When to use Scott’s $\pi$ or Krippendorff’s $\alpha$, if ever?

Xinshu Zhao
Hong Kong Baptist University, zhao@hkbu.edu.hk

Follow this and additional works at: https://repository.hkbu.edu.hk/coms_conf

Part of the Communication Commons

This document is the authors' final version of the published article.

Recommended Citation
Zhao, Xinshu (2011). “When to use Scott’s $\pi$ or Krippendorff’s $\alpha$, if ever?” paper presented at the annual conference of Association for Education in Journalism and Mass Communication, St. Louis, USA, August
When to Use Scott’s π or Krippendorff’s α, If Ever?

WHEN TO USE SCOTT’S π OR KRIPPENDORFF’S α, IF EVER?

Xinshu Zhao

Paper presented at the 2011 annual conference
of Association for Education in Journalism and Mass Communication, St. Louis, USA.

Suggested citation

Zhao, Xinshu (2011). “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, USA, August.

Xinshu Zhao, Ph.D., Chair Professor of Communication, Hong Kong Baptist University. He is also Cheung Kong Professor of Journalism and Research Fellow, Center for Information and Communication Studies, Fudan University. He was Professor and Director, Center for Research in Journalism and Mass Communication, University of North Carolina at Chapel Hill, when this study was conducted.

Acknowledgment

While this paper was presented at the AEJMC conference in August 2011, St. Louis, USA, it was archived and made available online through Hong Kong Baptist University Institutional Repository in August 2016. The archiving benefited from a donation by Dr. Lee Shau Kee, GBM, through HKBU Research Grant Scheme (LSK/14-15/P13, Zhao PI), and a China Ministry of Education Major Grant for Social Sciences through Fudan University Center for Information and Communication Studies (11JJD860007, Zhao PI).

鳴謝

本研究 2011 年 8 月在香港浸會大學「網上檔案」存檔上網。存檔上網工作部分得益於香港浸會大學「李兆基博士資助研究計劃」 (LSK/14-15/P13，趙心樹主持) 及復旦大學信息與傳播研究中心「教育部人文社會科學重點研究基地重大項目」（編號 11JJD860007，趙心樹主持）的資助。
WHEN TO USE SCOTT’S $\pi$ OR KRIPPENDORFF’S $\alpha$, IF EVER?

Xinshu Zhao

Author Note

Xinshu Zhao is Cheung Kong Chair Professor of Journalism, Fudan University, Shanghai, China and Chair Professor and Dean of Communication, Hong Kong Baptist University. He was Professor and Director, Center for Research in Journalism and Mass Communication, University of North Carolina at Chapel Hill, when the main theses of this study were formulated.

This study was supported in part by grants from NICHD (R24 HD056670, Henderson PI), UNC-CH Center for AIDS Research (#07-1191, Brown PI) and Panmedia Institute (2010, Zhao PI). The author also thanks Visne K. C. Chan, Juntao Kenneth He, Sri Kalyanaraman, Jing Lucille Li, and Ning Mena Wang for their support and assistance.

Correspondence concerning this article should be addressed to: Xinshu ZHAO, School Office, School of Communication, Hong Kong Baptist University, CVA1018D, Communication and Visual Arts Building, 5 Hereford Road, Kowloon Tong, Hong Kong. Tel: 852-3411-7492, Fax: 852-3411-7375, zhao@hkbu.edu.hk

Last Revision March 7, 2011
Abstract

Scott’s π (1955) and Krippendorff's α (1980) have been among the most-often used or recommended general indicators of reliability. This article presents paradoxes showing that neither can be a general indicator. We show that π or α should be used only when (a) coders enforce a predetermined quota as the first priority and (b) coders maximize chance coding as the second priority. Because the two conditions rarely hold, π or α should rarely be used.
WHEN TO USE SCOTT’S \( \pi \) OR KRIPPENDORFF’S \( \alpha \), IF EVER?

Inter-coder or test-retest reliability is widely used to assess measurement quality, in such disciplines as communication, medical studies, and marketing research. Various methods have been proposed to calculate reliability, including Yule’s \( Y \) (1912), Bennett et al’s \( S \) (1954), Scott’s \( \pi \) (1955), Osgood’s index (1959), Cohen’s \( \kappa \) (1960), Holsti’s coefficient (1969), Maxwell’s RE (1977), Krippendorff’s \( \alpha \) (1980), Perreault & Leigh’s I, (1989), Feinstein & Cicchetti’s \( \psi \) (1990b), Potter & Levine-Donnerstein’s redefined \( \pi \) (1999), and Gwet’s \( \gamma \) (2006). Each has been recommended and used as the general indicator of the same concept, reliability. Yet they offer different formulas and produce different, sometimes drastically different, results. This paper focuses on two of them, Scott's \( \pi \) and Krippendorff’s \( \alpha \).

I. Reliability vs. Reliabilities

Across various disciplines of social sciences and medical studies, Cohen’s \( \kappa \) (1960), which I have discussed in a separate paper, is the most often used indicator of reliability. Scott’s \( \pi \) (1955) is the second most popular, especially among communication researchers.

In Social Sciences Index (SSCI), between 1994 and 2010, Scott (1955) was cited 261 times and Krippendorff (1980) 243 times. During the same period, in Communication and Mass Media Complete (CMMC), “Scott’s Pi” had 597 citations, which rose from 11 in 1994 to 61 in 2009; “Krippendorff’s Alpha” had 145 citations, which also rose from 2 in 1994 to 26 in 2009. At the time of our search, the full year statistics were not yet available for 2010.

If Krippendorff’s \( \alpha \) has not been the most frequently used, it may be among the most strongly recommended. Respected communication methodologists such as Hayes & Krippendorff (2007) and Krippendorff (2004b) argued that \( \alpha \) should be the only general indicator to use.
The assumption is that the various indicators measure the same concept of reliability. Yet the indicators disagree with each other, as they produce different—sometimes drastically different—results, even when the underlying between-coder agreement is the same. As reliability means agreement (Riffe, Lacy, & Fico, 1998), these indicators of reliability themselves are not reliable, if they indeed measure the same concept.

Under the assumption of “various indicators for one reliability,” methodologists debated which indicator is the best, whether to use any or several of them in a study, and how to fix or cope when some indicators, especially Cohen’s κ, behave unexpectedly (e.g., Brennan & Prediger, 1981; Zwick, 1988; Lombard, Snyder-Duch & Bracken, 2002; Krippendorff, 2004b).

This paper ventures with a different assumption. Given that the indicators produce different results, there might be several reliability concepts, each having its own indicator. Only one can be the general indicator, while the others specialized. Under the assumption of “various indicators for various reliabilities,” we should analyze each indicator to determine whether it is general or special and, if special, to outline its boundaries.

As I have discussed Cohen’s κ elsewhere, this paper will focus on Scott's π and Krippendorff’s α. I will first briefly review the formulas of π and α. I will then introduce 18 paradoxes to show that neither measures general reliability. Assuming that they might measure reliability in special circumstances, I will analyze the mathematics of π and α to pin point these circumstances. I will use the simplest case—binary scales with two coders. The conclusions also apply to nominal scales of more categories with more coders.

