within highly restrictive boundaries defined by various assumptions and scenarios, beyond which the paradoxes and abnormalities arise.

VI.5. Benini's β

Nearly sixty years before Cohen (1960), Italian sociologist Benini (1901) designed a chance-estimating formula that is identical to Cohen's Equation 14. Benini's chance removing formula is also similar to Cohen's (Eq. 3), except it subtracts an extra$|n_{pn}-n_{np}|$from the denomenator:

$$\boldsymbol{\beta = \frac{a_o - a_c}{1 - a_c - |\, n_{pn} - n_{np}\,|}} \qquad (\ 17\ )$$

Here n_{pn} is percent of cases Coder 1 judges as positive while Coder 2 judges as negative, and n_{np} is percent of cases Coder 1 judges as negative while Coder 2 judges as positive. They are two components of between-coder disagreements. If all disagreements are strictly random, $n_{pn}=n_{np}$, hence $|n_{pn}-n_{np}|=0$. So some may see $|n_{pn}-n_{np}|$ as non-random disagreements.

The denominator of Equation 3 is a reference scale. Benini's β (Equation 17) has a shorter reference scale than κ, which means β tends to be higher than κ across a scale when κ is above zero. So Benini's β is an elevated κ in the important 0-1 range, like I_r is an elevated ρ. So β adopts all assumptions, paradoxes, and abnormalities of κ, and adopts Assumption 8 of I_r.

VI.6. Goodman and Kruskal's λ_r

Goodman and Kruskal (1954) proposed an agreement index, λ_r, based an a_c that behaves in some ways similarly to Cohen's (1960):

$$\boldsymbol{a_c = \frac{1}{2}\left(\frac{N_{l1}}{N} + \frac{N_{l2}}{N}\right)} \qquad (\ 18\ )$$

One may interpret N_{l1} and N_{l2} as, respectively, individual modal frequency reported by each coder. Suppose on a binary scale Coder 1 reports 85 cases in Category 1 and 15 cases in Category 2, while Coder 2 reports 45 cases in Category 1 and 55 cases in Category 2, N_{l1}=85, N_{l2}=55, and a_c=(.85+.55)/2=0.7. Goodman and Kruskal's λ_r shares Equations 1, 3, 4, and 6 with other chance-adjusted indices. Replacing a_c in Equation 3 with the right side of Equation 18, we have Goodman and Kruskal's λ_r.

An alternative interpretation appears equally plausible, according to Fleiss, 1975. (N_{l1}+N_{l2})/2 may be the modal average frequency reported by two coders, which in the above example would instead produce an a_c=(.85+.45)/2=.65. As Goodman and Kruskal did not provide a numerical example, we are unable to decide with certainty which interpretation they meant. The differences between the two interpretations would be analogous to the differences between κ and π, one assuming individual behaviors while the other presuming collective action. Given the limited space we will assume individual modal interpretation in the following discussion, and analyze the modal average interpretation in more details in a future study.

As N_{l1} and N_{l2} are a part of two coders' individual distributions, λ_r shares almost all assumptions and paradoxes we have discussed of Cohen's κ. Most notably, λ_r shares with κ the individual quota assumption (17). Goodman and Kruskal (1954) were the first we know to make Equation 3 explicit. Their λ_r also started the practice of sharing the chance-removing procedure while creating a unique chance-estimating formula.

Goodman and Kruskal's λ_r makes a set of unique assumptions, which we will put under one title, "modal color assumption." We analyze the assumption using κ as a reference:

Basic Assumption 22 : Coders code randomly when they draw the modal color.

While κ assumes that coders code randomly every time marbles' colors match, λ_r assumes that coders code randomly some of the time when one or both coders draw a certain color. Specifically, λ_r assumes: a) In addition to placing marbles into the urns according to

individual quotas, each coder also notes which color has the largest number of marbles, which we call "mode color," in his or her urn. b) The coders would code randomly every time both draw the modal color(s). c) The coders would code randomly half the time when one draws his or her modal color but the other does not.

Equation 18 of λ_r uses addition to estimate chance agreement, while Equation 14 of κ uses multiplication. Consequently, Goodman and Kruskal's chance agreement is equal to or larger, often much larger, than Cohen's, which can be seen by comparing Table 8 with Table 5. Further comparison of Table 8 with Tables 3 and 6 and other estimates by other indices show that Goodman and Kruskal provide the highest estimation for chance agreement, which makes λ_r the most conservative estimation among the 22 indices reviewed in this article.

[INSERT TABLE 8 Goodman and Kruskal's Chance Agreement (a_c) as a Function of Two Distributions* HERE]

VII. A Double-Based Index -- Gwet's AC_1

Gwet's (2008, 2010) theory about coder behavior differs from the stated theories behind all other indices reviewed in this article. Gwet separated difficult cases from easy cases, in a way that appears much closer to the Grove-Riffe Scenario than Guttman-Goodman Scenario. By adopting Equations 3, 4, and 6, however, Gwet's index, AC_1, adopts the maximum randomness assumption and the related paradoxes just like other chance-adjusted indices. Gwet's chance-estimating formulas are unique. While all other chance-adjusted indices use either category or distribution to estimate chance agreement, AC_1 uses both, hence "double-based." For a binary scale with two coders, Gwet's Equation 19 looks similar to Scott's Equation 11, except it switches one positive distribution rate (N_p/N) with a negative one (N_n/N):

$$\boldsymbol{a_c = \left(\frac{N_p}{N}\right)\left(\frac{N_n}{N}\right) + \left(\frac{N_p}{N}\right)\left(\frac{N_n}{N}\right)} \qquad (\ 19\ )$$

All chance-adjusted indices before Gwet assume coders code randomly when marbles match, and code honestly when marbles mismatch. Accordingly, Scott's Eq. 11 multiplies the positive rate by itself, and the negative rate by itself. In contrast, Gwet's Equation 19 multiplies the positive rate by the negative rate, implying a unique assumption: coders code randomly when marbles mismatch, and code honestly when marbles match.

The multiplication is done twice because the mismatches include black-white and white-black. A practical implication is that Gwet's coders have to agree on which color of which coder represents which category when the marbles mismatch, in a similar fashion that Scott's coders agree on which color represents which category when the marbles match. The choice of color-category pairing does not affect probability calculation.

While Scott had extended Equation 11 to three or more categories, Gwet also needed to extend Equation **19**. But Gwet could not do a simple extension like Scott had done. More categories mean more marble colors hence more mismatches, which mean more random coding under Gwet's unique assumption discussed above. A simple extension of Equation 19 would lead to intolerably high a_c and intolerably low AC_1, especially when number of categories is large. To counter the effect, Gwet re-introduced categories (K) as a main parameter:

$$\boldsymbol{a_c = \frac{1}{(K-1)} \sum_{q=1}^{K} \left(\frac{N_q}{N} * \frac{N - N_q}{N} \right)} \qquad (\ 20\ )$$

In Equation 20, the part after the summation sign (Σ) is a simple extension of Equation 19 from two to K categories. N_q/N represents percent of targets in the qth category while $(N\text{-}N_q)/N$ represents percent of other targets. With a binary scale, N_q/N and $(N\text{-}N_q)/N$ become respectively N_p/N and N_n/N in Equation 19. The part before the summation sign is at least equally important. Multiplying by 1/(K-1) effectively lowers the estimated chance agreement, but it also implies another unique assumption:

Basic Assumption 23 : Double drawing.

