Hong Kong Baptist University
HKBU Institutional Repository

HKBU Staff Publication

11-2018

We agreed to measure agreement - Redefining reliability de-justifies Krippendorff's alpha

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

Guangchao Charles Feng
Shenzhen University

Jun S. Liu
Harvard University

Ke Deng
Tsinghua University

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

Part of the Communication Commons

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

APA Citation

ZHAO, X., Feng, G., Liu, J., & Ke Deng. (2018). We agreed to measure agreement - Redefining reliability de-justifies Krippendorff's alpha. *China Media Research, 14* (2), 1-15. Retrieved from https://repository.hkbu.edu.hk/hkbu_staff_publication/6803

This Journal Article is brought to you for free and open access by HKBU Institutional Repository. It has been accepted for inclusion in HKBU Staff Publication by an authorized administrator of HKBU Institutional Repository. For more information, please contact repository@hkbu.edu.hk.

We Agreed to Measure Agreement – Redefining Reliability De-justifies Krippendorff's Alpha

Xinshu Zhao, Hong Kong Baptist University / University of North Carolina at Chapel Hill
Guangchao Charles Feng, Shenzhen University
Jun S. Liu, Harvard University
Ke Deng, Tsinghua University

Abstract: Zhao, Liu, & Deng (2013) reviewed 22 inter-coder reliability indices, and found that each makes unrealistic assumption(s) about coder behavior, leading to paradoxes and abnormalities. Krippendorff's α makes more of such assumptions, consequently produces more paradoxes and abnormalities than any other index.

Professor Krippendorff (2013) countered that "most of the authors' discoveries are the artifacts of being led astray by strange, almost conspiratorial uses of language." The commentary reiterated Krippendorff's long-standing position that Krippendorff's α is the standard reliability measure, and is the only index qualified to serve the function (Hayes & Krippendorff, 2007; Krippendorff, 2004b, 2016).

This paper continues this dialogue. We offer a review of literature to show that the scientific community, including Krippendorff, has defined intercoder reliability as intercoder agreement, and Krippendorff's α, like all its main competitors, was designed and declared to measure intercoder agreement. Now that evidences are mounting that α, like Scott's π and Cohen's κ, does not accurately measure intercoder agreement, Krippendorff chose to redefine intercoder reliability and, furthermore, redefine information, variation, sensitivity, and specificity.

By redefining reliability, we argue, Prof. Krippendorff has redefined the function of Krippendorff's α, thereby disqualified α as an indicator of intercoder agreement. The search for a better index of intercoder agreement aka intercoder reliability should continue.

We, however, also note a spiral of inertia in science communication in general, and reliability research in particular. The powerful spiral, we argue, should not forever keep up the appearances for α, π or κ.

[Xinshu Zhao, Guangchao Charles Feng, Jun S. Liu and Ke Deng (2018). **We Agreed to Measure Agreement – Redefining Reliability De-justifies Krippendorff's Alpha.** *China Media Research, 14*(2):1-15

Keywords: spiral of inertia, selective spiral, reliability, inter-coder reliability, inter-rater reliability, agreement, Cohen's kappa, Scott's pi, Krippendorff's alpha, multi-signification, multi-concepts, multi-signified, multi-signs, multi-signifiers, sensitivity, specificity, mechanical information, human information, aggregate estimation, individual classification, individual prediction.

More than a century after Benini (1901) documented percent agreement (a_o) and introduced β, two of the earliest known indices of intercoder reliability, new indices continue to emerge (e.g. Cousineau & Laurencelle, 2016; Kirilenko & Stepchenkova, 2016), and reliability experts continue to debate whether any indices are legitimate, or which ones are more legitimate (Conger, 2016; Feng, 2015; Feng & Zhao, 2016; Flight & Julious, 2015; Grant, Button, & Snook, 2017; Lombard, Snyder-Duch, & Bracken, 2002, 2004; Krippendorff, 2004b, 2016; Shankar & Bangdiwala, 2014; Xu & Lorber, 2014).

Scores of indices are available. Popping (1988) identified no less than 39. We analyzed the behavioral assumptions of 23 (Zhao et al, 2012, 2013). Grant et al. (2017) simulated the performance of five. Most indices were advertised as *the* index, yet they are often drastically different from each other. While Cohen's κ (1960) has been by far the most popular chance-adjusted index across disciplines, it is also the most often debated, due to the numerous paradoxes and abnormalities that it produces (Bakeman, 2000; Bloch & Kraemer, 1989; Brennan & Prediger, 1981; Dewey, 1983; Feinstein & Cicchetti, 1990a, 1990b; Feuerman & Miller, 2008; Grove, Andreasen, McDonald-Scott, Keller, & Shapiro, 1981; Gwet, 2008, 2010; Kraemer, 1979; Kraemer & Bloch, 1988; Kraemer, Periyakoil, & Noda, 2002; Lombard et al., 2002; Perreault & Leigh, 1989; Roberts, 2008; Uebersax, 1987, 2009, Vach, 2005; Williamson, Lipsitz, & Amita K. Manatunga, 2000; Wongpakaran et al., 2013; Zhao, 2011a; Zhao et al., 2013). But every other index has its own advocate(s) and critic(s) (Zhao, 2011b; Zhao et al., 2013).

Some found that the root problem is the indices' assumptions of whimsical coder behaviors. After examining 22 indices, Zhao et al. (2013) reported that each index makes at least one such assumption(s); Krippendorff's α makes more of such assumptions, consequently produces more paradoxes and abnormalities than any other index. By contrast, the supposedly primitive and flawed percent agreement makes fewer and less whimsical assumptions, and

produces fewer paradoxes. Zhao et al. (2013) called for better indices based on more realistic assumptions.

To avoid the indices' most severe symptoms, Zhao et al. (2013, Table 19.13) recommended different indices for different situations. The approach, referred to as the *best available for a situation* (BAFS), is intended to help until a better index becomes available.

To counter, Krippendorff (2013, p. 481, parenthesis added) stated "most of the authors' (Zhao et al., 2013) discoveries are the artifacts of being led astray by strange, almost conspiratorial uses of language." Krippendorff accordingly maintained his longstanding position that Krippendorff's α is the only index qualified to serve as the standard indicator of intercoder reliability.

The disagreements are not just about numbers or formulas, but more about functions that reliability plays in scientific enquiry. For practical purposes it boils down to two questions:

1) Do we need better indices? Some say yes, because all indices examined so far show deficiencies. Krippendorff says no, because α is superb and perfect.

2) Should we use α, and only α, in all situations (α-only) or the Best Available for a Situation (BAFS)? Zhao et al. (2013) recommended BAFS, because different indices have shown different strengths and weaknesses in different situations. Krippendorff (Hayes & Krippendorff, 2007; Krippendorff, 2011; Krippendorff, 2012, 2013, 2016) recommended α only, because it is the only one that's perfect in all situations.

We agree with Krippendorff that all other 21 indices have deficiencies. So the two questions become one: is α perfect?

This article provides a non-mathematical analysis of Krippendorff's defenses of α. The defenses, we will show, implies a concession by Krippendorff that α does not measure intercoder *agreement*, which α was originally designed and declared to measure (Krippendorff, 1970a, 1970b, 1980, 2011b); instead, α measures a mixture of several concepts (Krippendorff, 2011a, 2013, 2016).

We will note that, in all disciplines, the defined mission or stated function of intercoder reliability is to estimate true agreement. For over four decades Krippendorff had been a prominent part of this *agreement on agreement*. We will show that, since κ was found to fail the mission, Krippendorff (2011, 2013) has redefined the mission by (1) redefining reliability as information and (2) redefining information as even distribution, smaller sample, larger variation, and higher sensitivity/specificity. The redefinitions have one function, which is to explain away the paradoxes and abnormalities, so that α continues to appear perfect.

We will argue that the redefinitions should not and probably will not forever keep up α's appearances. We will call for better indices based on more realistic assumptions.