II. Scott’s π and Krippendorff's α

Agreement rate \( (a_r) \) has been the simplest indicator of reliability recommended (Budd & Thorp 1963; Holsti 1969).

**Equation 1**

\[
a_r = \frac{A}{N}
\]
A is the number of cases on which two coders agree, and N is total number of cases analyzed. To simplify the analysis, I assume no missing value.

Many authors (Scott 1955; Krippendorff 1980; Riffe, Lacy, & Fico 1998) pointed out that \( a_r \) includes agreements due to chance \( (c_a) \), therefore overestimates reliability. Scott (1955) recommended Equation 2 to produce a reliability indicator, \( \pi \), that is corrected for chance agreement:

\[
\text{Equation 2} \quad \pi = \frac{a_r - c_a}{1 - c_a}
\]

Alternatively, \( \pi \) is often expressed as (Krippendorff, 1980, p.138; 2004a, p. 417) :

\[
\text{Equation 3} \quad \pi = 1 - \frac{d_r}{c_d}
\]

Here \( d_r \) is the observed disagreement rate and \( c_d \) is the expected disagreement rate due to chance. Equation 3 is equivalent to Equation 2 because:

\[
\text{Equation 4} \quad 1 = a_r + d_r \quad \text{and}
\]

\[
\text{Equation 5} \quad 1 = c_a + c_d
\]

To calculate the indicator \( \pi \) using Equation 2 or Equation 3, we need to know \( a_r \) and \( c_a \). While \( a_r \) is provided by the coders using Equation 1, \( c_a \) can only be estimated.

Scott (1955) calculated the average \( (M) \) of two coders' \( (i \text{ and } j) \) positive answers \( (M_i \text{ & } M_j) \),

\[
\text{Equation 6} \quad M = \frac{M_i + M_j}{2}
\]

and the average \( (W) \) of their negative answers \( (W_i \text{ & } W_j) \):

\[
\text{Equation 7} \quad W = \frac{W_i + W_j}{2}
\]

then use the averages to estimate Scott's \( c_a \):

\[
\text{Equation 8} \quad c_a = \left( \frac{M}{N} \right)^2 + \left( \frac{W}{N} \right)^2
\]
Krippendorff (1980) shares with Scott (1955) Equation 1, Equation 4, and Equation 5. The main formulas (Equation 9 & Equation 10) for Krippendorff's indicator corrected for chance agreement, α, are also identical to Scott's Equation 2 & Equation 3:

Equation 9 \[ \alpha = \frac{a_r - c_a}{1 - c_a} \]

Equation 10 \[ \alpha = 1 - \frac{d_r}{c_d} \]

Equation 9 and Equation 10 are equivalent to each other, for the same reason that Equation 2 and Equation 3 are equivalent to each other. The difference is in \( c_a \), expected chance agreement. To estimate \( c_a \), Scott (1955) multiplied average positive rate \((M/N)\) by itself, and multiplied the average negative rate \((W/N)\) by itself, as shown in Equation 8, which can be also expressed as:

Equation 11 \[ c_a = \left( \frac{2M}{2N} \right)^2 + \left( \frac{2W}{2N} \right)^2 \]

By contrast, Krippendorff (1980) subtracted 1 from the two multipliers’ numerators and denominators:

Equation 12 \[ c_a = \left( \frac{2M}{2N} \right) \left( \frac{2M-1}{2N-1} \right) + \left( \frac{2W}{2N} \right) \left( \frac{2W-1}{2N-1} \right) \]

Equation 9, Equation 10, and Equation 12 constitute Krippendorff's \( \alpha \) for binary scale with two coders. With multiple coders and multi-category nominal scales, Krippendorff's \( \alpha \) takes more complicated form (Hayes and Krippendorff, 2007; Krippendorff 2004a, 2004b). While this paper focuses on binary scale with two coders to outline the boundaries of legitimate application of the two indicators, these boundaries should also apply to more categories and more coders.

III. Fourteen Paradoxes of \( \pi \) and \( \alpha \)
While Scott’s π and Krippendorff’s α have been recommended and used as general indicators of reliability, their behavior often deviates from what’s expected from such an indicator. Here we list fourteen paradoxes.

**Paradox 1: High agreement rate, low π and α.** Suppose two coders coded 1,000 magazine advertisements for cigarettes in the United States. Their task was to see whether the Surgeon General's warning had been inserted as required by law. Suppose each coder found 999 “yes” and one “no,” with 998 positive agreements and two disagreements, generating a 99.8% agreement rate. But α and π would be below zero (-.001 and -.0005). Zero means totally unreliable. How could the instrument be so bad, when it produces near-perfect agreement rates? Is the symptom an indictment on the instrument, or on π and α?

**Paradox 2: Zero change in ar causing radical drop in π and α.** The two indicators are supposed to measure reliability, which is defined as agreement. Feinstein and Cicchetti (1990a) argued that a reliability indicator should rise and fall with agreement rate, ar.

Revising the Paradox 1 example, suppose two coders initially agreed on 998 “yes” and one “no,” plus one disagreement, producing ar=99.9%, π=.6662, and α=.6663. Suppose the two coders found a clerical mistake in the negative agreement and flipped it, resulting in 999 agreed positives and one disagreement, and increasing the average of the positive frequency from 99.85% to 99.95%. While ar remains 99.9%, π drops radically from .6662 to -.0005 and α from .6663 to .0000. The drop, which incidentally resulted from an improvement in coding instrument, covers two thirds of the distance between “perfectly reliable” and “totally unreliable.”

The “frequency” mentioned above was referred to as “prevalence” by Spitznagel & Helzer (1985) and Shrout, Spitzer & Fleiss (1987). Perreault & Leigh (1989) called it “distribution,” which we will also use hereafter. On a binary scale, a 50&50% distribution is the most even, while 0&100% is the most uneven.

**Paradox 3: Undefined π and α.** When two coders agree that the distribution of one category is 100% and another is 0%, π or α is undefined. If they are general indicators, they should have no blind spots. 0&100% and 100&0% are the two ends of all possible distributions, like the two ends of a ruler that define its length.
When to Use Scott’s \( \pi \) or Krippendorff’s \( \alpha \), If Ever?

and scope. If a ruler is completely messed up at both ends, we probably shouldn't hold much confidence that it would be accurate everywhere in between. It's possible that, as general indicators, \( \pi \) and \( \alpha \) have gene defects permeating everywhere, but only manifesting their worst symptoms at the two ends.

Krippendorff (2004b, p. 425) appeared to argue that this phenomenon is not paradoxical: “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.”

I argue it is paradoxical if \( \pi \) or \( \alpha \) is to be a general indicator. A general indicator must accommodate typical situations. The typical situation is that all coders do not always code all cases using broken instruments, falling asleep, or agreeing in advance. But Krippendorff might have touched upon something important. Maybe \( \pi \) and \( \kappa \) should be used only when coders agree in advance, which I will return to when attempting to resolve the paradoxes.