While other chance-adjusted indices all assume one round of marble drawing in the first stage of the two-stage coding (see Section IV), Gwet's AC_1 assumes two rounds of marble drawing from two urns during the first stage. Two coders first draw with replacement from the first urn, which has *K* minus one colors and an equal number of marbles per color. If colors differ, they go to the second stage to code honestly. If the colors match, they draw with replacement from the second urn that has *K* colors and a distribution that equals the observed target distribution. Coders go to the second stage after this second drawing, and they code honestly if the colors match, and code by chance if the colors mismatch. This implies another unique assumption:

Basic Assumption 24 : Marble mismatch or double-match equals honesty.

Gwet's AC_1 assumes that color mismatch in the first round or color matches in both rounds leads to honest coding, while color match in the first round followed by mismatch in the second round leads to chance coding.

By adopting the maximum random equations and using average distribution as a parameter in Equations 19 and 20, Gwet's AC_1 adopts all of Scott's assumptions except replacing Scott's Assumption 8, which is about color mismatch and honest coding, with Assumptions 23 & 24.

The Gwet assumptions lead to the following *Gwet Scenario,* which is another case of the broader Guttman-Goodman Scenario, for two coders and *K* categories:

1. Two coders prepare two urns.
2. They place marbles of (*K*-1) colors into the first urn. Each color has an equal number of marbles.
3. They set a quota for the marble distribution in the second urn, and fill the second urn accordingly. They also agree on which color of which coder represents which category, which we will call *color-category scheme*.

4. They take a target to be coded.
5. One coder draws a marble randomly from the first urn, notes the marble's color, and puts it back. The other coder does the same from the same urn.
6. If the two colors differ, each coder codes and reports objectively, then skips to Step 9. If the colors match, they go to the next step.
7. One coder draws a marble randomly from the second urn, notes the color, and puts it back. The other coder does the same from the same urn.
8. If the two colors differ, each reports the results according to the pre-determined color-category scheme, without looking at the target under coding. If the two colors match, each codes and reports objectively.
9. The two coders calculate the averages of the positive and negative cases they've reported. If one of the two averages reaches the predetermined quota, they stop drawing, report the remaining targets according to the quota, and end all coding. If neither average reaches the quota, they repeat Step 4 and the subsequent steps.

Which is right, one round or two rounds, color match or mismatch? If coders code as AC_1 assumes they do, two rounds and mismatch-or-double-match are right. If coders code like π, κ, or α assume they do, one round and mismatch are right. But if coders code like the Grove-Riffe Scenario assumes they do, none of them is right.

With a binary scale, K-1=1, which means all marbles in the first urn are of the same color, so the colors always match, and the coders always go to the second urn for the second drawing. So the mismatch-or-2-matches-equals-honesty assumption can be simplified as match-equals-honesty assumption, as we discovered while analyzing Equation 19 above.

[INSERT TABLE 9 Gwet's Chance Agreement (a_c) as a Function of Two Distributions* HERE]

Comparing Table 9 with Table 3, we see that, with a binary scale, Gwet's chance agreement is a mirror image of Scott's, with the "mirror" positioned at the 50% & 50%

distribution line. When each individual distribution is exactly 50% & 50%, Gwet's a_c is identical to Scott's, because here the probabilities of color match and mismatch are equal. When average distribution deviates from 50% & 50%, Scott's a_c increases while Gwet's a_c decreases at the same rate. When distribution becomes more uneven, Scott's a_c continues to increase toward 100%, while Gwet's a_c continues to decrease toward 0%. As Krippendorff's a_c and Cohen's a_c behave in the same pattern as Scott's, Gwet's a_c also behaves in opposite directions of Krippendorff's or Cohen's, as can be seen by comparing Table 9 with Table 6 or 5.

With a binary scale, Gwet's a_c assumes that color mismatch equals random coding while Scott, Cohen and Krippendorff's a_c assume the opposite, and Bennett et al's a_c is a constant at 0.5. So Gwet's a_c tends to be lower than the other four, hence Gwet's AC_1 tends to be higher than S, π, κ, and α. One extreme is when distribution is 0% or 100%, where π, κ, and α cannot be calculated because they all assume 100% chance coding and 0% honest coding, while, in contrast, AC_1 assumes 0% chance coding and 100% honest coding, producing a perfect AC_1=1.

There are a few exceptions to this general pattern. The first exception is when individual distributions are 50% & 50% where Gwet's a_c and the other four all equal 0.5, assuming a large enough sample for α. With a large enough sample, Gwet's a_c also equals Scott's and Krippendorff's when average distribution is 50% and 50%, even when individual distribution is not even. The second exception is when N is very small, leading to very low a_c by Krippendorff hence higher α than AC_1. The third exception is when two coders give highly uneven distributions at the opposite directions, which could lead to very low a_c by Cohen hence higher κ than AC_1.

When categories increase to three, Bennett et al's a_c is 1/3, while Gwet's a_c ranges from 0, when the coders report that all targets fall into one category, to (2/9+2/9+2/9)/2=1/3,

when the targets distribute evenly into three categories. So Gwet's a_c is usually smaller and never larger than Bennett et al's a_c, hence AC_1 is usually larger and never smaller than S. As categories increase further, the margins of AC_1 over S increase further. That means that AC_1 is more liberal than S and the equivalents.

Comparing AC_1 with I_r is more complicated. With even distribution and a_o=0.5, I_r may be higher than AC_1. With uneven distribution and and a_o closer to 0 or 1, AC_1 may be higher than I_r. A simulation by Guangchao Charles Feng, a doctoral student at Hong Kong Baptist University School of Communication, shows I_r is more often higher than AC_1, and the difference is statistically significant.

Low estimate of a_c means that AC_1 assumes less chance agreement and more honest coding. So even though AC_1 still assumes maximum randomness, its specific type of randomness is closer to complete honesty under a Grove-Riffe Scenario. Consequently, even though AC_1 shares most of its assumptions with π, κ and α (see Table 4), AC_1 does not generate as many or as dramatic paradoxes or abnormalities (see Table 10) when used under a Grove-Riffe Scenario.

But there are still paradoxes and abnormalities. Most notably, by reintroducing category as a major parameter, AC_1 brought back the classic paradox that Scott (1955), Cohen (1960) and Krippendorff (1970) worked hard to avoid, which is that empty categories increase reliability. In Scott's example (see Paradox 6) that originally had "male" and "female," by adding "hermaphrodite" and "indeterminant," S increases from .2 to .47, while AC_1 increases from .2 to .52. The larger increase means an even more dramatic paradox. Gwet's AC_1 also shares Paradoxes 2~5 with other chance-adjusted indices, and shares Paradoxes 19 & 20 with π, κ and α. It also suffers a couple abnormalities of its own:

Abnormality 21 : Same quality, same agreement, higher reliability.

Suppose, as a way of testing our instrument, we give two coders 100 news stories, and ask the coders to judge whether the stories contain commentary opinions. We put in 80 easy

cases, 40 of them having obvious commentaries, and other 40 obviously not. We put in 20 difficult cases that even experienced teachers can't judge with certainty. As expected, the two coders agree on 40 clearly positive cases, 40 clearly negative cases, and disagree on 20 difficult cases. Also as expected, of the twenty disagreements, each coder reports half positive and half negative. This generates an a_o=0.8 and AC_1=0.6.