1. Intercoder Reliability = Intercoder Agreement

This section shows that Krippendorff redefined reliability as information, and redefined information as statistical variation, to make α appear useful.

Reliability has always been defined as *consistency*. The Marketing Accountability Standards Board, for example, endorsed a Wikipedia definition: "Reliability in statistics and psychometrics is the overall consistency of a measure. A measure is said to have a high reliability if it produces similar results under consistent conditions." ("Reliability (Statistics)," 2017)

Inter-measure reliability, often gauged by Cronbach's alpha (1951), is consistency between measures. Test-retest reliability is consistency between repeated tests. Intercoder reliability is consistency between coders. The latter two are typically measured by *agreement* indices, so much so that *inter-rater reliability* and *inter-rater agreement* are seen as synonyms (Saal, Downey, & Lahey, 1980). Notable exceptions include Tinsley & Weiss (1975, 2000), who used correlation as the indicator of reliability, and Neuendorf (2002), who considered agreement and covariation as two indicators of reliability. As correlation and covariation are still among the broader consistency concepts, these views do not contradict the general consensus that reliability means consistency.

The near consensus on the sign-concept relationship existed from early on. While Cohen's κ has been by far the most popular index of intercoder reliability across disciplines, Cohen (1960, 1968) himself consistently called it a "coefficient of agreement." Popping (1988), who reviewed 39 reliability indices, called them "agreement indices" in the title. Gwet (2010), Maxwell (1977) and Rogot & Goldberg (1966) also labeled theirs "agreement" indices in the titles of their book and articles.

For more than four decades, Krippendorff was a prominent part of this near consensus. He labeled α an "agreement coefficient" in the title of the article that introduced α (Krippendorff, 1970a). In the opening paragraph of another article, Krippendorff (1970b, p. 61, emphasis by Krippendorff) stated "the reliability of a population of data must be *estimated from the agreement* among many observers regarding a sample." Krippendorff (e.g. 2004b, p. 415) opposed, time and again, defining intercoder reliability as association, correlation or anything else other than agreement. Up to quite recently, Krippendorff continued to define α as "a reliability coefficient developed to measure the agreement ..." (Krippendorff, 2011b, p. 1).

It was against this backdrop that α, π and κ's paradoxes and abnormalities appeared so troublesome to so many, including the high-agreement-low-index phenomenon (Abnormality 10 for α, π & κ, Zhao et al., 2013; also see Feinstein & Cicchetti, 1990a; Feinstein &

Cicchetti, 1990b; Grove et al., 1981; Lombard et al., 2002; Spitznagel & Helzer, 1985). If reliability measures agreement, how could the trio deviate so drastically from agreement? Zhao et al. (2013) showed that α shares most of π and κ's troublesome assumptions, paradoxes, and abnormalities, then adds more of its own.

In response, Krippendorff (2013, pp. 482-484) emphasized the difference between reliability and agreement, and criticized the "one-to-one" relationship between the two. Reliability is no longer consistency, and intercoder reliability is no longer agreement. So what is reliability now, according to Krippendorff?

Recall that π, κ and α are significantly affected by distribution skew. Higher skews produce lower indices, and the highest skews produce undefined indices. With identical agreement rates, a skewer distribution can make the three indices' chance estimates hundreds or even tens of thousands times higher than a more even distribution. Zhao et al. (2013) reported that some of the paradoxes and abnormalities happen because π, κ and α are systematically and negatively affected by skew, and the linkage is caused by unrealistic assumptions, including maximum randomness and predetermined quota.

To deny that α is based on these assumptions, Krippendorff needs a more friendly explanation for the negative skew-α correlation. On a nominal scale, skew is statistically linked with variance/variation – higher skews produce smaller variances, and the highest skew produces zero variance. So if one can justify reliability being positively affected by variance/variation, he can justify the index being negatively affected by skew.

Larger variance or variation, however, does not indicate higher reliability. On the contrary, common sense and classic theories associate larger variance with larger error, therefore *lower* reliability. For example, everything else being equal, a larger variance would lead to smaller correlation (Cohen, 1988, 1992), lower Alpha (Cronbach, 1951), and lower probability for statistical significance (Cohen, 1988, 1992).

To avoid directly and overtly contradicting common sense and well established theories, Krippendorff (2011a; 2013, pp. 484-485, 493) inserted (more) *information* between (larger) *variation* and (higher) *reliability*, asserting that a larger variation provides more information, which means higher reliability, creating a chain of conceptual equations: *(lower) skew* ≈ *(larger) variation* ≈ *(more) information* ≈ *(higher) reliability*, where each "≈" represents a redefinition. The objective is to redefine (higher) reliability as (lower) skew, aka (more even) distribution.

Unfortunately for Krippendorff, two of the three redefinitions fail. First, reliability may not be defined as information, although both are desirable. Research is to create knowledge, which is to provide information. The knowledge needs to be reliable, so it may be valid. Therefore, knowledge/information and validity /reliability are different concepts playing different roles. Information is a goal of research; reliability is a quality of the instrument employed to achieve the goal. Ends should not be confused with means.

Conceptually equating information with reliability brings troublesome questions. If more information is considered more reliable, can coders improve reliability by adding more variables and more data, which would increase information? If a reviewer asks for vital information about a variable that the researcher failed to preserve, can the researcher somehow produce a higher reliability and argue it is equivalent to more information? Our answers are no, and no. Reliability and information are distinct concepts. One cannot replace the other. We should not equate the two just to justify α.

Second, for human communication, information may not be defined in terms of variance, variation or any other purely statistical characteristics, even though computer engineers often do so for their needs.

Engineers and computer scientists use *bit*, a physical characteristic of a disk, cable, or transmitter, to quantify information (Hartley, 1928; Kolmogorov, 1968; Shannon, 1948). They can do so legitimately because they do not have to consider the content and context of human messages.

Communication researchers and other social scientists, however, must consider content and context. Compare two computer files, one occupying 40 kb (thousands bits), the other 4 kb. Does the former contain nine times more information? For computer engineers, probably yes; for social scientists and especially communication researchers, probably no. Depending on the content in the files and their meanings to a reader(s), either file could contain more information. The *bits* tell us little, if anything. The same is true for *variance* or *variation*.

Compare two sentences from news: "Santorum and Romney are tied at 25%" (Condon, 2012) and "Romney edges Santorum by 8 Votes" (Khan, Friedman, & Shushannah, 2012). While one sentence reports a more even distribution, entailing a larger statistical variance / variation, one does not necessarily carry more information for readers, politicians, journalists, or others.

In human communication, a receiver's mind set is also important. For someone who knew nothing and cared nothing about American politics, or someone who had already known the news, the two stories carried no information. The stories were informative only for those who cared but did not know. Krinppendorff's *information* ignores such human contexts.

There are two concepts of information. One is *mechanical information*, which is universally quantifiable by bits, entropy, variance, variation or other physical or mathematical patterns without referring to human context. The other is *human information*, which can only be assessed within particular contexts.

Even if we accept the *variation* ≈ *information* ≈ *reliability* chain of redefinitions, it can at best avert a couple best-known paradoxes or abnormalities, such as Abnormalities 10 and 11, but not the others, such as Abnormalities 12, 14 & 15 (all described in Zhao et al., 2013). In the latter three, π, κ and α changed radically while agreement and Krippendorff-defined information both changed little. Had α truly measured information, it should have at least co-varied with Krippendorff-defined information. The three indices were originally designed to measure agreement, not distribution or information. Each index was designed incorrectly because of unrealistic assumptions. Consequently each measures partially agreement, partially distribution skew, and partially something else.