**Paradox 4:** Eliminating disagreements producing no improvement in \( \pi \) or \( \alpha \). Extending the Paradox 1 example: With just two disagreements, eliminating one or both should improve \( \pi \) and \( \alpha \) significantly, right? Wrong. Suppose one coder found an error in his only negative finding and flipped it, reducing disagreements by half, and increasing agreements to 99.9%. One might expect \( \pi \) and \( \alpha \) to improve half way toward 1, to be around 0.5. Instead, \( \alpha \) barely moves, to 0.0000, and \( \pi \) even remains negative, at -.0005. Suppose the other coder also flipped his only negative finding, improving agreement to 100%. Would \( \pi \) and \( \alpha \) jump to 1 for perfect reliability? No. Neither \( \pi \) nor \( \alpha \) can be calculated, like in Paradox 3!

**Paradox 5:** Tiny rise in \( a_r \) causing radical rise in \( \pi \) and \( \alpha \). With 998 agreements on “yes,” suppose one coder flipped his originally positive decision in one of the two disagreements. Now disagreements decrease to one and agreements increase to 999. \( a_r \) improves slightly from 99.8% to 99.9%. Given what we have seen in Paradox 4, one might expect \( \pi \) and \( \alpha \) to change little. Instead, \( \pi \) jumps from -.001 to .6662, while \( \alpha \) jumps from -.0005 to .6663, both covering two-thirds of the distance between “totally unreliable” and “perfectly reliable.” How can we rely on \( \pi \) or \( \alpha \) as an indicator of reliability when each jumps so radically with such small changes in the underlying agreement rate?
Paradox 6: Rise in $a_r$ causing radical drop in $\pi$ and $\alpha$. Suppose two coders initially had two disagreements and 998 agreements, with 997 positive and one negative, producing an $a_r=99.8\%$, $\pi=.499$, and $\alpha=.4992$. Suppose one coder found all his three negative decisions erroneous, and flipped each, resulted in 999 positive agreements and one disagreement. While $a_r$ increases to 99.9%, $\alpha$ drops drastically to 0.00%, and $\pi$ drops even more, to -0.0005.

Paradox 7: Radically and erratically moving bar. To highlight the dramatic paradoxes, 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 the symptoms would not appear as dramatic. Scott’s and Krippendorff’s chance agreements ($c_a$) are both functions of distribution. Uneven distribution produces higher $c_a$, as is shown in Tables 1a and 1b.

---

Table 1a. Scott’s Chance Agreement ($c_a$) as a Function of Two Distributions
Table 1b. Krippendorff’s Chance Agreement ($c_a$) as a Function of Two Distributions
---

A cell-by-cell comparison shows that Tables 1a and 1b are almost identical. When $N$ is 100 or above, Scott's and Krippendorff’s $c_a$ are basically the same, hence $\pi$ and $\alpha$ are essentially the same.

$c_a$ is the bar that $a_r$ must pass in order to produce an above-zero $\pi$ or $\alpha$, according to Equation 2 and Equation 9. $c_a$ and $a_r$ both have 100% as the maximum. The closer is $c_a$ to 100%, the less room above it for $a_r$, the less chance for a high $\pi$ or $\alpha$. When distribution reaches 0&100%, $c_a$ reaches 100%, leaving no chance for $a_r$ to pass $c_a$. That’s the technical reason $\pi$ and $\alpha$ are undefined there (Paradox 3).

Tables 1a and 1b give comprehensive views of how $c_a$ changes with the distributions reported by the two coders. Figures 1a and 1b visualize how large and erratic the changes are. Figure 2 presents one slice from Figures 1a & 1b, and Figure 3 presents five additional slices.

---

Figure 1a. Scott’s Chance Agreement ($c_a$) as a Function of Two Distributions
Figure 1b. Krippendorff’s Chance Agreement ($c_a$) as a Function of Two Distributions
---
These six figures and two tables show how gradually, yet sweepingly, the two $c_a$ bars move — both can reach as high as 100%, but neither jumps there suddenly. They instead move gradually with no gap or abrupt turn, from 50% ($\pi$) or below 50% ($\alpha$) as the minimum.

This supports our suspicion that “undefined $\pi$” and “undefined $\alpha$” should not be seen just as isolated incidents under extreme circumstances. They are also crying symptoms of pervasive defects hidden in $\pi$ and $\alpha$ as. The moving bars also explain why $\pi$ and $\alpha$ change with distribution (Paradox 2). So the questions become: why should the “bars” move so much, covering half (Scott’s) or more (Krippendorff’s) of the 0–1 scale? Why should the bar move at all depending on distributions?

**Paradox 8: Punishing agreement.** $c_a$ not only moves significantly, it also moves to punish the good and reward the bad. Table 1a and Figures 1a and 3 show that when one coder’s distribution $M_j/N$ is 100%, Scott's $c_a$ is positively linked to the other coder’s distribution $M_i/N$; an increase in $M_i/N$ brings it closer to $M_j/N$, producing a higher agreement $a_r$ and at the same time a higher $c_a$, which means a higher bar. The same pattern exists when $M_i/N =100\%$, $M_j/N=0\%$, or $M_i/N =0\%$. The maximum agreement at the lower left and upper right corners makes $c_a=100\%$, which is impossible to pass. On the other hand, as agreement rate decreases from either corner along any of the four sides, $c_a$ decreases at an averaged half rate, until maximum disagreement at the upper left or lower right corner where $c_a=50\%$.

Krippendorff’s $c_a$ behaves in a very similar pattern, except more erratically and the paradox more dramatic with smaller $N$, as shown in Tables 1b and 2, and Figures 1b, 3, and 4.

While the paradoxical pattern is strongest in the four sides encompassing Tables 1a and 1b and Figures 1a and 1b, it also manifests inside. Scott’s $\pi$ and Krippendorff’s $\alpha$ are supposedly indicators of reliability, which means agreement. Why do they *systematically* punish agreement and reward disagreement?
Paradox 9: Honest work worse than coin flipping. Suppose we show at normal speed 60 television segments, 50 of which contain subliminal advertisements barely recognizable. One coder found the ads in all 60 segments, making 10 false alarms, while the other recognized only 40, calling 10 false negatives. The 40 positive agreements and 20 disagreements produce a 66.667% \( \alpha \) and a 83.333% average distribution, which perfectly matches the target distribution that we knew. While the instrument may seem adequate, especially considering the difficult task, Scott's \( \pi \) and Krippendorff's \( \alpha \) are both negative, at miserable -0.2 and -0.19.

Now suppose we ask the coders to flip coins without looking at any television segments, ever. Their \( \alpha \) 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 totally dishonest coding, however, produces \( \pi=0 \) and \( \alpha=0.0083 \).

Honest coding that produces more accuracy and agreement is worse than dishonest coding that produces less accuracy and agreement, according to Scott's \( \pi \) or Krippendorff's \( \alpha \).

Paradox 10: Punishing Improved Coding. In Paradox 6, the improvement in \( \alpha \) is due to one coder correcting his errors, hence a case of improved coding causing a drastic drop in \( \pi \) and \( \alpha \). (from halfway reliable (0.5) to not at all reliable (0)!) Of all the paradoxical symptoms of \( \pi \) and \( \alpha \), this one may be among the most troublesome.