Now we delete the commentaries from the 40 clearly-positive cases, so they become clearly negative. With no other changes, we give the 100 stories to the same coders to be coded again. The two coders again agree on 80 easy cases and disagree on 20 difficult cases. Of the 20, each coder again reports half positive and half negative. The only change is that all 80 easy cases are now negative. Again a_o=0.8. But AC_1 jumps from 0.6 to 0.7561.

The same coders, the same procedure, the same targets, the same quality of work, and the same agreement rate. Why the jump?

Abnormality 22 : Lower quality, less agreement, higher reliability.

Suppose, instead of switching all 40 easily positive to easily negative, we switch only 36, and switch the other four to be difficult by making the commentaries ambiguous. Now we have 76 obviously positive and 24 difficult cases. As expected, the same two coders agree on 76 and disagree on 24, and each reports half and half for the difficult 24. As the task is more difficult, the quality of the coding and the agreement rate is understandably lower, a_o=0.76. Gwet's AC_1, however, is 0.69574, higher than the original 0.6 by nearly 1/6. Why?

We found the results "abnormal" because, again, we assumed the coders code honestly under the Grove-Riffe Scenario. AC_1 assumes that the coders conspire to set quotas, place marbles into the second urn according to the quotas, and draw from it. They code randomly when marbles mismatch. In both abnormalities, target distribution moves from even to uneven, which means uneven marble distribution, less chance for color mismatches, less random agreement, lower bar, and therefore higher AC_1. The results would have seemed "normal" had coders indeed followed the Gwet Scenario.

VIII. When to Use Which Index?

Tables 4, 10 and 11 summarize our findings from various angles. A contrast emerges in Tables 4 and 10 – the long list of assumptions, paradoxes, and abnormalities for what we believed to be the sophisticated and rigorous measures, such as α, and the much shorter list, just one unreasonable assumption and one paradox, for the supposedly primitive and flawed percent agreement a_o. To avoid this one assumption and one paradox, we adopted more and stronger assumptions, which created more and stagier paradoxes and abnormalities. Are the medicines worse than the disease?

[INSERT TABLE 10 Paradoxes and Abnormalities of 22 Inter-coder Reliability Indices AND TABLE 11 What's Missing in the Map of Reliabilities? HERE]

The "medicines" cause not only more symptoms, but also more severe symptoms. Under a Grove-Riffe Scenario, the zero-chance-agreement assumption underlying a_o may hold sometimes, namely for "easy" and "textbook" cases with "well-developed protocols," while the maximum-randomness and other assumptions of the chance-adjusted indices may never hold.

Methodologists talk about chance agreement (a_o) as what *would* have happened, as a reference for comparison, but not what really happens in typical research. Following this thinking, each methodologist could have selected several hypothetical scenarios, such as flipping coins or throwing dice, drawing marbles of 60% or 90% distribution, from one or multiple urns, with or without replacement, in one, two or more rounds, and code randomly with color match or mismatch, etc. and etc. Each scenario can produce a unique chance agreement. As there is an unlimited number of ways for "random coding," we could have unlimited number of chance agreements, as reference lines for comparison with just one

index, which is percent agreement. Had we done that, we would not have assumed so many whimsical coders, and we would not have had so many paradoxes and abnormalities.

The methodologists, instead, used maximum-randomness equations (3, 4, and 6) to "remove" and "correct for" chance agreement. Each of them chose one hypothetical scenario of randomness, yet each believed his index applied to all real studies. This created a gap between theoretical understanding, which sees maximum randomness as hypothetical, and the actual computation, which treats maximum randomness as real, leading to the paradoxes, abnormalities, and confusions. We need to close this gap by developing a reliability index based on *complete honesty* and *variable randomness* assumptions under a Grove-Riffe Scenario.

Table 11 shows 18 cells under Column 1 titled "maximum random," seven of which occupied and 11 empty. Each empty cell represents an opportunity to propose a new index, and spend years advocating it. There are even more opportunities for creativity outside the table – e.g. rounds of drawing or number of urns could increase to three or more; marble colors could be any positive constant or variable; and marble distribution could be any percentage.

What we really need, however, is to fill the empty Column 2 titled "variable random," representing typical studies under a Grove-Riffe Scenario. We need reliability formulas based on empirical facts, rather than hypothetical imagination.

VIII.1. Liberal vs Conservative Estimates of Reliabilities

Do some indices regularly give higher scores than others? Earlier, by comparing chance agreements estimated by Scott (Table 3) and Cohen (Table 5), we established that Scott's π is more conservative than Cohen's κ. By comparing Goodman and Kruskal's Table 8 with other counterpart estimates, we found that λ_r is more conservative than all others.

Lombard et al (2002) used the "liberal" vs "conservative" concepts. Krippendorff (2004b, p. 412) objected, arguing that "trying to understand diverse agreement coefficients by their numerical results alone, conceptually placing them on a conservative-liberal continuum, is seriously misleading." We contend that patterns of numerical results can be helpful if they are grounded on an analysis of the underlying concepts and assumptions. Suppose we know that, with a large sample, λ_r is always lower than or equal to α, which is always lower than or equal to I_r, which is always lower than or equal to a_o, then if a researcher gets a very low λ_r, low α, high I_r, and very high a_o, she may look into the possibility that this is an artifact of the four indices, rather than focusing exclusively on possible deficiencies in her data, calculation or coding instrument.

The key is that this pattern or continuum must be based on a systematic and comprehensive comparison, rather than anecdotal observations of isolated cases. Such a comparison is now feasible, because --

First, of the 11 unique indices, the only difference between seven (*percent agreement* and equivalents, *S* and equivalents, λ_r, π and two equivalents, κ and an equivalent, α, and AC_1) is in chance agreement a_c. The other four are more complicated but still comparable, as β is an elevated κ, I_r is an elevated *S,* ρ is an approximate of *S,* and A_1 is a reweighted a_o.

Second, there is an inverse relation between chance agreement a_c and agreement index r_i. This can be proven by assuming $a_{c1} \geq a_{c2}$, replacing a_o in Equation 3 with a_{c1} and a_{c2} to obtain $r_{i1}=(a_o-a_{c1})/(1-a_{c1})$ and $r_{i2}=(a_o-a_{c2})/(1-a_{c2})$. Rearranging the equalities and inequalities, we have $a_{c1} \geq a_{c2} \rightarrow r_{i1} \leq r_{i2}$. So if Index A's a_c is often larger and never smaller than Index B's a_c, we may conclude with confidence that A is more conservative than B.

Third, chance agreement a_c for all indices have been calculated for binary scale with two coders. Five of them are in Tables 3, 5, 6, 8, and 9. We also know $a_c=0$ for a_o and A_1,

a_c=0.5 for S, I_r is an elevated S with the same a_c, β is an elevated κ with the same a_c, and ρ is an approximate of S with the same a_c.

So we can and should compare these a_c. If a hierarchy emerges for the nine a_c, it implies a reversed hierarchy for the nine groups of indices listed in Table 4.

[INSERT TABLE 12 Liberal vs Conservative Estimates of Reliability for Binary Scale, Two Coders, and Sufficiently Large Sample HERE]

The result of this comparison is in Table 12, which shows two hierarchies. The relative positions of any two indices in two different hierarchies are also meaningful, e.g., ρ is *generally* more liberal than β because ρ is in a higher cell in one hierarchy than β is in another hierarchy. They are in two different hierarchies because strict mathematical comparison between them does not yield stable results, i.e., in less frequent or less important situtions, an index in a lower cell in one hierarchy could produce a higher number than another index in a higher cell in another hierarchy. We assume two coders, binary scale, and reasonably large samples. When categories increase to three or more, category and double-based indices can be very liberal. When a sample reduces to 20 or below, Krippendorff's α can be very liberal.