2. *Reliability ≠ Sensitivity* & *Reliability ≠ Specificity*

This section shows that Krippendorff redefined reliability as sensitivity, to make α appear useful. Krippendorff (2013, pp. 484-485) wrote:

> Let us use the authors' (Zhao et al., 2013) numerical example: Suppose two separate doctors administer the test to the same 1,000 individuals. Suppose each doctor finds one in 1,000 to have the disease and they agree in 998 cases on the outcome of the test. The authors note that Cohen's (1960) κ, Scott's (1955) π, and Krippendorff's α (1980, 2004a, 2012) are all below zero (-.001 or -.0005). They … proclaim that chance-adjusted indices entail the paradox (their abnormality 10) of "high agreement but low reliability" as sure proof of the inadequacy of these coefficients. ... I contend that a test which produces 99.8% negatives, 2% disagreements and not a single case of an agreement on the presence of the disease is totally unreliable indeed. Nobody in her right mind should trust a doctor who would treat patients based on such test results. The inference of zero reliability is perfectly justifiable. The paradox of "high agreement but low reliability" does not characterize any of the reliability indices cited but resides entirely in the authors' conceptual limitations.

This medical example, however, was not authored by us (Zhao et al., 2013). Our example was a content analysis of 1,000 magazine advertisements (Zhao et al., 2013, p. 454). Why did Krippendorff replace our example and critique his own example as if it was ours? A communication scholar, critiquing another communication scholar's communication example in a communication journal for a communication audience, replaced the communication example with a medical example. Why? What's to be gained from this strawman?

Individual precision is more important for medical diagnoses than communication studies. Suppose a procedure finds 55% people having a disease, with 9% of the patients and 11% of the non-patients misdiagnosed. As the two types of misdiagnoses offset each other, the 55% aggregate estimate is accurate. But the errors do not offset for individuals. Given that more than 16% of the patients were not treated and more than 24% of the non-patients were treated, the diagnostic procedure is inadequate despite the aggregate accuracy.

In contrast, when communication researchers found 55% TV programs featuring romantic scenes, they did not estimate false alarms or false negatives, because their aim was to estimate overall distribution, not to classify individual programs (Brown et al. 2013).

The differences in foci are understandable – assuming accurate aggregate estimation, misclassifying individual contents are often non-consequential, while misdiagnosing patients can be detrimental. Furthermore, misdiagnoses of different directions can have different consequences. Failing to diagnose a SARS patient may cause a pandemic, while misdiagnosing a healthy person would mean just a few days of unnecessary hospital stay. Conversely, there may be situations where false alarms entail life-threatening operations and unnecessary loss of organs, while failed detections mean only delayed treatment without long term harm.

Therefore, medical researchers differentiate *sensitivity*, the accuracy in detecting a disease, from *specificity*, the accuracy in deciding the absence of a disease (Altman & Bland, 1994). In other words, a higher sensitivity reduces false negatives, while a higher specificity reduces false positives. Besides *validity*, therefore, a diagnostic instrument's quality is assessed by three indicators, *reliability*, *sensitivity*, and *specificity* (Altman & Bland, 1994; Feng, 2013) .

In contrast, due to our emphasis on aggregate accuracy, communication researchers have rarely differentiated types of misclassifications. Sensitivity or specificity is rarely mentioned in communication studies or classes, while reliability is taught regularly.

Krippendorff focused on the two cases of possible disease while ignoring the 998 cases of agreed non-disease, placing more weight on sensitivity than specificity, in effect redefining α as a sensitivity measure. We call this a *sensitivity defense*. Medical researchers have used the sensitivity defense to explain the phenomenon of "high agreement, low κ" (Hoehler, 2000; Kraemer & Bloch, 1988; Shrout, Spitzer, & Fleiss, 1987). Later, π and α were found to produce the same phenomenon, "high agreement, low π and α" (Lombard et al., 2002; Zhao, 2011b). Zhao et al.'s (2013) communication example illustrated that the root cause was the indices' maximum randomness and quota

assumptions. To refute these explanations, Krippendorff (2013) was to evoke the sensitivity defense.

But Krippendorff faced an obstacle. He was to critique a communication example, where different misclassifications appear equally bad. The sensitivity defense would appear appealing only when false negatives are far more detrimental than false alarms, making sensitivity far more important than specificity, which is often the case in medical studies. So the communication example was switched quietly to a medical example. Strawman stood. Obstacle removed. Sensitivity defense launched.

Krippendorff had another obstacle. There were far more positive cases in Zhao et al.'s (2013, p.454) example, while the sensitivity defense would appear appealing only when positive cases are rare. So Krippendorff made sure there were few positive cases when he reconstructed the target of his criticism.

Even for medical studies, however, sensitivity defense does not really justify κ, π or α as consistent reliability indicators. The indices were designed to measure intercoder/interrater reliability, which assume equal weight for a false positive and a false negative, hence equal importance for sensitivity and specificity. Krippendorff's arguments may at best reclassify the trio as makeshift sensitivity indices in these limited situations, beyond which the trio's behaviors are shifty. They approximate sensitivity measures when positive cases are rare and sensitivity is far more important; they approximate specificity measures when negative cases are rare and specificity is far more important (Hoehler, 2000). They covary with distribution skew when sensitivity and specificity are fixed while distribution varies; they covary with agreement rates when distribution is fixed and approximately even (cf., Vach, 2005).

Reliability has been defined as consistency ("Reliability (Statistics)," 2017). It is ironic that π, κ and α, the most consistently respected indices of consistency, are consistently inconsistent. The inconsistent behavior is, again, due to the indices' unrealistic assumptions of coder behavior, as outlined by Zhao (2011a, 2011b) and Zhao et al. (2013).

Medical researchers would be better served with three families of indices respectively for reliability, sensitivity, and specificity, and π, κ and α might play useful roles after rethinking and remaking. As they are now, however, researchers should be highly cautious, especially in high-risk medical research. But the long-term solution is to develop indices based on more realistic assumptions.

3. Smaller Sample ≠ More Information & Smaller Sample ≠ Higher Reliability

This section shows that Krippendorff redefined *more information* as *smaller sample*, to make a unique defect of α look like a unique virtue. Krippendorff (2013, p. 492) wrote:

> They (Zhao et al., 2013) consider it counterintuitive that under conditions of a constant a_o, α becomes larger when sample sizes become smaller. I am suggesting that this intuition is due to the confusion, discussed in the authors' first conceptual problem, of the role of the amount of information (variation and sample size) and of the agreement coefficients involved, and moreover, not recognizing what π and α differentially count as chance agreement.

What Zhao et al. (2013, p.450, emphasis added) said was "*everything else* being equal, a smaller sample produces a smaller a_c, hence a higher α." "Everything else" includes true reliability and measurement quality in general. Krippendorff in effect further redefined reliability, information, and sample size: everything else being equal, smaller samples have more information, hence are more reliable. If one is skeptical about this extraordinary assertion, he or she has "confusion" and "conceptual problem" (Krippendorff, 2013; p. 486, 492).

That a larger sample contains more information than a smaller sample, everything else being equal, is a basic principle behind many procedures of scientific sampling and data analysis. Krippendorff turned the principle upside down in the effort to make α different from π. By the time this uncanny feature was discovered, it was probably too late and too difficult to remove it from α. Krippendorff chose to defend the feature by redefining the fundamental concepts. Is α so precious, that we must reject or reverse every principle and every common sense in its way?

Let's explain again why sample size affects α but not π, and why smaller samples lead to higher α. Both indices assume that coders code randomly when marble colors match, and code honestly when colors mismatch. But π assumes drawing with replacement, while α assumes no replacement. With no replacement, more marbles produce more color matches. For example, two coders draw once from two marbles, one black and one white. The second coder has 0% chance matching the color of the first. Drawing once from four marbles, again half black and half white, the second coder has a 33.33% chance matching that of the first. Ten marbles would increase the chance to 44.44%, and 1,000 marbles would increase it to slightly lower than 50%. As α assumes the number of marbles is linked to the target sample (Equation 16, Zhao et al., 2013), a larger target sample means more marbles, more color matches, therefore more random coding and more chance agreement, leading to a lower α.

If the drawing is with replacement like π assumes, sample size does not affect the probability of color match. For example, two coders draw once from two marbles, one black and one white. Because the marble is replaced after each drawing, the second coder has 50% chance matching the color of the first coder. Drawing from 4, 10 or 1,000 marbles would produce the same probability. The size of target sample or marble population does not affect the π-estimated random coding or chance agreement, hence does not affect π.