Paradox 2 is another example: While the coding improved after coders corrected their errors, \( \pi \) and \( \alpha \) drop even more drastically, from nearly .67 to about zero.

Paradox 11: Circular logic. Scott’s \( \pi \) and Krippendorff’s \( \alpha \) are functions of the observed distribution, whose quality is dependent on the quality of the coding instrument. But that’s the very instrument that \( \pi \) and \( \alpha \) are to evaluate. Scott’s \( \pi \) and Krippendorff’s \( \alpha \) depend on the instrument’s reliability to assess the instrument’s reliability!

Paradox 12: Nothing but chance. Equation 5, which says \( 1=c_a+c_d \), represents a critical assumption in Scott's \( \pi \) and Krippendorff’s \( \alpha \): chance coding, which includes chance agreements \( (c_a) \) and chance
disagreements ($c_d$), constitutes 100% of the expected outcomes. But where is honest coding? Are there at least some coders who have at least some honesty? If so, should the formula instead look like the following?

**Equation 13**  \[ 1 = c_a + c_d + h_a + h_d \]

Here $h_a$ stands for “honest agreements” and $h_d$ for “honest disagreements.”

**Paradox 13:** *Apples compared with oranges.* The numerator in Equations 2 and 9, $(a_r-c_a)$, represents “honest agreements.” The denominator, $(1-c_a)$, represents “expected chance disagreements.” The division, which defines $\pi$ and $\alpha$, compares the numerator as a part with the denominator as the whole to produce a percentage figure. So $\pi$ and $\alpha$ are actually the results of comparing honest agreements with chance disagreements. Why are apples compared with oranges? Why not compare some apples with all apples, e.g. honest agreements with honest coding ($h_a$+$h_d$), or with all coding ($c_a$+$c_d$+$h_a$+$h_d$)?

**Paradox 14:** *Pandas and humans as subgroups of women?* When we divide women by humans, we are asking “what percent of humans are women?” which assumes women are a subgroup of humans. When $\pi$ (Equation 3) and $\alpha$ (Equation 10) divide $d_r$ by $c_d$, they are asking “what percent of chance disagreements are observed disagreements?” They are assuming that all observed disagreements are a subgroup of chance disagreements. But shouldn’t it be the other way around, that chance disagreements be a subgroup of observed disagreements? Why shouldn’t chance disagreements and honest disagreements be the two subgroups of observed disagreements? If so, we should divide chance disagreements by observed disagreements, not vice versa. Were $\pi$ and $\alpha$ built on assumptions like “humans are a subgroup of women?”

From Equation 4 or Equation 5, we can derive $(a_r-c_a)+d_r=c_d$, which indicates that chance disagreements ($c_d$) have two subgroups, "observed disagreements" ($d_r$) discussed above and “honest agreements" $(a_r-c_a)$. How can agreements be a subgroup of disagreements? Were $\pi$ and $\alpha$ built on assumptions like “pandas and humans are the two subgroups of women?”

Any of the paradoxes may invalidate $\pi$ or $\alpha$ as a general indicator of reliability. But $\alpha$ posseses several unique paradoxes, which I present in the next section.
IV. Four More Paradoxes of $\alpha$

**Paradox 15: More of the same quality less reliable?** Suppose two coders code four online messages purported to be news stories, to see if they are actually commentaries in disguise. With an $N=4$, they generated two positive agreements, one negative agreement, and one disagreement. This means 37.5\&62.5\% distribution, a decent agreement rate $a_r=75\%$, and Krippendorff's $\alpha=0.5333$, which looks encouraging, especially given the tiny $N$.

The researcher instructs the coders to expand their work 10 fold by coding 36 more online messages. With a bigger $N=40$, the two coders replicated and multiplied almost everything by ten: 20 positive agreements, 10 negative agreements, and 10 disagreements, leading to another 37.5\&62.5\% distribution and another $a_r=75\%$. The only exception is Krippendorff's $\alpha$, which drops to 0.4733. This is a 11.25\% drop in $\alpha$ for 10 times of repetition of the same quality!

Reliability is often understood as replicability. How can $\alpha$ be a general indicator of reliability when it punishes replicability? The same paradox does not exist for $\pi$ or $\kappa$. Both remain the same with $N=4$ or $N=40$.

Two examples by Krippendorff (1980, pp. 133-135; 2007, pp. 2-3) can be adapted to illustrate the same phenomenon. In both examples, $N=10$, distribution is 70\&30\%, $a_r=.6$, $\alpha=0.09524$. If $N$ increases to 100 while distribution and $a_r$ remain the same, one might expect $\alpha$ to improve or at least remain the same. Instead, $\alpha$ drops to 0.05238.

**Paradox 16: More of better quality less reliable?** Extending the above example, suppose the coders expand the work to code a total 4,000 messages, generating 2,004 positive agreements, 1,004 negative agreements, and 992 disagreements. With $N$ exploding 100 times from 40 to 4,000, distribution remains at 37.5\&62.5\%, while $a_r$ improves slightly, from 75\% to 75.2\%. One might expect $\alpha$ also improves from 0.473. Instead, it drops further, to below 0.471

The same paradox does not exist for $\pi$ or $\kappa$, both of which improve as expected.
Paradox 17: The bar punishes replicability. The phenomena in the above two paradoxes are not isolated. They may be due to the paradoxical behavior of Krippendorff’s \( c_a \). Table 1b and Figure 4 show that Krippendorff’s \( c_a \) is positively correlated with \( N \): given distribution, bigger \( N \) leads to higher \( c_a \). \( c_a \) is the bar that \( a_r \) must pass to produce a good looking \( \alpha \) (cf. Paradox 7). Higher \( c_a \) means less chance for \( \alpha \) to look good. But why? Bigger \( N \) means more cases coded hence higher replicability. How can \( \alpha \) be a general indicator of reliability when it systematically punishes replicability?

Table 2 and Figure 4 about here

Paradox 18: Totally random coding not totally unreliable? Suppose two coders code four cases completely by flipping coins. The coins behave exactly as probability theory says most likely to happen -- head-head, tail-tail, head-tail, and tail-head, with \( a_r=0.5, N=4 \). As one might expect, most of the reliability indicators, including Scott's \( \pi \) and Cohen's \( \kappa \), are exactly 0.00. Krippendorff’s \( \alpha \), however, stands out at 0.125. It's not a spectacular number. But still much higher than zero. Why? How can a completely random result from a completely random process be anything but totally unreliable?

Further, this \( \alpha=0.125 \), from \( a_r=.5, N=4 \) and a totally random coding, is better than \( \alpha=0.095 \) from two Krippendorff examples, each having \( a_r=.6, N=10 \) and honest coding (Krippendorff, 1980, pp. 133-135; 2007, pp. 2-3). Again, how can more and better agreement be less reliable?