To the extent that these indices have to be used, the liberal-conservative hierarchies in Table 12 may be helpful. If a researcher gets high scores from the most liberal indices, she should not assume everything is fine. If she gets low scores from the most conservative indices, she should not immediately abandon the study. In both cases, check what other indices say. Researchers might pay more attention to the more liberal indices at early stages of a study when the protocols are formulated and coders are trained, and pay more attention to the more conservative indices in the later stages, so as to be cautious before publication. We developed software to assist researchers to calculate the various indices. The software is available at http://reliability.hkbu.edu.hk.

VIII.2. Discussions and Recommendations

Reliability assesses the empirical foundation of research. Ironically, the foundation of inter-coder reliability calculation is more imaginative than empirical. Scientists and scholars tend to be skeptical that our findings are sound. We tend to guard against Type I errors more than Type II errors. We want to be rigorous, which often means conservative. This usually helpful tendency may have contributed to the development of some inter-coder reliability indices. But can we be too conservative? Are we overcorrecting?

Perhaps some designers of the indices wanted to estimate and remove the occasional dishonesty, and used maximum randomness as a surrogate. They probably did not realize their formulas assume that all coders maximize randomness, hence were all dishonest, in every study. We know dishonesty does not exist in large amounts in all data. Even if it exists, it has no consistent patterns that can be modeled or estimated mathematically.

We need an index of inter-coder reliability to accommodate typical research where coders try to be accurate but sometimes involuntarily allow some randomness. The existing indices do not meet this need. They assume either no or maximum randomness. The maximum-randomness assumption also entails other whimsical behaviors, such as setting quota or matching categories with marble colors. The chance-adjusted indices assume category, distribution or both as the factors affecting chance agreement, causing various paradoxes and abnormalities.

While a zero-random assumption likely overestimates reliability, we do not know when it overestimates or by how much. While maximum-random assumption may underestimate reliability in many situations, it may also overestimate in other situations, and, again, we do not know when it errs, at which direction, or by how much. We do know that some indices are more liberal than others, and the differences can be drastic.

When agreement is 100% and distribution is not 0% or 100%, major indices produce the same result -- r_i=1. The indices start to differ when a_o is lower than 100%. This implies that researchers can help to overcome deficiencies of the indices by perfecting their protocols, assuming their distributions are not skewed. The difficulty is that researchers cannot always expect perfect agreement or even distribution.

Researchers want the appearance of high reliability. The various indices and easy software allow shopping around until hitting the highest number. The two newer indices, I_r and AC_1, are more liberal than other chance-adjusted indices and are gaining in popularity. It should worry those striving to maintain high standards in academic publications. On the other hand, we should not equate low estimates with rigor, or complex calculations with sophistication. We should not require π or λ_r just for their low estimates. Given its unusual assumptions, we also should not require universal application of α, especially when the distribution is highly uneven or the sample is very small. We should not condemn a research just because the observed distribution is uneven, presuming that the coders have fallen asleep, agreed in advance, or had a broken instrument. We also should not reward small sample sizes.

The frequent use of π, κ and α may have had an undesired effect. All three favor more even distributions. Since the three have been applied by so many for so long, it may have reduced the publication of more uneven distributions of communication content and other things coded, rated, assessed, or diagnosed, making the world appear a bit more even than it actually is.

Our century-old concern over the zero-randomness assumption is legitimate. Our century-long search for a remedy assuming maximum-randomness and dishonest coders needs to stop. We need an index based on assumptions of variable-randomness and honest coders that uses degree of difficultiy, rather than category or distribution, as the main factor.

[INSERT TABLE 13 When to Use or Not Use Which Index of Reliability HERE]

Before such an index is established, researchers have to choose from the existing indices. We hope the practical recommendations in Table 13 can be of some help. As the table recommends various indices for various situations, we developed software available at http://reliability.hkbu.edu.hk to help researchers to calculate the indices. It is not a long-term solution. If and when the better index(es) is established, we should stop using Table 13 and the existing indices.

A major difference between indices is in their assumptions about coder behavior. Percent agreement indices assume coders never do any random coding, while chance-adjusted indices assume coders maximize random coding. Category-based indices assume coders draw from marbles of equal distribution, while distribution-based indices assume quotas. This article derived these assumptions through mathematical analysis. Social scientists may be more receptive of empirical evidences. Future research may test these assumptions as empirical hypotheses, through simulations and controlled experiments. For instance, a researcher may assign some participants to code according to a Bennett Scenario, and others to follow a Scott Scenario, yet others follow other scenarios. We may consider derived assumptions supported if the observed "wrong" agreements produced by a scenario, e.g. Cohen Scenario, are closest to or best correlated with the predictions of the corresponding index, i.e., κ.

References

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

Bennett, E. M., Alpert, R., & Goldstein, A. C. (1954). Communication 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.

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa, *Journal of Clinical Epidemiology, 46,* 423-429.

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

Cohen, J. (1968). Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70,* 213-220.

Feinstein, A. R., & Cicchetti, D.V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology, 43,* 543–549.

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

Glander, T. (2000). *Origins of Mass Communications Research during the American Cold War: Educational Effects and Contemporary Implications*. Mahwah, NJ: Erlbaum.

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

Grove, W. M., Andreasen, N. C., McDonald-Scott, P., Keller, M. B., & Shapiro, R. W. (1981). Reliability studies of psychiatric diagnosis: Theory and practice. *Archives of General Psychiatry, 38,* 408–413.

Guilford, J. P. (1961, November). *Preparation of item scores for correlation between*

individuals in a Q factor analysis. Paper presented at the annual convention of the Society of Multivariate Experimental Psychologists.

Guttman, L. (1946). The test-retest reliability of qualitative data. *Psychometrika, 11,* 81-95.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology, 61,* 29–48.

Gwet, K. L. (2010). *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters* (2nd ed.). Gaithersburg, MD: Advanced Analytics, LLC.

Hayes, A. F. (2009). Beyond Baron and Kenny: statistical mediation analysis in the new millennium. *Communication Monographs, 76,* 408-420.

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*, 749-753.

Holsti, O. R. (1969). *Content Analysis for the Social Sciences and Humanities.* Reading, MA: Addison-Wesley.

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

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

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

Krippendorff, K. (1980). *Content Analysis: An Introduction to its Methodology.* Newbury Park,CA: Sage Publications.

Krippendorff, K. (2004a). *Content Analysis: An Introduction to its Methodology* (2nd ed.).

Thousand Oaks,CA: Sage Publications.

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

Krippendorff, K. (2007). Computing Krippendorff's Alpha Reliability. *University of Pennsylvania Scholarly Commons,* http://repository.upenn.edu/asc_papers/43

Lasswell, H. D. (1948). The structure and function of communication in society. In L. Bryson (Ed.), *The Communication of Ideas*. NY: The Institute for Religious & Social Studies.

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.

Maxwell, A. E. (1977). Coefficients of agreement between observers and their interpretation. *British Journal of Psychiatry, 130,* 79–83.

Neuendorf, K. (2002). *The Content Analysis Guidebook*. Thousand Oaks, CA: Sage.