Assuming replacement or no replacement is the only difference between Scott's π and Krippendorff's α with categorical scales and two coders. The difference in probabilities is the largest when the sample is the smallest (N=2). As the sample gets larger, α approaches π. When the sample is over 100, α and π are almost the same for most of the practical purposes.

So, everything else being equal, α defines smaller sample as more information, therefore higher reliability, while π assumes sample size unrelated to reliability or information. Now that we are reminded this is the *only* difference between the two indices for nominal scales with two coders, we need to revise Zhao et al.'s (2013) recommendation to sometimes use α. Researchers should almost never use α. Where Zhao et al. (2013) recommended α, researchers should use its better twin π, unless the coders' behavior follows the highly restrictive Krippendorff Scenario and the data are too precious to discard.

4. Target Invariance ≠ Instrument Invariability

This section shows that Krippendorff confused *target variation* with *instrument variability*, to make a defect of α look like the defect of the coding instrument

Krippendorff's α, like π, κ and their equivalents, cannot be calculated when only one category is used (Zhao et al., 2013, Abnormality 11). To defend α, Krippendorff (2013) provided an example (p. 483-484):

> Fire extinguishers tend to have a pressure gauge. Usually the pointer on the dial does not change. Their owners have no ability to vary the gauge and therefore no clue whether it indicates hydrostatic pressure or is stuck and dysfunctional. Therefore, by law, fire extinguishers need to be checked by professionals who can de- and re-pressurize the extinguisher. Without demonstrable variability, there is no evidence of the reliability of the gauge.

Here Krippendorff confused *target variation* with *instrument variability*. With the distinction understood, Krippendorff's example further illustrates π, κ and α's deficiencies. Suppose a gauge responds properly to de- and re-pressurization, it is not stuck, and the professional should report that the gauge is functional. Indices of intercoder reliability should behave like this professional. Some indices do. But κ, π and α don't.

Zhao et al. (2013, pp. 454-455) provided an example showing that, when both coders found Surgeon General's Warnings in all 1,000 (or any other number of) cigarette advertisements, α, κ and π were incalculable. Krippendorff (2013) blamed the *no report* on *instrument invariability* – the coders were incapable of reporting "no" when advertisements fail to display the Warning. By shifting the blame, Krippendorff (2013) avoided responding to the real criticism. That is, α, κ and π are incalculable when both of the following are true: 1) the *coding target is invariant*, e.g., every advertisement displays the Warning, and 2) *the coding instrument is capable of varying* but stay invariant with the invariant target, e.g., both coders report yes every time, not because they are unable to say no, but because the Warning is in every advertisement.

The three indices are like professionals you hired to test your gauge, but refused to report whether the gauge works. When you complained, they gave you a sophisticated discussion about "need for information" and "demonstrable variability," to convince you that you should not have asked for a report. You compared notes with your neighbors and found that these professionals are actually *incapable of producing a report* until fire flares. Don't hire them again.

A coder's coding should vary when the target of the measurement varies, and not vary when the target does not. When two coders do so consistently, an index should find the coding instrument reliable, including when the coders' coding stays invariant in accordance with its invariant target. Indices α, κ and π are incapable of giving a report in the last situation, including when the instrument is confirmed capable of varying.

A reader may do a simple experiment. Find a coding partner. Look at any 10 people to see if each has a third eye. Suppose neither of you finds any, producing 100% agreement. Calculate π, κ and α, and you will find them incalculable. What's wrong? The Krippendorff logic would blame the invariable instrument – you the coders are incapable of seeing the extra eyes, which amounts to another sensitivity defense. Somebody must grow a third eye or your research is worse than unreliable. We, by contrast, blame unreliable indices – π, κ and α are incapable of seeing the perfect agreement on the absence of extra eyes. Who is right? You decide.

Now do the opposite: Look for noses in the 10 faces. If each of you finds a nose in every face, π, κ and α are again incalculable. The Krippendorff Logic would again blame instrument invariability, but this time through *specificity defense*, accusing you the coders being incapable of seeing the absence of noses. Somebody must lose a nose or your coding is beyond bad. If you are confident of the coders' ability to decide the presence or

absence of human noses, you should be skeptical of π, κ and α's ability to measure reliability.

The blame is not on the coders or their instruments' invariability. The blame is on the indices' unrealistic assumptions about coder behaviors (Zhao et al., 2013).

Indeed, only the indices' behavioral assumptions can properly explain the indices' incalculability. Indices π, κ and α assume coders draw marbles, and marbles' color distribution equals target distribution. Reliability is calculated only for honest coding, which occurs only when the colors mismatch. Invariant target implies a single marble color, therefore no color mismatch, no honest coding, and no calculable π, κ or α.

Krippendorff (2013, p. 484) offered a second example of a thermometer. Readers may take the example to further illustrate π, κ and α' deficiencies. Remember to assume 1) target invariance, i.e., the target temperature stays invariant; 2) instrument variability, i.e., the thermometer is capable of varying with temperature.

5. Maximum Random Assumptions ≠ Metaphors

This section shows that Krippendorff re-characterized a key assumption of α as a metaphor, so that the assumption appears beyond analysis, evaluation, or criticism.

Krippendorff (2013, p. 487) wrote: "I consider all assumptions and paradoxes based on the metaphorical scenarios that the authors (Zhao et al., 2013) have constructed to be flawed." And (p. 492) "the authors' references to 'randomness,' 'random guessing,' 'random coding,' 'randomly drawing marbles from an urn,' are completely metaphorical, unrelated to how a_c is obtained in fact, …."

Numerous authors including Krippendorff discussed "flipping a … coin" or "throwing dice" (e.g., Goodman & Kruskal, 1954, p. 757; Krippendorff, 2004a, p.114, 226; Krippendorff, 2004b, p. 413; Riffe, Lacy, & Fico, 1998, p. 129, 130). Klebanov & Beigman (2009, p. 496) made it explicit, stating "the main assumption is that … annotators … (sometimes) flip a coin." Preemptively, Zhao et al., (2013) clarified (p. 425, emphasis added):

> 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.

While the indices do not necessarily assume physical coins, dice, marbles, or urns, they assume the coders maximize random coding using equivalent devices, physical, mental, electronic, virtual, or of other forms, following the procedures that Zhao et al. (2013) detailed. These procedures are not metaphors.

Following the procedures, each index equates something in the coding scheme or target sample with something in the marbles, namely "a population of real or conceptual collection of random elements." The category-based indices equate the number of categories in the coding scheme with the number of (real or virtual) marble colors, leading to the category-related paradoxes. Distribution-based indices equate distribution in the target sample with (real or virtual) marble distribution, leading to distribution-related paradoxes. In the real world, marbles, whether real or virtual, are rarely linked with research targets or coding schemes. The discrepancy between the assumptions and the reality is the culprit.

Estimating and removing random agreements have been *the* central concern of all chance-adjusted indices, including Krippendorff's α. By calling randomness and random coding (not just coins, dice, marbles or urns) "completely metaphorical" and categorically denying their role in the indices, Krippendorff was to make α untouchable, i.e., beyond serious analysis.

This action, however, also ripped the heart out of α and the other indices. While labeling the description and justification of the random drawing procedure "completely metaphorical," Krippendorff did not offer any alternative description or justification for the procedures. Now, at least for α, this core assumption is described and justified solely by a so called "metaphor." Yet the prescribed procedure is deemed beyond analysis, evaluation, or criticism, because it is only a metaphor.

Readers can decide for themselves whether the indices indeed assume what Zhao et al. (2013) said they assume. Take the example of α. Examine the Krippendorff Scenario on p. 451 of Zhao et al. (2013). Do a math exercise: Produce an equation to calculate a_c based on this scenario. If needed, find help from a high school student good at math. Assuming no derivation errors, if the resulted formula is equivalent to Eq. 15 in p. 445, the Scenario does describe the assumptions behind α and the description is not just a metaphor. If substantively different, the Scenario describes something else, possibly just a metaphor.