These additional paradoxes are additional evidences that \( \alpha \) cannot be a general indicator of reliability. Scott’s \( \pi \) and Krippendorff’s \( \alpha \) might be useful only within a certain boundaries, beyond which the paradoxes would arise. The following sections will define these boundaries, and test their validity by applying them to resolve the paradoxes.
V. Assumptions and Implications

To explain chance agreement, methodologists (Krippendorff, 1980, pp. 133-134; Riffe et al., 1998, pp. 129-130) talked about two coders drawing from urns with black and white marbles. If both draw black, they would agree the object is positive; if both draw white, they would agree the object is negative; in both cases without looking at the object being coded. That’s the chance agreements that we calculate using Equation 8 or Equation 12, and remove using Equation 2, Equation 3, Equation 9, or Equation 10. Of course here “marbles” are any physical or electronic objects of equal probability, and “urns” are any real or virtual containers.

These sounded reasonable, therefore widely accepted, as they were told as hypothetical stories. Few authors or readers believe random drawing takes place regularly in actual research. As Riffe, Lacy, & Fico (2005, p.151) pointed 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.”

The mathematics of Scott’s π and Krippendorff’s α, however, treats marble drawing as real, therefore entails strong assumptions about the coders’ behavior and intent:

**Assumption 1: Conspired quota.** The two numerators in Equation 8 and Equation 12 are supposed to be “marble distribution” in the urn. As marble distribution is not known, Scott (1955) and Krippendorff (1980) inserted the average of the “observed distributions” reported by the two coders. This is a key to understand π and α: they mathematically equates marble distribution with observed distribution.

The objective of a research, however, is to gauge “target distribution” among objects under coding. Coders may draw from an urn of 40&60% marbles while coding a pile of 90&10% commercials. If a research is done reasonably well, its observed distributions should be related to the target distribution; and if it is done honestly, the observed distributions should not always equal marble distributions.

A marble distribution must be set before drawing, which has to take place before a coding that will produce an observed distribution. There is only one way a marble distribution could equal an observed
distribution regularly and precisely—someone preset a quota that’s accurately executed in both marble placing and coding. While ordinary marble drawing contains sampling errors, Equation 8 and Equation 12 leave no room for error, implying that π and α assume a strict quota — the two coders execute the quota so faithfully that the observed average distribution they report is identical to the marble distribution in the urn.

Equation 8 and Equation 12 use the average of two coders’ observed distributions, implying that the two coders set one quota, share one urn, and work together to assure that the average they report meets the quota, hence “conspired quota,” or “collectively strict quota.”

Assumption 2: Maximum random. By removing chance agreement (Equation 2 and Equation 9), π and α assume that deliberate chance coding was not hypothetical, but real — no empirical research should “remove” or “correct for” anything that’s not real. Further, by requiring marble drawing to precede and confine honest coding (again Equation 2 and Equation 9), π and κ assume that coders always maximized random coding under the quota(s) described in Assumption 1.

Assumption 2a for π: Maximum random with replacement: Equation 8, which is for π, multiplies positive rate \((M/N)\) by itself, and multiplies negative rate \((W/N)\) by itself, implying that every time a coder draws a marble from the urn, he or she puts it back into the urn before next drawing, hence "maximum random with replacement."

Assumption 2b for α: Maximum random without replacement. Equation 12, which is for α, subtracts 1 from the second multiplier's numerators and denominators \((M-1, W-1, \text{ and } N-1)\), implying that coders do not put back the drawn marbles before next drawing, hence "maximum random without replacement." While this is the only mathematical difference between α and π, it has other consequences.

When drawing without replacement, the number of marbles \((N_m)\) becomes important. If two coders draw from two marbles, half black and half white, without replacement, the probability of their getting the same color is zero. If they draw from four marbles, again half black and half white, the probability rises to nearly 0.167. If \(N_m\) rises further, the probability rises further; and it approaches 0.5 as \(N_m\) approaches infinity. Therefore α assumes a fixed \(N_m\) within a study.
But $N_m$ is never known. In actual research most of us never requested or detected marble drawing by coders. Even if there is a drawing, the number of marbles could be anything between one and positive infinity. So Krippendorff's $\alpha$ makes another crucial assumption: number of marbles is twice the number of objects under coding, because each drawing by two coders takes two marbles, and supposed by covers one object to be coded:

\textit{Equation 14} \hspace{1em} N_m = 2N

As $N_m$ and $N$ are thus linked, $N_m$ being fixed means $N$ is also fixed. Suppose two coders code 20 cases, calculate the reliability, add 80 more cases, and then calculate the reliability of the 100 cases combined. Krippendorff's $\alpha$ assumes such simple "adding" cannot happen; instead, the only possibility is: (a) the coders treat the coding of 20 and the coding of 100 as two separate studies, (b) the coders put 20 marbles in an urn and draw from the 20 without replacement, to guide their coding of the 20 cases, and (c) the coders put another 100 marbles into another urn, to guide their coding of the 100 cases, including re-coding the 20 cases already coded in the "separate" study.

By contrast, Scott's $\pi$ or Cohen's $\kappa$ assume replacement. Consequently the number of marbles ($N_m$) is not as important. If two coders draw from an equal number of black and white marbles, with replacement, the probability of their getting the same color is 50% regardless of $N_m$. So $\pi$ or $\kappa$ does not assume a fixed $N_m$ or $N$ within a study, although, they also link $N_m$ to $N$ just like $\&$ does.

\textbf{Assumption 3: Limited honesty.} By estimating reliability using Equation 2 or Equation 9, $\pi$ and $\alpha$ assume that honest coding takes place, but is confined to times when marbles' colors mismatch. Scott’s $\pi$ assumes that the probability of the mismatch ranges from 0%, when all marbles in the urn have the same color, to 50%, when the two colors are half and half (See Table 1a and Figure 1a). Krippendorff’s $\alpha$ assumes that the probability ranges from 0% to 100% depending on the distribution and the number of cases coded ($N_m$). It's like a special traffic light — you cannot go on green; you cannot go on red; you can go on green-red flashing, which is on half of the time sometimes, and never on sometimes.
When to Use Scott’s $\pi$ or Krippendorff's $\alpha$, If Ever?

The three assumptions together portray the following Scott Scenario:

1a. Two coders set a quota for the black and white marbles, and fill the urn accordingly (Assumption 1).
1b. They take an object to be coded.
1c. One coder draws a marble randomly from the urn, notes marble’s color, and puts it back into the urn.
1d. The other coder draws a marble randomly from the same urn, notes marble’s color, and puts it back into the urn.
1e. If both draw black, each reports positive; if both draw white, each reports negative; in either case they do not look at the object being coded (Assumption 2). Only when one draws a black and the other draws a white will they code objectively, at which point they may honestly agree or disagree, and report accordingly (Assumption 3).
1f. The two coders calculate the average of positive cases and the average of negative cases they’ve reported. If one of the two averages has reached the predetermined quota, they stop random drawing and coding, and report the remaining objects in such a way that both averages meet the quota (Assumption 1). If neither average has reached the quota, they go to Step 1g.
1g. The coders repeat Step 1b and the subsequent steps until all objects are "coded."

The three assumptions also portray the following Krippendorff Scenario:

2a. Two coders set a quota for the black and white marbles, decide the total number of marbles ($N_m$), and fill the urn accordingly (Assumptions 1 and 2).
2b. They take an object to be coded (Assumption 2b).
2c. One coder draws a marble randomly from the urn, notes marble’s color, and puts it aside without placing it back into the urn (Assumption 2b).