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

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

Popping, R. (1988). On agreement indices for nominal data. In W. E. Saris & I. N. Gallhofer (Eds.), *Sociometric research: Volume I, data collection and scaling* (pp. 90-105). New York: St. Martin's.

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

R Development Core Team (2011). R: A language and environment for statistical computing. Vienna, Austria, http://www.R-project.org/ (ISBN 3-900051-07-0)

Riffe, D., Lacy, S., & Fico, F. G. (1998). *Analyzing Media Messages: Using Quantitative*

Content Analysis in Research. Mahwah, N J: Lawrence Erlbaum Associates.

Riffe, D., Lacy, S., & Fico, F. G. (2005). *Analyzing Media Messages: Using Quantitative Content Analysis in Research* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Rogot, E. & Goldberg I. D. (1966). A proposed index for measuring agreement in test-retest studies. *Journal of Chronic Diseases, 19(9)*, 991-1006

Schramm, W. L. (1973). *Men, Messages, and Media: A Look at Human Communication*. New York: Harper & Row.

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

Shrout, P. E., Spitzer, R. L., & Fleiss, J. L. (1987). Quantification of agreement in psychiatric diagnosis revisited. *Archives of General Psychiatry, 44,* 172–177.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioural Sciences*. New York: McGraw-Hill.

Spitznagel, E. L., & Helzer, J. E. (1985). A proposed solution to the base rate problem in the kappa statistic. *Archives of General Psychiatry, 42,* 725–728.

Tankard, J. W., Jr. (1988). Wilbur Schramm: definer of a field. *Journalism Educator, 43(3),* 11-16.

Tinsley, H. E. A., & Weiss, D. J. (1975). Interrater Reliability and Agreement of Subjective Judgments. *Journal of Counseling Psychology, 22,* 358-376.

Tinsley, H. E. A., & Weiss, D. J. (2000). Interrater reliability and agreement. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 95-124). San Diego, CA: Academic Press.

Zhao, X. (2011a). *When to use Cohen's κ, if ever?* Paper presented at the 61st annual conference of International Communication Association, Boston.

Zhao, X. (2011b). *When to use Scott's π or Krippendorff's α, if ever?* Paper presented at the

annual conference of Association for Education in Journalism and Mass Communication, St. Louis.

Zhao, X., Lynch, J. G., & Chen, Q. (2010). Reconsidering Baron and Kenny: myths and truths about mediation analysis. *Journal of Consumer Research, 37,* 197-206.

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

Table 1
Reliability and Related Concepts

		Concepts of Consistency	
		Multi-Measure Reliability	Inter-Coder & Test-Retest Reliability
Scales	Categorical	Association / Covariation e.g. χ^2	Agreement / Proximity e.g. percent agreement
	Numerical	Correlation / Covariation. e.g. Pearson r & r^2	Agreement / Proximity e.g. closeness measure*

* Correlation indices, such as Pearson r or r^2, is at present the most often used indicator of inter-coder or test-retest reliability for numerical scales. Closeness measure would be a more appropriate measure, which we will discuss in another paper.

Table 2
A Typology of 22 Inter-coder Reliability Indices

		Adjusted for chance agreement?	
		Yes	No
On what basis is chance agreement estimated?	Category	ρ, S, (G, RE, C, k_n, $PABAK$, $rdf\text{-}Pi$)*, I_r	a_o, (Osgood's, Holsti's CR)*
	Distribution	β, λ_r, π, ($Rev\text{-}K$, BAK)*, κ, (A_2), α	
	Category & Distribution	AC_1	A_1

* Index(es) in parentheses is a mathematical equivalent(s) of the preceding index

	Index symbol	Author, Year	other known name of the index
1	α	Krippendorff, 1970, 1980.	
2	A_1	Rogot & Goldberg, 1966	
3	A_2	Rogot & Goldberg, 1966	
4	AC_1	Gwet, 2008, 2010.	
5	a_o	unknown author, pre 1901.	*Percent agreement*
6	β	Benini, 1901.	
7	BAK	Byrt et al., 1993.	
8	C	Jason & Vegelius, 1979	
9	CR	Holsti, 1969.	*Holsti's*
10	G	Guilford, 1961; Holley & Guilford, 1964.	
11	I_r	Perreault & Leigh, 1989.	
12	κ	Cohen, 1960.	
13	k_n	Brennan & Prediger, 1981.	
14	λ_r	Goodman & Kruskal, 1954.	
15	*Osgood's*	Osgood, 1959.	
16	π	Scott, 1955.	
17	$PABAK$	Byrt et al., 1993.	
18	$Rdf\text{-}Pi$	Potter & Levine-Donnerstein, 1999.	*Redefined Pi*
19	$Rev\text{-}K$	Siegel & Castellan, 1988.	*Revised K*
20	ρ	Guttman, 1946.	
21	RE	Maxwell, 1977.	
22	S	Bennett et al., 1954.	

Table 3

*Scott's Chance Agreement (a_c) as a Function of Two Distributions**

Distribution 2: Percent of Positive Findings by Coder 2 (N_{p2}/N)**	Distribution *1*: Percent of Positive Findings by Coder *1* (N_{p1}/N)**										
	0	10	20	30	40	50	60	70	80	90	100
100	50.0	50.5	52.0	54.5	58.0	62.5	68.0	74.5	82.0	90.5	100.0
90	50.5	50.0	50.5	52.0	54.5	58.0	62.5	68.0	74.5	82.0	90.5
80	52.0	50.5	50.0	50.5	52.0	54.5	58.0	62.5	68.0	74.5	82.0
70	54.5	52.0	50.5	50.0	50.5	52.0	54.5	58.0	62.5	68.0	74.5
60	58.0	54.5	52.0	50.5	50.0	50.5	52.0	54.5	58.0	62.5	68.0
50	62.5	58.0	54.5	52.0	50.5	50.0	50.5	52.0	54.5	58.0	62.5
40	68.0	62.5	58.0	54.5	52.0	50.5	50.0	50.5	52.0	54.5	58.0
30	74.5	68.0	62.5	58.0	54.5	52.0	50.5	50.0	50.5	52.0	54.5
20	82.0	74.5	68.0	62.5	58.0	54.5	52.0	50.5	50.0	50.5	52.0
10	90.5	82.0	74.5	68.0	62.5	58.0	54.5	52.0	50.5	50.0	50.5
0	100.0	90.5	82.0	74.5	68.0	62.5	58.0	54.5	52.0	50.5	50.0

*: Main cell entries are Scott's Chance Agreement (a_c) in %.

**: M_1 is the number of positive answers by Coder *1*, M_2 is the number of positive answers by Coder *2*, and N is the total number of cases analyzed. See also Table 4 for various assumptions behind Scott's π.