And we invite Prof. Krippendorff to do the same: produce an equation mathematically consistent with the Krippendorff Scenario but inconsistent with Krippendorff's α, so as to convince the skeptics that the Krippendorff Scenario alleged by Zhao et al., (2013) does not represent the coder behaviors assumed by α, and α is indeed flawless.

6. Stable Sign-Meaning Pairing: Intercoder Reliability = Intercoder Reliability

To redefine a term or to relabel a concept is to de-pair a conventional pair of a sign (word, symbol, signifier, term) and its meaning (concept, signified, signification), and to make a new pair(s). *Intercoder reliability*, for

example, has been conventionally defined as intercoder agreement. Krippendorff did not deny that α does not accurately measure agreement, or that α is significantly affected by skew. The discrepancies between what α is supposed to measure (agreement) and what it actually measures (partially distribution skew, partially agreement, and partially some other concepts) produce paradoxes and abnormalities. Facing criticism, Krippendorff switched a key sign, *reliability*, away from its traditional meaning *agreement* and toward a new meaning *distribution skew*.

It's shrewd. A function of language is to express. The expresser has the arbitrary discretion to decide the match between his *signs* and his *meanings* (de Saussure, 1916, 2004; Keller, 1998). An expresser may match any sign(s) with any concept(s), and may change the match anytime in anyway. If a man places *his* sign *shirts* on *his* shirts, and later relabels the shirts as *socks*, these are his choices, and are not right or wrong. When Krippendorff delinked *his* sign *reliability* from *his* original meaning "agreement," and relinked *reliability* with "distribution skew," "information" and "variation," these were his choices, and were not right or wrong.

The more important function of language, however, is to communicate. Here the sign-meaning relationship is no longer the sole property of a speaker, writer, sender, releaser or any one stake holder. Rather, the sign-meaning associations are socially constructed, physiologically conditioned, and practically contracted among all parties concerned (Berger & Luckmann, 1966; Searle, 1995; Zhao, Chen, & Tong, 2011). If a buyer ordered shirts but instead received socks, the seller would be ill-advised to argue that he has the right to arbitrarily decide the meaning of *shirts*.

Switching a sign away from its agreed meaning is not a new tactic. It has been seen in academic debates in Chinese, for example, so often that 11 rules of concept naming had been put forward, including Rule #6: *Respect words' traditional meanings* (Zhao, 2004, 2005, 2008, p. 75).

As mentioned, we as a scholarly community once had an agreement to measure "agreement" under the sign *reliability*. Krippendorff proposed to redo this agreement, so that *reliability* would not (or not only) represent "agreement," but instead (or also) represent "information," "variation," and "distribution skew." The objective is to justify α, which is heavily influenced by distribution skew (Feng, 2013a, 2013c). If we accept this proposal, it would change the concepts of reliability, agreement, and information as we know them. For social sciences, the foreseeable benefit of this dramatic change, which is to save α, is too small, and cost too large. We vote to stay with the extant agreement of defining intercoder reliability as agreement. If we are serious about measuring information, variation, or distribution skew, we may do so more effectively under the existent signs *information*, *variation*, and *distribution skew*, rather than mixing them with *reliability*.

7. Multi-Signification: $Reliability_a \neq Reliability_b$, & $Information_a \neq Information_b$

Then there is Rule #3: *Avoid using one word with multiple meanings*, or one signifier to signify multiple concepts (Zhao, 2004, 2005, 2007, 2008, 2009). The phenomenon is called *multi-signification*, multi-signified, or multi-concepts (Zhao et al., 2011). Since Socrates, logicians have pointed out time and again that multi-signification can cause equivocation fallacies (Fischer, 1970, p. 274). For example: a feather is light; what is light cannot be dark; therefore a feather cannot be dark (Malloch & Huntley, 1966). Such word-plays intrigue as the multi-significations are not immediately obvious; they do no harm as the faults in the conclusions are obvious; they are fun as they intrigue while doing no harm.

In an earnest effort to defend α, Krippendorff (2013) fell for similar multi-significations. The sign *information* represents at least two concepts, *mechanical information* and *human information*. When Krippendorff (2013, pp. 483-484) talked about "informational requirement" and the "need to have enough information," we may be sympathetic, but only because we thought of *information* as "human information," which is assessed in particular contexts in reference to particular human meanings. What Krippendorff (2011a, 2013) actually calculated, however, was "mechanical information," which Krippendorff believes is universally calculable based on statistical variation with no reference to content or context. Larger variation is the necessary and sufficient condition for more *information*. Had this been explicit, we would have been less sympathetic. That is, a measurement is no more reliable or informative simply because it reports a larger variation.

Krippendorff did not create the two meanings of *information*. He used what's available. He did, however, create multiple meanings for the sign *reliability*. As said, reliability has been traditionally defined as agreement, and recently Krippendorff advocated switching *reliability* away from agreement and toward information, variation, and distribution skew. He criticized the "one-to-one" relationship "between agreement and reliability." But he never said reliability is unrelated to agreement, perhaps because he knew α is still heavily influenced by agreement, although it is as heavily influenced by distribution skew (Feng, 2013a, 2013b; Zhao, 2012a). The result is another multi-signification in Krippendorff's writings, where the word *reliability* may represent different concepts at different times, *agreement*, *information*, *variation*, and *distribution skew*. What's constant is the imperative conclusion, that α is the best and perfect.

We urge, again, to refrain from multi-signifying, although it's tempting to do in debates. Multi-significations confuse authors as much as they confuse audiences (Zhao, 2004, 2005, 2007, 2009; Zhao, Chen, & Tong, 2011).

8. Multi-Signifiers: Quota = Pre-coding Agreement = a priori Marginal = Assigned Prevalence = Fixed Probability = Known Distribution

While *multi-signification* or *multi-signified* refer to a single sign representing two or more concepts, *multi-signs* or *multi-signifiers* refer to two or more signs representing a single concept. This section shows that several groups of reliability experts separately sensed or detected what is now known as the *quota assumptions* underlying π, κ and α. Each group used a different term. Due in part to the multi-signs, the later authors did not cite the earlier ones, and none of them appeared to be aware that others shared their views on the quota assumptions.

Zhao et al., (2013) discussed the quota assumptions extensively; the word *quota* appeared more than 60 times in the paper. Krippendorff (2013), however, responded with half of sentence embedded in the middle of a long paragraph: "coefficients do not *predefine* quotas *for* coders" (p. 487, emphases added). The word *quota* appeared nowhere else in the 19-page commentary.

In this only response, Krippendorff refuted yet another strawman – a position that we never expressed or espoused. Our position was and is: the coefficients (indices) *presume* quotas *by* coders, that is, π, κ and α assume that coders predetermine quotas, which we call *quota assumptions* (Zhao, 2011a; 2011b; Zhao et al., 2013). More than five years after Zhao et al. (2013) discussed the quota assumptions, the advocates of π, κ or α have yet to explicitly deny or acknowledge them.

Krippendorff appeared reluctant to do either. Acknowledging the assumptions would disqualify α (and π and κ) because coders rarely if ever observe quotas. Denying is also untenable because Krippendorff had years earlier recognized the assumption for α.

To justify "total agreement, undefined α," which occurs with 0% & 100% observed distributions, Krippendorff (2004b, p.425) found it necessary to assume "coders … agreed in advance of the coding effort to make their task easy." To *agree pre-coding* on 0% & 100% distributions, or any other distributions, is to set quotas.

Others recognized the quota assumptions decades earlier. Brennan & Prediger (1981, p. 687) gave "special consideration … to assumptions about whether marginals are fixed a priori, or free to vary." They concluded "when marginals are fixed, coefficient kappa is found to be appropriate," but "when either or both of the marginals are free to vary," *S* is more appropriate. The only way to fix a priori (pre-coding) cross-table marginals (frequency distributions) is to set and execute quotas.