2d. The other coder draws another marble randomly from the same urn, notes marble’s color, and puts it aside without placing it back into the urn (Assumption 2b).

2e. If both draw black, each reports positive; if both draw white, each reports negative; in either case they do not look at the object being coded (Assumption 2b). Only if one draws a black and the other draws a white will they code the object objectively, at which point they may honestly agree or disagree, and report accordingly (Assumption 3).

2f. The two coders calculate the average of positive cases and the average of negative cases they’ve reported. If one of the two averages has reached the predetermined quota, they stop random drawing and coding, and report the remaining objects in such a way that both averages meet the quota (Assumption 1). If neither average has reached the quota, they go to Step 2g.

2g. The coders repeat Step 2b and the subsequent steps until all objects are "coded."

The three assumptions have four implications for both $\pi$ and $\alpha$.

**Implication 1:** *Constrained task.* Assumptions 1 and 3 imply that the purpose of a study is not to find out how many objects are in one category or another, which has been pre-decided by the quota. The only purpose is to place objects into appropriate categories under the quota.

**Implication 2:** *Fixed distributions.* These assumptions also imply that the observed distribution do not change within a study when the coders improve their work, as their “work” is not to assess distribution between categories, but only to assign objects to categories according to the quotas.

**Implication 3:** *Variable benchmark.* Assumptions 1 and 2 imply that chance agreement $c_a$ is a function of the marble distributions predetermined by the quota. When almost all marbles are of the same color, the coders have close to 100% probability agreeing by chance, and close to 0% opportunities to code
honestly. Hence the bar, $c_a$, should be set close to 100%. At the higher extreme, when the quota sets all marbles to the same color, the coders have no chance to do any honest coding, therefore $c_a = 100\%$, and $\pi$ and $\alpha$ are both undefined. Although $\alpha_r$ would be 100%, all of it is deemed due to chance.

The lower extreme is different between $\pi$ and $\alpha$. For $\pi$, when marbles are half black and half white, the coders have 50% chance agreeing by chance and 50% chance of coding honestly, therefore the bar is set at 50%, which is the lowest $c_a$ under $\pi$. Table 1a and Figures 1a, 2 & 3 display the entirety of the changing benchmark for $\pi$ from various angles in various details.

For $\alpha$, the lower extreme varies according to the number of cases ($N$) coded. When $N=2$, Krippendorff’s $c_a$ is either 0% (when distribution is 50&50%) or 100% (when distribution is 0&100%). As $N$ increases from 2 to approach infinity, this lower end of the bar increases from 0%; to approach 50%. Figure 1b, 2, 3, & 4 and Tables 1b & 2 display the changing pattern from different angles with various details.

**Implication 4:** Specified random: The differences between Assumptions 2a and 2b imply that $\pi$ and $\alpha$ each recognizes only one specific type of random as the random. For example, $\alpha$ recognizes only “drawing from a shared urn without replacement.” When another random process, e.g., flipping coins, generates more agreements, $\alpha$ attributes the difference to honest coding, but not to another type of random. This is the source of Paradox 18, which we will explain further.

Scott’s $\pi$ recognizes only drawing from a shared urn *with replacement*. Cohen’s $\kappa$ recognizes only drawing from separate urns with replacement. While both assume “specified random” just like $\alpha$ does, Paradox 18 does arise for or $\pi$ or $\kappa$, because drawing from 50 & 50% marbles with replacement generates the same probability of agreements as coin flipping.

There is an additional implication that applies only to $\alpha$ but not to $\pi$:

**Implication 5 for $\alpha$: Fixed sample.** Coders are assumed to fix the number of cases ($N$) within a given study. This is because $\alpha$ assumes no replacement, hence $N$ affects the probability calculation, as explained in
Assumption 2b. This implication does not apply to Scott’s π or Cohen’s κ because both assume replacement, hence N does not affect probability.

If a study is conducted in a fashion described by the three assumptions and the four or five implications, π or α would be an appropriate indicator of reliability. If the study is not done this way, neither π nor α is appropriate.

VI. Resolving Paradoxes

How would these boundaries explain the 18 paradoxes?

**Paradox 1:** High agreement, low π and α. We found this paradoxical because we assumed the coders coded honestly. Scott's π and Krippendorff’s α assume differently. According to Assumption 1 and Implication 3, all of the observed agreement (ar=99.8%) is due to chance because each coder drew from 999 or 998 black marbles and one or two white marbles. The marbles show different colors only twice, which were the only opportunities to do honest coding. The coders disagreed both times. Low π and α are right.

**Paradoxes 2, 4, 5, 6, and 10:** π and α move erratically. In each of the five situations, the coders coded honestly without quota. Distributions changed as the coders improved their work. Scott's π and Krippendorff’s α, however, assume the coders apply strict quotas (Assumption 1), therefore distributions do not change with improved work (Implication 2). Scott's π and Krippendorff’s α appear paradoxical because they were used beyond their boundaries. Again, there is no paradox.

**Paradox 3:** π and α cannot be calculated. We thought a perfect agreement should be credited with a decent π or α, because we assumed at least some of the agreements came from honest coding, whether the distribution is 0 & 100%, 50 & 50%, or anything in between. Scott's π and Krippendorff’s α, however, assume that when distribution is 0 & 100%, the coders do not have an opportunity to be honest (Implication 3), therefore all observed agreements are due to chance, therefore π or α should not be calculated.
In defense of $\pi$ and $\alpha$ being undefined with 0 & 100% distribution, 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.”

For those of us who assume that most of the coders code honestly most of the time, this explanation could be as puzzling as the paradox. We usually understand “population” as the target population, such as cigarette advertisements we study. Suppose:

- 100% of the population and 100% of the sample we drew have the Surgeon General’s warning
- we know there was no broken instrument, coders (ourselves) did not fall asleep or agree on an outcome in advance; there is only honest and diligent coding
- the two coders agree 100% that 100% of the advertisements have the Surgeon General’s warning.

Why is this not “evidence” for “reproducibility of a population of data”? Isn’t the “evidence that it could be otherwise” obvious, that the two coders could have disagreed any number of times, or agreed that any number of the ads don’t have the warning? Why should a reliability index invariably require “variation” in observation when there may not be variation in the population?

Now that we know $\pi$ or $\alpha$ is to be used only under conditions of strict quota and maximum random, Krippendorff’s (2004b) explanation appears more sensible. The “population” would be better understood as the urn population of marbles, not the target population of advertisements. Under the strict quota and maximum random assumptions, “no variation” in observed results is sufficient “evidence” for no variation in the population of marbles. There is indeed “agreeing in advance” — the coders agree in advance to make the marbles all black, and agree in advance to code honestly only when the marbles’ colors mismatch. There is no chance for mismatch, therefore no chance for any honest coding. Krippendorff was right to say there is no “evidence that it (the observation) could be otherwise. … hence no evidence of reliability.” In other words,
Krippendorff’s explanation as I interpreted it could be seen as another corroborating evidence that $\pi$ and $\alpha$ assume strict quota and maximum random.