Table 4

Assumptions of 22 Inter-coder Reliability Indices

Down: Assumption name (assumption #)	% Agreement a_o (Osgood, Holsti's *CR*), Rogot & Goldberg's A_1	Benini's β	Guttman's ρ	Bennett' et al's *S* (*C, G,* k_n, *PABAK, rdf-Pi, RE*)	Goodman & Kruskal's λ_r	Scott's π (*Rev-K*, *BAK*)	Cohen's κ (A_2)	Krippen-dorff's α	Perreault & Leigh's I_r	Gwet's AC_1
Random chance agreement (1, 2)	zero	maximum	maximum	maximum	maximum	maximum	maximum	maximum	maximum	maximum
Honesty (3)	complete	limited	limited	limited	limited	limited	limited	limited	limited	limited
Specified random (4)	no	yes	yes	yes	yes	yes	yes	yes	yes	yes
Rounds of marble drawing (23)	zero	one	one	one	one	one	one	one	one	two
Drawing with replacement (7, 18)	N/A	yes	yes	yes	yes	yes	yes	no	yes	yes
What marble pattern = honesty? (8, 22, 24)	N/A	mismatch	mismatch	mismatch	mode color	mismatch	mismatch	mismatch	mismatch	mismatch or 2 matches
Categories=colors (5)	no	yes	yes	yes	yes	yes	yes	yes	yes	yes
Equal number per color (6)	no	no	yes	yes	no	no	no	no	yes	yes
Categories reduce chance agreements a_c (9)	no	no	yes	yes	no	no	no	no	yes	yes
Agreement observed or approximated (10)	observed	observed	approximated	observed	observed	observed	observed	observed	observed	observed
Elevated index (11)	no	yes	no	no	no	no	no	no	yes	no
Quota (12, 17)	no	individual	no	no	individual	conspired	individual	conspired	no	conspired
Trinity distribution (13)	no	yes	no	no	yes	yes	yes	yes	no	yes
Constrained task (14)	no	yes	no	no	yes	yes	yes	yes	no	yes
Predetermined distribution (15)	no	yes	no	no	yes	yes	yes	yes	no	yes
Quota & Distribution affects a_c (16)	no	yes	no	no	yes	yes	yes	yes	no	yes
Trinity size (19)	no	no	no	no	no	no	no	yes	no	no
Predetermined target size (20)	no	no	no	no	no	no	no	yes	no	no
Larger samples increase a_c (21)	no	no	no	no	no	no	no	yes	no	no

Table 5

*Cohen's Chance Agreement (a_c) as a Function of Two Distributions**

Distribution 2: Positive Findings by Coder 2 (N_{p2}/N) in %**	Distribution *1:* Positive Findings by Coder *1* (N_{p1}/N) in %**										
	0	10	20	30	40	50	60	70	80	90	100
100	0.0	10.0	20.0	30.0	40.0	50.0	60.0	70.0	80.0	90.0	100.0
90	10.0	18.0	26.0	34.0	42.0	50.0	58.0	66.0	74.0	82.0	90.0
80	20.0	26.0	32.0	38.0	44.0	50.0	56.0	62.0	68.0	74.0	80.0
70	30.0	34.0	38.0	42.0	46.0	50.0	54.0	58.0	62.0	66.0	70.0
60	40.0	42.0	44.0	46.0	48.0	50.0	52.0	54.0	56.0	58.0	60.0
50	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
40	60.0	58.0	56.0	54.0	52.0	50.0	48.0	46.0	44.0	42.0	40.0
30	70.0	66.0	62.0	58.0	54.0	50.0	46.0	42.0	38.0	34.0	30.0
20	80.0	74.0	68.0	62.0	56.0	50.0	44.0	38.0	32.0	26.0	20.0
10	90.0	82.0	74.0	66.0	58.0	50.0	42.0	34.0	26.0	18.0	10.0
0	100.0	90.0	80.0	70.0	60.0	50.0	40.0	30.0	20.0	10.0	0.0

*: Main cell entries are Cohen's Chance Agreement (a_c) in %.

**: N_{p1} is the number of positive answers by Coder *1*, N_{p2} is the number of positive answers by Coder *2*, and N is the total number of cases analyzed. See also Table 4 for various assumptions behind Cohen's κ.

Table 6

*Krippendorff's Chance Agreement (a_c) as a Function of Two Distributions (N=100)**

		Distribution *1*: Percent of Positive Findings by Coder *1* (N_{p1}/N)**										
		0	10	20	30	40	50	60	70	80	90	100
Distribution 2: Percent of Positive Findings by Coder 2 (N_{p2}/N)**	100	49.7	50.3	51.8	54.3	57.8	62.3	67.8	74.4	81.9	90.5	100.0
	90	50.3	49.7	50.3	51.8	54.3	57.8	62.3	67.8	74.4	81.9	90.5
	80	51.8	50.3	49.7	50.3	51.8	54.3	57.8	62.3	67.8	74.4	81.9
	70	54.3	51.8	50.3	49.7	50.3	51.8	54.3	57.8	62.3	67.8	74.4
	60	57.8	54.3	51.8	50.3	49.7	50.3	51.8	54.3	57.8	62.3	67.8
	50	62.3	57.8	54.3	51.8	50.3	49.7	50.3	51.8	54.3	57.8	62.3
	40	67.8	62.3	57.8	54.3	51.8	50.3	49.7	50.3	51.8	54.3	57.8
	30	74.4	67.8	62.3	57.8	54.3	51.8	50.3	49.7	50.3	51.8	54.3
	20	81.9	74.4	67.8	62.3	57.8	54.3	51.8	50.3	49.7	50.3	51.8
	10	90.5	81.9	74.4	67.8	62.3	57.8	54.3	51.8	50.3	49.7	50.3
	0	100.0	90.5	81.9	74.4	67.8	62.3	57.8	54.3	51.8	50.3	49.7

*: Main cell entries are Krippendorff's Chance Agreement (a_c) in %.

**: N_{p1} is the number of positive answers by Coder *1*, N_{p2} is the number of positive answers by Coder *2*, and *N* is the total number of cases analyzed. See also Table 4 for various assumptions behind Krippendorff's α.

Table 7

Krippendorff's Chance Agreement Rate (a_c) as a Function of Coded Targets (N) and Average Distribution (N_p/N)