Shrout, Spitzer, & Fleiss (1987, p. 173), who defended κ, prominently assumed --

> Neither clinician interviewed any of the subjects but both simply randomly assigned 6% of them to the case group – perhaps because they expected that the prevalence of a current DSM-III disorder in a general population would be low …

Assigning prevalence (6%) before interviewing patients is to set a quota.

Feinstein & Cicchetti (1990a, p. 548), who criticized κ, noted that κ "makes the assumption that each observer has a relatively fixed probability of making positive or negative responses." *Fixed probability* is another word for quota.

Interestingly, Cohen (1960) appeared just one question away from recognizing the quota assumptions behind π and κ. Advocating κ over π, Cohen (1960, pp. 40-41, emphases added) observed:

> (Scott's π, 1955) assumes…the *distribution of proportions over the categories* for the population *is known* and is taken to *be equal for the judges*. The former assumption is reasonable in survey research, but the latter may be questioned in more general applications … .

How can the distribution be known before coding? Only one way -- coders (judges) observe quotas. As Cohen (1960) considered the assumption reasonable, he made it the foundation of κ. Consequently, κ assumes quota just like π does. What Cohen disputed was π's second assumption, that the quota-based distribution be equal for the coders, which we now call *conspired quota*. Cohen replaced this assumption with another, that the "judges … distribute their judgments differently," which we now call *individual quota*.

Krippendorff's α adopts both of Scott's assumptions. Consequently α assumes conspired quota just like π does (Zhao et al. 2013). Sharing the very similar quota assumptions make the three indices behave like one index in simulations and experiments (Feng, 2013a; Zhao, 2011a,b; 2012a).

Curiously, Cohen (1960, pp. 40-41) considered the assumption "reasonable" only for "survey research." Does it imply that Cohen in 1960 was not sure κ was appropriate for non-survey research, which he called "general applications"? Communication researchers don't usually consider content analysis survey research, and medical doctors rarely call patient diagnoses survey research. What would Cohen have said had he known that κ is now used more often in content analysis and medical research than in what we call "survey research"?

9. Redefining Reliability Disqualifies α

The scientific community agreed *intercoder reliability* meant "intercoder agreement." For decades, Krippendorff (1970a, 1970b, 1980, 2011b) was a prominent part of the near consensus. When α was found to be often low with high agreement, Krippendorff (2011a, 2013) redefined reliability to delink it from agreement and relink it with information.

Furthermore, when skewed distribution was found to produce low α, skewed distribution was redefined as less information. When larger sample was found to produce lower α, larger sample was redefined as less information. When smaller variation was found to produce lower α, smaller variation was redefined as less information. When low sensitivity or low specificity was found to produce low α, low sensitivity or low specificity was redefined as less information. When zero variation was found to lead to incalculable α, zero variation was redefined as no information. As reliability was redefined as information, higher reliability was redefined as more even distribution, smaller sample, larger variation and higher sensitivity or specificity. The stated mission of the flurry of redefinitions is to defend α as the standard measure of intercoder reliability.

When a tool fails its defined mission, we repair or replace the tool. Task dictates tool. Mission directs means. When Krippendorff's α failed its defined mission, Krippendorff (2011a, 2011b, 2013) revised and redefined the mission. The tool dictated the tasks. The means directed the missions.

While the other indices also have their deficiencies, several are closer than α to measuring agreement. The evidences do not justify α as an adequate index, even less a better index, and further less the standard index.

We join the calls for a new index(es) based on more realistic assumptions. We are making progress developing such an index, and we invite others to join (Zhao, 2012a). But it will take time to gain acceptance. The decades-long dominance of α in communication studies and κ in other disciplines create a spiral of inertia that will likely propel and sustain itself for years, during which content researchers will have to make use of the existing indices (Feng & Zhao, 2016).

Zhao et al. (2013, Table 19.13) suggested guidelines on which indices to avoid or use in various situations. As no index is adequate for all situations, our objective has to be modest, which is to avoid the most severe paradoxes, abnormalities and other known deficiencies, by recommending *the best available for a situation* (BAFS). For those who use BAFS scheme to select indices, Zhao et al. (2012a) developed several hierarchies that shows which indices tend to be more liberal or conservative under which circumstances, which may be helpful for interpreting the indices.

10. Spiral of Inertia in Reliability Research

Evidences are mounting that Scott's π, Cohen's κ and Krippendorff's α are inferior indicators of intercoder reliability, and theories are becoming clearer why they are (Conger, 2016; Feng, 2013a; Flight & Julious, 2015; Grant et al., 2017; Gwet, 2002; Shankar & Bangdiwala, 2014; Wongpakaran et al., 2013; Zhao, 2011a, 2011b, 2012a; Zhao et al., 2013) . Nevertheless, α may continue to dominate communication content research, and κ may continue to dominate some other disciplines, thanks to *spiral of inertia*, a form of *selective spiral* (Slater, 2007, 2015; Slater, Henry, Swaim, & Anderson, 2003; Zhao, 2000b, 2002, 2009b, 2012b; Zhao, 2017).

Ideally, science advances in a revolving and evolving process. Researchers conceive new ideas, write up proposals, reviewers review, funded researchers investigate, write up findings, and submit manuscripts for publication; editors assign reviewers, reviewers review, editors request revision, authors revise, manuscripts get published; readers read, and recommend others to read; more readers read, some accept the new knowledge, based on which some conceive their own ideas, seek funding, and investigate, so goes another round.

In our experience teaching and analyzing π, κ and α over the years, however, we saw new ideas resisted every step (Zhao, 2000a). The more different an idea was from the prevailing view, the stronger the resistance. Scott's π, Cohen's κ and Krippendorff's α have dominated various disciplines for decades. Of the qualified reviewers, few have not advocated, recommended, used or taught π, κ or α, or accepted its main premises.

Proposals and manuscripts criticizing one of the trio were routinely rejected. Committees disapproved such proposals anticipating publication difficulties. Editors desk rejected the manuscripts anticipating reviewer oppositions, or demanded unanimous and unreserved support from reviewers anticipating post-publication controversies. Time and again, we switched our attention elsewhere, and discouraged others from getting interested, fearing jeopardizing their careers. The authors and editors who publish studies mentioning α's defects learned to anticipate energetic reactions from α's author (Feng, 2015; Hayes & Krippendorff, 2007; Krippendorff, 2004b, 2013, 2016, Lombard et al., 2002, 2004; Zhao et al., 2013). Consequently, in the few cases when criticisms of the trio were to be published, some editors insisted on removing the sharpest criticisms and blurring the main conclusions. After perusing the published debates, some authors told us they now understood the serious flaws of the trio; they then went on to calculate and report one of the trio, because that's what reviewers, editors, and readers likely like to see.

Days after seeing a draft of Krippendorff's (2013) criticism of Zhao et al. (2013), we finished the first draft of this response. It took over five years to find a journal able and willing to publish the response.

These actions and inactions foster an impression that the three indices are more reliable or acceptable than they are. The impression encourages more actions and inactions favoring the trio, which further strengthen the false impression, leading to a *spiral of inertia*, a more active and aggressive version of *static inertia* or *spiral of silence* (Feng & Zhao, 2016). It's remarkable that, in spite of the powerful spiral, criticisms of the trio persisted, and some even managed to be published (Feinstein & Cicchetti, 1990a, 1990b; Feng, 2015; Feng & Zhao, 2016; Flight & Julious, 2015; Grant et al., 2017; Grove et al., 1981; Gwet, 2002; Lombard et al., 2002; Shankar & Bangdiwala, 2014; Zhao et al., 2013).

****Authors' Note**

This research is supported in part by a donation from Dr. Lee Shau Kee, GBM, through HKBU Research Grant Scheme (LSK/14-15/P13, Zhao PI), and grants from China Ministry of Education (Social Sciences) through Fudan University Center for Information and Communication Studies (11JJD860007, Zhao PI), the 13th Five-Year Plan of Shenzhen for Philosophy and Social Sciences 2017-2018 (135A011, Feng PI), China Ministry of Education (Humanities and Social Sciences) 2015-2018 (15YJA860004, Feng PI), and National Science Foundation of China (11401338, Deng PI).