**Paradox 7:** *Erratically moving bar.* We found it paradoxical because we didn’t think the bar, as a part of the general indicator applied to typical studies, should move with distribution. But $\pi$ and $\alpha$ are not general indicators. They are to be used only when Assumptions 1–3 are all met. Under these assumptions, especially those derived from Implication 3, the bar should move.

**Paradox 8:** *Punishing Agreement.* We found this paradoxical because we compared $c_a$, hence $\pi$ and $\alpha$, across different distributions. Under Assumptions 1 and 3 and Implications 1 and 2, $\pi$ and $\alpha$ would reward higher agreement, but only within a fixed distribution. When distributions change, higher observed agreement is deemed from higher *chance* agreement, hence less opportunities for honest coding, which $\pi$ and $\alpha$ should punish according to Implication 3.

**Paradox 9:** *Honest coding is as bad as coin flipping.* We found the negative $\pi$ and $\alpha$ paradoxical in contrast to the nearly 67% agreement rate, because we knew the coders coded honestly. Assumptions 2 and 3, however, stipulate that coders started by drawing marbles. Both $\pi$ and $\alpha$, based on the 83% average distribution, presumes the coders put 50 black marbles and 10 white marbles into an urn, and drew randomly from the urn (Assumptions 1 and 2). Without a single glance at the objects, they should have generated a 72% agreement rate, much higher than the 67% they actually got, leading to a perfectly justifiable $\pi=-.2$.

**Paradox 11:** *Circular logic.* We found it circular because we thought the reported distributions embedded in $\pi$ or $\alpha$ came from coders’ observations. We were wrong. The distributions came from pre-determined quotas independent of the observations, according to Assumption 1 and Implication 1. The logic is not circular.

**Paradox 12:** *Nothing but chance.* We found Equation 5 paradoxical because we believe coders coded objectively, *before and beyond* chance. Assumptions 2 and 3, however, stipulate that coders maximize chance coding under quotas, and coders code honestly only when marbles instruct them to. “Nothing but chance” is an operating boundary for $\pi$ and $\alpha$, within which Equation 5 is fine.
**Paradox 13**: *Apples compared with oranges.* We found it paradoxical because we did not realize that, under Assumptions 2 and 3, chance disagreement is honest coding ($c_{e} = h_{o} + h_{d}$) — coders code honestly when and only when marbles disagree.

**Paradox 14**: *Humans a subgroup of men?* We found Equation 2 and Equation 9 paradoxical because we thought chance disagreement is a subgroup of observed disagreement. Under Assumption 2, however, coders agree whenever marbles agree (chance agreement); they code honestly (honest coding) only when marbles disagree (chance disagreement), during which they may agree (honest agreement) or disagree (observed disagreement=honest disagreement). Therefore observed disagreement is a subgroup of chance disagreement. No paradox.

**Paradoxes 15 and 16 for $\alpha$**: *More of the same quality or better quality less reliable.* These cases appear paradoxical only because we assumed these are normal studies in which researchers pretest with a few cases, calculate its reliability, expand to a larger number, and test reliability again, all in full honesty. In this "honest coding" scenario, more of the same quality deserve no punishments, and more of the better quality deserve rewards. Krippendorff's $\alpha$ assumes differently. Coders maximize random coding in a particular way, that is, drawing marbles with no replacement (Assumption 2b). So the number of cases and number marbles are fixed for a given study (Implication 5). The larger study takes more marbles than the pre-test. More marbles produce more chance agreements, which has to be punished.

**Paradox 17**: *The bar punishes replicability.* We called it "replicability" because, when we see a larger $N$, we see more of the honest coding. We could not understand why the bar should be raised. When $\alpha$ sees a larger $N$, it sees more marbles, which mean more deliberate random (Assumption 2b and Implication 5), hence more random agreement, for which the bar must be raised.

**Paradox 18**: *Totally random coding not totally unreliable?* We thought flipping coins is being "random." In Krippendorff's $\alpha$, however, only thing that can be called random is drawing from a shared urn with no replacement (Assumption 2b and Implication 4). Coin flipping does not qualify, because it is with replacement. We know we actually did nothing but flipping coins, and it generated more agreement than
When to Use Scott’s $\pi$ or Krippendorff’s $\alpha$, If Ever?

marble drawing, just like probability theory predicts. We disregard the difference as “another kind of random.” Krippendorff’s $\alpha$ credits the difference to honest coding, and reward us with a higher $\alpha$.

Our coin flipping generated higher $\alpha$ than Krippendorff’s (1980, 2007) honest coding because (a) $\alpha$ assumes some of our coin flipping is honest coding, (b) Krippendorff’s examples have larger $N$ than our coin flipping, and $\alpha$ assumes larger $N$ generate more chance agreements, which has to be “corrected for,” as we explained above.

VII. Discussion: Source of Confusions

Some reliability methodologists talked about chance agreement ($a_c$) as what *would have been* obtained from a completely random coding, indicating that they saw maximum random only as a reference for comparison, not what really happens in typical research.

Following this thinking, each methodologist could have selected several hypothetical scenarios, such as flipping coins or throwing dices, drawing marbles of 60&40% or 90&10% distribution, from one or two urns, with or without replacement. Each scenario can produce a chance agreement to be compared with the observed agreement. Had the methodologists done that, we would not have had the various indicators of reliabilities, nor would we have had the paradoxes. We would have many hypothetical chance agreements to be compared with observed agreement. Because there is an unlimited number of ways we could practice “random,” (at least the possible distribution, which is a continuum, is unlimited), there could be an infinite number of chance agreements as references for comparison.

The methodologists did not do that. Instead, they recommended equations like Equation 2, Equation 3, Equation 9 or Equation 10 to “remove” or “correct for” chance agreement. Each of them chose a different scenario of maximum random, yet each believed he was recommending *a or the* general indicator of reliability. This created a gap between the conceptual understanding, which sees maximum random as hypothetical, and the actual computation, which treats maximum random as real. This gap is a major source
of our confusions and paradoxes. We should close this gap by stop using \( \pi \) or \( \alpha \) as a general indicator of reliability.

VIII. Conclusion

The eight assumptions and implications can resolve the 18 paradoxes, supporting our speculation that \( \pi \) or \( \alpha \) does not measure general reliability.

Scott's \( \pi \) measures reliability under special conditions when coders (1a) enforce conspired quota, (1b) exercise maximum random with replacement, and (1c) practice limited honesty. These assumptions imply that coders (1d) perform constrained task, (1e) require fixed distribution, (1f) entail variable benchmark, and (1g) conduct specified random.

Krippendorff’s \( \alpha \) measures reliability under special conditions when coders (2a) enforce conspired quota, (2b) exercise maximum random without replacement, and (2c) practice limited honesty. These assumptions imply that coders (2d) perform constrained task, (2e) require fixed distribution, (2f) entail variable benchmark, (2g) conduct specified random, and (2h) fix the number of cases.