		Average Distribution of Positive Cases (N_p/N in %)										
		0	10	20	30	40	50	60	70	80	90	100
Number of Coded Targets (N)	1	100.00	64.00	36.00	16.00	4.00	0.00	4.00	16.00	36.00	64.00	100.00
	2	100.00	76.00	57.33	44.00	36.00	33.33	36.00	44.00	57.33	76.00	100.00
	3	100.00	78.40	61.60	49.60	42.40	40.00	42.40	49.60	61.60	78.40	100.00
	4	100.00	79.43	63.43	52.00	45.14	42.86	45.14	52.00	63.43	79.43	100.00
	5	100.00	80.00	64.44	53.33	46.67	44.44	46.67	53.33	64.44	80.00	100.00
	6	100.00	80.36	65.09	54.18	47.64	45.45	47.64	54.18	65.09	80.36	100.00
	7	100.00	80.62	65.54	54.77	48.31	46.15	48.31	54.77	65.54	80.62	100.00
	8	100.00	80.80	65.87	55.20	48.80	46.67	48.80	55.20	65.87	80.80	100.00
	9	100.00	80.94	66.12	55.53	49.18	47.06	49.18	55.53	66.12	80.94	100.00
	10	100.00	81.05	66.32	55.79	49.47	47.37	49.47	55.79	66.32	81.05	100.00
	11	100.00	81.14	66.48	56.00	49.71	47.62	49.71	56.00	66.48	81.14	100.00
	12	100.00	81.22	66.61	56.17	49.91	47.83	49.91	56.17	66.61	81.22	100.00
	13	100.00	81.28	66.72	56.32	50.08	48.00	50.08	56.32	66.72	81.28	100.00
	14	100.00	81.33	66.81	56.44	50.22	48.15	50.22	56.44	66.81	81.33	100.00
	15	100.00	81.38	66.90	56.55	50.34	48.28	50.34	56.55	66.90	81.38	100.00
	16	100.00	81.42	66.97	56.65	50.45	48.39	50.45	56.65	66.97	81.42	100.00
	17	100.00	81.45	67.03	56.73	50.55	48.48	50.55	56.73	67.03	81.45	100.00
	18	100.00	81.49	67.09	56.80	50.63	48.57	50.63	56.80	67.09	81.49	100.00
	19	100.00	81.51	67.14	56.86	50.70	48.65	50.70	56.86	67.14	81.51	100.00
	20	100.00	81.54	67.18	56.92	50.77	48.72	50.77	56.92	67.18	81.54	100.00
	21	100.00	81.56	67.22	56.98	50.83	48.78	50.83	56.98	67.22	81.56	100.00
	22	100.00	81.58	67.26	57.02	50.88	48.84	50.88	57.02	67.26	81.58	100.00
	23	100.00	81.60	67.29	57.07	50.93	48.89	50.93	57.07	67.29	81.60	100.00
	24	100.00	81.62	67.32	57.11	50.98	48.94	50.98	57.11	67.32	81.62	100.00
	25	100.00	81.63	67.35	57.14	51.02	48.98	51.02	57.14	67.35	81.63	100.00
	26	100.00	81.65	67.37	57.18	51.06	49.02	51.06	57.18	67.37	81.65	100.00
	27	100.00	81.66	67.40	57.21	51.09	49.06	51.09	57.21	67.40	81.66	100.00
	28	100.00	81.67	67.42	57.24	51.13	49.09	51.13	57.24	67.42	81.67	100.00
	29	100.00	81.68	67.44	57.26	51.16	49.12	51.16	57.26	67.44	81.68	100.00
	30	100.00	81.69	67.46	57.29	51.19	49.15	51.19	57.29	67.46	81.69	100.00

Table 8

Goodman and Kruskal's Chance Agreement (a_c) as a Function of Two Distributions[*]

Distribution 2: Positive Findings by Coder 2 (N_{p2}/N) in %**	Distribution *1:* Positive Findings by Coder *1* (N_{p2}/N) in %**										
	0	10	20	30	40	50	60	70	80	90	100
100	100.0	95.0	90.0	85.0	80.0	75.0	80.0	85.0	90.0	95.0	100.0
90	95.0	90.0	85.0	80.0	75.0	70.0	75.0	80.0	85.0	90.0	95.0
80	90.0	85.0	80.0	75.0	70.0	65.0	70.0	75.0	80.0	85.0	90.0
70	85.0	80.0	75.0	70.0	65.0	60.0	65.0	70.0	75.0	80.0	85.0
60	80.0	75.0	70.0	65.0	60.0	55.0	60.0	65.0	70.0	75.0	80.0
50	75.0	70.0	65.0	60.0	55.0	50.0	55.0	60.0	65.0	70.0	75.0
40	80.0	75.0	70.0	65.0	60.0	55.0	60.0	65.0	70.0	75.0	80.0
30	85.0	80.0	75.0	70.0	65.0	60.0	65.0	70.0	75.0	80.0	85.0
20	90.0	85.0	80.0	75.0	70.0	65.0	70.0	75.0	80.0	85.0	90.0
10	95.0	90.0	85.0	80.0	75.0	70.0	75.0	80.0	85.0	90.0	95.0
0	100.0	95.0	90.0	85.0	80.0	75.0	80.0	85.0	90.0	95.0	100.0

*: Main cell entries are Goodman and Kruskal's Chance Agreement (a_c) in %.

**: N_{p1} is the number of positive answers by Coder *1*, N_{p2} is the number of positive answers by Coder *2*, and N is the total number of cases analyzed. See also Table 4 for various assumptions behind Cohen's κ and Goodman and Kruskal's λ_r.

Table 9

*Gwet's Chance Agreement (a_c) as a Function of Two Distributions**

Distribution *2*: Percent of Positive Findings by Coder *2* (N_{p2}/N)**	Distribution *1*: Percent of Positive Findings by Coder *1* (N_{p1}/N)**										
	0	10	20	30	40	50	60	70	80	90	100
100	50.0	49.5	48.0	45.5	42.0	37.5	32.0	25.5	18.0	9.5	0.0
90	49.5	50.0	49.5	48.0	45.5	42.0	37.5	32.0	25.5	18.0	9.5
80	48.0	49.5	50.0	49.5	48.0	45.5	42.0	37.5	32.0	25.5	18.0
70	45.5	48.0	49.5	50.0	49.5	48.0	45.5	42.0	37.5	32.0	25.5
60	42.0	45.5	48.0	49.5	50.0	49.5	48.0	45.5	42.0	37.5	32.0
50	37.5	42.0	45.5	48.0	49.5	50.0	49.5	48.0	45.5	42.0	37.5
40	32.0	37.5	42.0	45.5	48.0	49.5	50.0	49.5	48.0	45.5	42.0
30	25.5	32.0	37.5	42.0	45.5	48.0	49.5	50.0	49.5	48.0	45.5
20	18.0	25.5	32.0	37.5	42.0	45.5	48.0	49.5	50.0	49.5	48.0
10	9.5	18.0	25.5	32.0	37.5	42.0	45.5	48.0	49.5	50.0	49.5
0	0.0	9.5	18.0	25.5	32.0	37.5	42.0	45.5	48.0	49.5	50.0

*: Main cell entries are Gwet's Chance Agreement (a_c) in %.

**: N_{p1} is the number of positive answers by Coder *1*, N_{p2} is the number of positive answers by Coder *2*, and N is the total number of cases analyzed. See also Table 4 for various assumptions behind Gwet's AC_1.

Table 10

Paradoxes and Abnormalities of 22 Inter-coder Reliability Indices

Paradox or Abnormality #	Paradox or Abnormality	% Agreement a_o, (Osgood, Holsti's *CR*), Rogot & Goldberg's A_1	Bennett et al's *S*, Guttman's ρ, Perreault & Leigh's I_r, (*C, G,* k_n, *PABAK, rdf-Pi, RE*)*	Scott's π, (*Rev-K*, *BAK*)	Cohen's κ, (Rogot & Goldberg's A_2), Benini's β, Goodman & Kruskal's λ_r,	Krippendorff's α	Gwet's AC_1
Prdx 1	Random guessing is reliable	yes					
Prdx 2	Nothing but chance		yes	yes	yes	yes	yes
Prdx 3	Apples compared with oranges		yes	yes	yes	yes	yes
Prdx 4	Humans are subgroup of men		yes	yes	yes	yes	yes
Prdx 5	Pandas are subgroup of men		yes	yes	yes	yes	yes
Prdx 6	Categories increase reliability		yes				yes
Prdx 7	Punishing larger sample & replicability					yes	
Prdx 8	Purely random coding is reliable					yes	
Prdx 9	Randomness more reliable than honesty					yes	
Abn 10	High agreement, low reliability			yes	yes	yes	
Abn 11	Undefined reliability			yes	yes	yes	
Abn 12	No change in a_o, large drop in reliability			yes	yes	yes	
Abn 13	Zero disagreement, no improvement in r_i			yes	yes	yes	
Abn 14	Tiny rise in a_o, huge rise in r_i			yes	yes	yes	
Abn 15	Rise in a_o, huge drop in r_i			yes	yes	yes	
Abn 16	Honest coding as bad as coin flipping			yes	yes	yes	
Prdx 17	Punishing improved coding			yes	yes	yes	
Prdx 18	Punishing agreement			yes	yes	yes	
Prdx 19	Moving bar			yes	yes	yes	yes
Prdx 20	Circular logic			yes	yes	yes	yes
Abn 21	Same quality, same a_o, higher r_i						yes
Abn 22	Lower quality, lower a_o, higher r_i						yes

Table 11

What's Missing in the Map of Reliabilities?