Xinshu Zhao is Chair Professor of Communication at Hong Kong Baptist University and Cheung Kong Chair Professor of Journalism at Fudan University. He is also Emeriti Professor of Media and Journalism at the University of North Carolina at Chapel Hill.

Guangchao Charles Feng is Distinguished Professor of Communication at Shenzhen University, Guangdong, China.

Jun S. Liu is Professor of Statistics and Biostatistics at Harvard University. He is an elected Fellow of the Institute of Mathematical Statistics and of the American Statistical Association. He has won the NSF Career Award, the COPSS Presidents' Award, and the Morningside Gold Medal. He has served as associate editor and co-editor for the *Journal of the American Statistical Association*, and is author of *Monte Carlo Strategies in Scientific Computing*.

Ke Deng is Associate Professor and Associate Director, Center for Statistical Science, Tsinghua University, Beijing, China. His research interests include statistical modeling, statistical computation and applications in bioinformatics, text mining and sociology.

Correspondence to:
Prof. Xinshu ZHAO
School of Communication
Hong Kong Baptist University
5 Hereford Road, Kowloon Tong
Kowloon, Hong Kong
Email: zhao@hkbu.edu.hk
Tel: (852) 3411-7481

References

Altman, D. G., & Bland, J. M. (1994). Statistics Notes: Diagnostic tests 1: sensitivity and specificity. *BMJ: British Medical Journal*, *308*(6943), 1552–1552. http://doi.org/10.1136/bmj.308.6943.1552

Bakeman, R. (2000). Behavioral observation and coding. *Handbook of Research Methods in Social and Personality Psychology*, 138–159. http://doi.org/10.1037/13619-013

Benini, R. (1901). *Principii di Demongraphia: Manuali Barbera Di Scienze Giuridiche Sociali e Politiche (No. 29)[Principles of demographics (Barbera Manuals of Jurisprudence and Social Policy)]*. Firenze, Italy: G. Barbera.

Berger, P. L., & Luckmann, T. (1966). *The Social Construction of Reality*. Anchor Books. http://doi.org/10.2307/323448

Bloch, D. A., & Kraemer, H. C. (1989). 2 x 2 Kappa Coefficients: Measures of Agreement or Association. *Biometrics*, *45*(1), 269–287. http://doi.org/10.2307/2532052

Brennan, R. L., & Prediger, D. J. (1981). Coefficient Kappa: Some Uses, Misuses, and Alternatives. *Educational and Psychological Measurement*, *41*(3), 687–699. http://doi.org/10.1177/ 0013164 48104100307

Brown, J. D., Zhao, X., Wang, M. N., Liu, Q., Lu, A. S., Li, L. J., … Zhang, G. (2013). Love is all you need: A content analysis of romantic scenes in Chinese entertainment television. *Asian Journal of Communication*, *23*(3), 229–247. http://doi.org/10.1080/01292986.2012.729148

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. http://doi.org/10.1177/001316446002000104

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220. http://doi.org/10.1037/h0026256

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Statistical Power Analysis for the Behavioral Sciences* (Vol. 2nd). http://doi.org/10.1234/12345678

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. http://doi.org/10.1037/0033-2909.112.1.155

Condon, S. (2012, January 4). Iowa caucus results: Santorum and Romney in dead heat. *CBS News*.

Conger, A. J. (2016). Kappa and Rater Accuracy: Paradigms and Parameters. *Educational and Psychological Measurement*, 13164416663277. http://doi.org/10.1177/0013164416663277

Cousineau, D., & Laurencelle, L. (2016). An unbiased estimate of global interrater agreement. *Educational and Psychological Measurement*, 13164416654740. http://doi.org/10.1177/0013164416654740

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. http://doi.org/10.1007/BF02310555

de Saussure, F. (1916). Nature of the linguistic sign. In C. Bally, A. Sechehaye, & W. (translator) Baskin (Eds.), *Course in general linguistics* (pp. 68–73). New York: McGraw-Hill Education.

de Saussure, F. (2004). Course in general linguistics. In J. Rivkin & M. Ryan (Eds.), *Literary Theory: An Anthology* (2nd ed., pp. 59–71). New York: Blackwell Publishing.

Dewey, M. E. (1983). Coefficients of agreement. *British Journal of Psychiatry*, *143*(5), 487–489. http://doi.org/10.1192/bjp.143.5.487

Feinstein, A. R., & Cicchetti, D. V. (1990a). High agreement but low Kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 543–549. http://doi.org/10.1016/0895-4356(90)90158-L

Feinstein, A. R., & Cicchetti, D. V. (1990b). High agreement but low Kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 551–558. http://doi.org/DOI: 10.1016/0895-4356 (90)90159-M

Feng, G. C. (2013a). Factors affecting intercoder reliability: A Monte Carlo experiment. *Quality and Quantity*, *47*(5), 2959–2982. http://doi.org/10.1007/s11135-012-9745-9

Feng, G. C. (2013b). *Indexing versus Modeling Intercoder Reliability*. Hong Kong Baptist University.

Feng, G. C. (2013c). Underlying determinants driving agreement among coders. *Quality and Quantity*, *47*(5), 2983–2997. http://doi.org/10.1007/s11135-012-9807-z

Feng, G. C. (2015). Mistakes and how to avoid mistakes in using intercoder reliability indices. *Methodology*, *11*(1), 13–22. http://doi.org/10.1027/1614-2241/a000086

Feng, G. C., & Zhao, X. (2016). Do not Force Agreement – A Response to Krippendorff. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *12*(4), 145–148. http://doi.org/10.1027/1614-2241/a000120

Feuerman, M., & Miller, A. R. (2008). Relationships between statistical measures of agreement: sensitivity, specificity and kappa. *Journal of Evaluation in Clinical Practice*, *14*(5), 930–933. http://doi.org/10.1111/j.1365-2753.2008.00984.x

Fischer, D. H. (1970). *Historian's Fallacies: toward a logic of historical thought*. New York: Harper Collins.

Flight, L., & Julious, S. A. (2015). The disagreeable behaviour of the kappa statistic. *Pharmaceutical Statistics*, *14*(1), 74–78. http://doi.org/10.1002/pst.1659

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classification. *Journal of the American Statistical Association*, *49*(268), 732–764.

Grant, M. J., Button, C. M., & Snook, B. (2017). An Evaluation of Interrater Reliability Measures on Binary Tasks Using d-Prime. *Applied Psychological Measurement*, *41*(4), 264–276. http://doi.org/10.1177/0146621616684584

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*(4), 408–413. http://doi.org/10.1001/archpsyc.1981 .01780290042004

Gwet, K. L. (2002). Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical Methods for Inter-Rater Reliability Assessment*.

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*(1), 29–48. http://doi.org/10.1348/000711006X126600

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

Hartley, R. V. L. (1928). Transmission of information. *Bell System Technical Journal*, *7*(3), 535–563.

Hayes, A. F., & Krippendorff, K. H. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77–89. http://doi.org/10.1080/19312450709336664

Hoehler, F. K. (2000). Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *Journal of Clinical Epidemiology*, *53*(5), 499–503. http://doi.org/10.1016/S0895-4356(99)00174-2

Keller, R. (1998). *A theory of linguistic signs*. (D. (translator) Duenwald, Ed.). Oxford, UK: Oxford University Press.

Khan, H., Friedman, E., & Shushannah, W. (2012, January 4). Iowa Caucus Results: Romney Edges Santorum by 8 Votes. *ABC News*.

Kirilenko, A. P., & Stepchenkova, S. (2016). Inter-Coder Agreement in One-to-Many Classification: Fuzzy Kappa. *PloS One*, *11*(3), e0149787.