The main differences between \( \pi \) and \( \alpha \) are in Conditions (b) and (h) — \( \pi \) assumes replacement and no fixed sample, while \( \alpha \) assumes fixed sample and no replacement. Both differ from Cohen’s \( \kappa \) in one way: \( \pi \) and \( \alpha \) assume a collective quota, while \( \kappa \) assumes individual quotas.

When any of the seven conditions (1a-1g) does not hold, \( \pi \) doesn’t apply. When any of the eight conditions (2a-2h) does not hold, \( \alpha \) does not apply.

It’s hard to imagine a circumstance under which the seven or the eight conditions all hold. And even if they do hold, the first option is to discard the data and redo the study honestly. Only in an extraordinary situation, when the uncontaminated part of the data is too precious to discard and too difficult to re-collect, should we apply \( \pi \) or \( \alpha \) as a part of rescue and repair. We should use Scott’s \( \pi \) or Krippendorff’s \( \alpha \) rarely, and hopefully never.
Future studies should ask: Are there similar boundaries on other indicators of reliability? What are the boundaries? Is there a true general indicator? Which one?
References


When to Use Scott’s π or Krippendorff’s α, If Ever?


Table 1a: Scott’s Chance Agreement ($c_a$) as a Function of Two Distributions*

<table>
<thead>
<tr>
<th>Distribution $j$: Positive Findings by Coder $j$ ($M_j/N$)**</th>
<th>Distribution $i$: Percent of Positive Findings by Coder A ($M_i/N$)**</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>0</td>
</tr>
<tr>
<td>100</td>
<td>50.0</td>
</tr>
<tr>
<td>90</td>
<td>50.5</td>
</tr>
<tr>
<td>80</td>
<td>52.0</td>
</tr>
<tr>
<td>70</td>
<td>54.5</td>
</tr>
<tr>
<td>60</td>
<td>58.0</td>
</tr>
<tr>
<td>50</td>
<td>62.5</td>
</tr>
<tr>
<td>40</td>
<td>68.0</td>
</tr>
<tr>
<td>30</td>
<td>74.5</td>
</tr>
<tr>
<td>20</td>
<td>82.0</td>
</tr>
<tr>
<td>10</td>
<td>90.5</td>
</tr>
<tr>
<td>0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Table 1b: Krippendorff’s Chance Agreement ($c_a$) as a Function of Two Distributions with One Hundred Cases ($N=100$)*

<table>
<thead>
<tr>
<th>Distribution $j$: Positive Findings by Coder $j$ ($M_j/N$)**</th>
<th>Distribution $i$: Percent of Positive Findings by Coder A ($M_i/N$)**</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>0</td>
</tr>
<tr>
<td>100</td>
<td>49.7</td>
</tr>
<tr>
<td>90</td>
<td>50.3</td>
</tr>
<tr>
<td>80</td>
<td>51.8</td>
</tr>
<tr>
<td>70</td>
<td>54.3</td>
</tr>
<tr>
<td>60</td>
<td>57.8</td>
</tr>
<tr>
<td>50</td>
<td>62.3</td>
</tr>
<tr>
<td>40</td>
<td>67.8</td>
</tr>
<tr>
<td>30</td>
<td>74.4</td>
</tr>
<tr>
<td>20</td>
<td>81.9</td>
</tr>
<tr>
<td>10</td>
<td>90.5</td>
</tr>
<tr>
<td>0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

*: Main cell entries are Scott’s Chance Agreement ($c_a$) in %.

**: $M_i$ is the number of positive answers by Coder $i$, $M_j$ is the number of positive answers by Coder $j$, and $N$ is the total number of cases analyzed. See also Equation 6, Equation 7, Equation 8, & Equation 12.
Figure 1a: Scott's Chance Agreement ($c_a$) as a Function of Two Distributions
When to Use Scott’s $\pi$ or Krippendorff’s $\alpha$, If Ever?

Figure 1b: Krippendorff's Chance Agreement ($c_a$) as a Function of Two Distributions ($N=100$)
Figure 2: Krippendorff's Chance Agreement ($c_a$) as a Function of Average Distribution ($M/N$) with One Hundred Cases ($N=100$)

(The curve for Scott's Chance Agreement ($c_a$) look the same.)
When to Use Scott’s $\pi$ or Krippendorff’s $\alpha$, If Ever?

Figure 3: Krippendorff’s Chance Agreement ($c_a$) as a Function of Distribution $i$ ($M_i/N$) Given Distribution $j$ ($M_j/N$)

(The curve for Scott’s Chance Agreement ($c_a$) look the same.)

- $M_j/N=0\%$
- $M_j/N=100\%$
- $M_j/N=25\%$
- $M_j/N=75\%$
- $M_j/N=50\%$

Distribution $i$ ($M_i/N$) -- Percent of Positive Answers by Coder $i$
Table 2: Krippendorff's Chance Agreement Rate ($c_a$) as a Function of Number of Cases Coded ($N$) and Average Distribution ($M/N$)

<table>
<thead>
<tr>
<th>Number of Cases Coded ($N$)</th>
<th>Average Distribution of Positive Cases ($M/N$ in %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>100.00, 64.00, 36.00, 16.00, 4.00, 0.00, 4.00, 16.00, 36.00, 64.00, 100.00</td>
</tr>
<tr>
<td>10</td>
<td>81.67, 81.63, 81.60, 81.58, 81.54, 81.49, 81.33, 80.80, 80.00, 78.40</td>
</tr>
<tr>
<td>20</td>
<td>81.68, 81.65, 81.62, 81.60, 81.57, 81.54, 81.49, 81.42, 81.40, 81.38</td>
</tr>
<tr>
<td>30</td>
<td>81.69, 81.66, 81.63, 81.60, 81.57, 81.54, 81.51, 81.46, 81.44, 81.42</td>
</tr>
<tr>
<td>40</td>
<td>81.70, 81.67, 81.64, 81.61, 81.58, 81.55, 81.52, 81.47, 81.45, 81.43</td>
</tr>
<tr>
<td>50</td>
<td>81.71, 81.68, 81.65, 81.62, 81.59, 81.56, 81.53, 81.48, 81.46, 81.44</td>
</tr>
<tr>
<td>60</td>
<td>81.72, 81.69, 81.66, 81.63, 81.60, 81.57, 81.54, 81.49, 81.47, 81.45</td>
</tr>
<tr>
<td>70</td>
<td>81.73, 81.70, 81.67, 81.64, 81.61, 81.58, 81.55, 81.51, 81.47, 81.45</td>
</tr>
<tr>
<td>80</td>
<td>81.74, 81.71, 81.68, 81.65, 81.62, 81.59, 81.56, 81.53, 81.49, 81.47</td>
</tr>
<tr>
<td>90</td>
<td>81.75, 81.72, 81.69, 81.66, 81.63, 81.60, 81.57, 81.54, 81.51, 81.49</td>
</tr>
<tr>
<td>100</td>
<td>81.76, 81.73, 81.70, 81.67, 81.64, 81.61, 81.58, 81.55, 81.53, 81.51</td>
</tr>
</tbody>
</table>
Figure 4: Krippendorff's Chance Agreement Rate ($c_a$) as a Function of Number of Cases Coded ($N$) and Average Distribution ($M/N$)