<table>
<tr><td colspan="4">1. Maximum Random</td><td>2.Variable Random</td><td>3. Zero Random</td></tr>
<tr><td colspan="2" rowspan="2"></td><td colspan="2">Observed Distribution = Marble Distribution</td><td rowspan="2">Categories = Colors</td><td rowspan="8"></td><td rowspan="8">Percent Agreement (a_o)
Osgood's coefficient,
Holsti's CR
Rogot and Goldberg's A_1</td></tr>
<tr><td>Individual Quota</td><td>Conspired Quota</td></tr>
<tr><td rowspan="2">Color Mismatch= Honesty</td><td>Replacement drawing</td><td>κ, A_2, β</td><td>π, Rev-K, BAK</td><td>ρ, S, G, RE, C, k_n I_r, PABAK, rdf-Pi.</td></tr>
<tr><td>Non-replacement drawing</td><td></td><td>α</td><td></td></tr>
<tr><td rowspan="2">Mismatch or Double Match = Honesty</td><td>Replacement Drawing</td><td></td><td>AC_1*</td><td>AC_1*</td></tr>
<tr><td>Non-replacement drawing</td><td></td><td></td><td></td></tr>
<tr><td rowspan="2">Largest Color= Honesty</td><td>Replacement Drawing</td><td>λ_r</td><td></td><td></td></tr>
<tr><td>Non-replacement drawing</td><td></td><td></td><td></td></tr>
</table>

* AC_1 occupies two cells because it is double based, on category *and* distribution.

Table 12

Liberal vs Conservative Estimates of Reliability for Binary Scale, Two Coders, and Sufficiently Large Sample

	Hierarchy 1	Hierarchy 2
More *liberal* estimates of reliability.	Percent Agreement (a_o) (pre 1901), Osgood's (1959), Holsti's *CR* (1969), Rogot & Goldberg's A_1 (1966)	Percent Agreement (a_o) (pre 1901), Osgood's (1959), Holsti's *CR* (1969) Rogot & Goldberg's A_1 (1966)
	Perreault & Leigh's I_r (1989)	
	Gwet's AC_1 (2008, 2010)	
	Guttman's ρ (1946), Bennett et al.'s *S* (1954), Guilford's *G* (1961), Maxwell's *RE* (1977), Jason & Vegelius' *C* (1979), Brennan & Prediger's k_n (1981), Byrt et al.'s *PABAK* (1993) Potter & Levine-Donnerstein's *rdf-Pi* (1999).	
		Benini's β (1901)
		Cohen's κ (1960) Rogot & Goldberg's A_2 (1966)
	Krippendorff's α (1970, 1980)	Krippendorff's α (1970, 1980)
More *conservative* estimates of reliability.	Scott's π (1955), Siegel & Castellan's *Rev-K* (1988), Byrt et al's BAK (1993)	Scott's π (1955), Siegel & Castellan's *Rev-K* (1988), Byrt et al's BAK (1993)
	Goodman & Kruskal's λ_r (1954)	Goodman & Kruskal's λ_r (1954)

Comparisons across the dotted lines are between the general patterns in situations that are more frequent and more important for typical research, e.g., when indices are zero or above, and when the distribution estimates of two coders are not extremely skewed in opposite directions. Comparisons involving Guttman's ρ, its eight equivalents, and Perreault & Leigh's I_r assume binary scale. Comparisons involving Krippendorff's α assume sufficiently large sample.

Table 13

When to Use or Not Use Which Index of Reliability

Down: observed condition	Indices that tend to produce ***unfairly low*** reliability scores	Indices that tend to produce ***unfairly high*** reliability scores	Indices ***not obviously unfair*** due to the observed condition at the left, hence may be considered for temporary use until a more reasonable index is available [iv, v, vi]
Low agreement		Percent Agreement a_o, Osgood's, Holsti's *CR,* Rogot and Goldberg's A_1	Gwet's AC_1, Perreault & Leigh's I_r, Bennett et al's *S*, Cohen's κ, Scott's π, Krippendorff's α
Highly uneven individual distribution	Benini's β [i], Goodman & Kruskal's λ_r, Scott's π, Cohen's κ [i], Rogot & Goldberg's A_2, Krippendorff's α, Byrt et al's *BAK*, Siegel and Castellan's *Rev-K* (1988)	Benini's β [i], Cohen's κ [i], Rogot & Goldberg's A_2, Gwet's AC_1,	Percent Agreement a_o, Perreault & Leigh's I_r, Bennett et al's *S*
Highly uneven average distribution	Benini's β, Goodman & Kruskal's λ_r, Scott's π, Byrt et al's *BAK*, Siegel and Castellan's *Rev-K* (1988), Cohen's κ, Rogot & Goldberg's A_2, Krippendorff's α	Gwet's AC_1	Percent Agreement a_o, Perreault & Leigh's I_r, Bennett et al's *S*
$\rho \approx 0.5$		Perreault & Leigh's I_r	Percent Agreement a_o, Gwet's AC_1, Bennett et al's *S*, Cohen's κ, Scott's π, Krippendorff's α
$N<20$ [ii]		Krippendorff's α	Percent Agreement a_o, Gwet's AC_1, Perreault & Leigh's I_r, Bennett et al's *S*, Cohen's κ, Scott's π
$K\geq 3$ [iii]		Guttman's ρ, Perreault & Leigh's I_r, Bennett et al.'s *S*, Guilford's *G*, Maxwell's *RE*, Jason & Vegelius' *C,* Brennan & Prediger's k_n, , Byrt et al's *PABAK*, Potter & Levine-Donnerstein's *redefined Pi*, Gwet's AC_1	Percent Agreement a_o, Cohen's κ, Scott's π Krippendorff's α

Table 13 (Continued)

i When individual distributions are highly uneven, Benini's β and Cohen's κ can be unfairly high when the two distributions are highly skewed at the opposite directions, e.g., one coder reports 95% positive while the other 95% negative; the two can be unfairly low when the two distributions are skewed at the same direction, e.g., both coders report 95% positive.

ii *N* is number of target cases analyzed.

iii *K* is number of categories in the nominal coding scale.

iv Use with caution! While the indices in the extreme right cells are not necessarily unfair due to the observed condition in the extreme left cells of the same row, they may be unfair due to other condition(s) present in a study. For example, when a study uses three or more categories (last row), it does not make Scott's π unfair. But the same study may also have highly uneven distribution (second and third rows), which makes π unfairly low, so the researcher may have to use percent agreement. Combination of conditions could make all available indices unfair for a given study, which is one of the reasons that a better index is needed.

v In each cell of this column, the indices are listed according to their positions in the liberal-conservative hierarchies shown in Table 12. The information may be useful for meta analysts and other content analysts who wish to better evaluate their reliability level.

vi We excluded all "equivalents" from this "not obviously unfair" column, as credits should go to the first designer(s).