Klebanov, B. B., & Beigman, E. (2009). From Annotator Agreement to Noise Models. *Computational Linguistics*, *35*(4), 495–503. http://doi.org/10.1162/coli.2009.35.4.35402

Kolmogorov, A. N. (1968). Three approaches to the quantitative definition of information. *International Journal of Computer Mathematics*, *2*(1–4), 157–169. http://doi.org/10.1080/00207166808803030

Kraemer, H. C. (1979). Ramifications of a population model for Kappa as a coefficient of reliability. *Psychometrika*, *44*(4), 461–472. http://doi.org/10.1007/BF02296208

Kraemer, H. C., & Bloch, D. A. (1988). Kappa coefficients in epidemiology: An appraisal of a reappraisal. *Journal of Clinical Epidemiology*, *41*(10), 959–968. http://doi.org/10.1016/0895-4356(88)90032-7

Kraemer, H. C., Periyakoil, V. S., & Noda, A. (2002). Tutorial in Biostatistics: Kappa coefficients in medical research. *Statistics in Medicine*, *21*(14), 2109–29. http://doi.org/10.1002/sim.1180

Krippendorff, K. H. (1970a). Bivariate agreement coefficients for reliability of data. *Sociological Methodology*, *2*, 139–150. http://doi.org/10.2307/270787

Krippendorff, K. H. (1970b). Estimating the Reliability, Systematic Error and Random Error of Interval Data. *Educational and Psychological Measurement*, *30*(1), 61–70. http://doi.org/10.1177/001316447003000105

Krippendorff, K. H. (1980). *Content Analysis: An Introduction to its Methodology*. Thousand Oaks, CA: Sage.

Krippendorff, K. H. (2004a). *Content Analysis: An Introduction to Its Methodology* (2nd ed.). Thousand Oaks,CA: Sage. http://doi.org/10.2307/2288384

Krippendorff, K. H. (2004b). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, *30*(3), 411–433. http://doi.org/10.1093/hcr/30.3.411

Krippendorff, K. H. (2011a). Agreement and Information in the Reliability of Coding. *Communication Methods and Measures*, *5*(2), 93–112. http://doi.org/10.1080/19312458.2011.568376

Krippendorff, K. H. (2011b). Computing Krippendorff's alpha-reliability. Retrieved from http://repository.upenn.edu/cgi/viewcontent.cgi?article=1043&context=asc_papers

Krippendorff, K. H. (2012). *Content Analysis: An Introduction to its Methodology* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Krippendorff, K. H. (2013). Commentary: A dissenting view on so-called paradoxes of reliability coefficients. In C. T. Salmon (Ed.), *Communication Yearbook 36* (pp. 481–499). Routledge.

Krippendorff, K. H. (2016). Misunderstanding reliability. *Methodology*, *12*(4), 139–144.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability. *Human Communication Research*, *28*(4), 587–604. http://doi.org/10.1111/j.1468-2958.2002.tb00826.x

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2004). A call for standardization in content analysis reliability. *Human Communication Research*, *30*(3), 434–437. http://doi.org/10.1093/hcr/30.3.434

Malloch, A. E., & Huntley, F. L. (1966). Some notes on equivocation. *Publications of the Modern Language Association*, 145–146.

Maxwell, A. E. (1977). Coefficients of agreement between observers and their interpretation. *British Journal of Psychiatry*, *130*(1), 79–83. http://doi.org/10.1192/bjp.130.1.79

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

Perreault, W. D., & Leigh, L. E. (1989). Reliability of nominal data based on qualitative judgments. *Journal of Marketing Research*, *26*(2), 135–148. http://doi.org/10.2307/3172601

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* (1st ed., pp. 99–105). New York, NY: St. Martin's / Springer.

Reliability (Statistics). (2017). Retrieved September 20, 2017, from https://en.wikipedia.org/wiki/ Reliability _(statistics)

Riffe, D., Lacy, S., & Fico, F. G. (1998). *Analyzing Media Messages: Using Quantitative Content Analysis in Research*. Mahwah, N J: Lawrence Erlbaum Associates.

Roberts, C. (2008). Modelling patterns of agreement for nominal scales. *Statistics in Medicine*, *27*(6), 810–830. http://doi.org/10.1002/sim.2945

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

Saal, F. E., Downey, R. G., & Lahey, M. a. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*(2), 413–428. http://doi.org/10.1037/0033-2909.88.2.413

Scott, W. A. (1955). Reliability of Content Analysis: The Case of Nominal Coding. *Public Opinion Quarterly*, *19*(3), 321–325.

Searle, J. R. (1995). *Construction of Social Reality*. *The Free Press* (Vol. 1).

Shankar, V., & Bangdiwala, S. I. (2014). Observer agreement paradoxes in 2x2 tables: Comparison of agreement measures. *BMC Medical Research Methodology*, *14*(1). http://doi.org/10.1186/1471-2288-14-100

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(July 1928), 379–423. http://doi.org/10.1145/

584091.584093

Shrout, P. E., Spitzer, R. L., & Fleiss, J. L. (1987). Quantification of agreement in psychiatric diagnosis revisited. *Archives of General Psychiatry*, *44*(2), 172–177. http://doi.org/10.1001/archpsyc.1987.01800140084013

Slater, M. D. (2007). Reinforcing spirals: The mutual influence of media selectivity and media effects and their impact on individual behavior and social identity. *Communication Theory*, *17*(3), 281–303. http://doi.org/10.1111/j.1468-2885.2007.00296.x

Slater, M. D. (2015). Reinforcing spirals model: Conceptualizing the relationship between media content exposure and the development and maintenance of attitudes. *Media Psychology*, *18*, 370–395. http://doi.org/10.1080/15213269.2014.897236

Slater, M. D., Henry, K. L., Swaim, R. C., & Anderson, L. L. (2003). Violent media content and aggressiveness in adolescents: A downward spiral model. *Communication Research*, *30*, 713–736. http://doi.org/10.1177/0093650203258281

Spitznagel, E. L., & Helzer, J. E. (1985). A proposed solution to the base rate problem in the kappa statistic. *Archives of General Psychiatry*, *42*(7), 725–8. http://doi.org/10.1001/archpsyc.1985.01790300093012

Tinsley, H. E., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, *22*(4), 358–376. http://doi.org/10.1037/h0076640

Tinsley, H. E., & Weiss, D. J. (2000). Interrater Reliability and Agreement. *Handbook of Applied Multivariate Statistics and Mathematical Modeling*, 95–124. http://doi.org/10.1037/h0076640

Uebersax, J. S. (1987). Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin*, *101*(1), 140–146. http://doi.org/10.1037/0033-2909.101.1.140

Uebersax, J. S. (2009). The Myth of Chance Corrected Agreement. Retrieved October 18, 2012, from http://www.john-uebersax.com/stat/kappa2.htm

Vach, W. (2005). The dependence of Cohen's kappa on the prevalence does not matter. *Journal of Clinical Epidemiology*. http://doi.org/10.1016/j.jclinepi.2004.02.021

Williamson, J. M., Lipsitz, S. R., & Amita K. Manatunga. (2000). Modeling kappa for measuring dependent categorical agreement data. *Biostatistics*, *1*(2), 191–202. http://doi.org/10.1093/biostatistics/1.2.191

Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K. L. (2013). A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Medical Research Methodology*, *13*, 61. http://doi.org/10.1186/1471-2288-13-61

Xu, S., & Lorber, M. (2014). Interrater agreement statistics with skewed data: Evaluation of alternatives to cohen's kappa. *Journal of Consulting and Clinical Psychology*, *82*(6), 1219–1227. http://doi.org/10.1037.a0037489

Zhao, X. (2000a). *Agreement Index as a Reliability Indicator for Nominal Scales with Two Coders*. Paper presented at the annual conference of Association for Education in Journalism and Mass Communication, Phoenix, Arizona, USA, August. https://works.bepress.com/xinshu-zhao/10/.

Zhao, X. (2000b, November). Media Coverage: Wen Ho Lee, China and Beyond. *Making the Global Local*, https://works.bepress.com/xinshu-zhao/15/.

Zhao, X. (2002). Informed Democracy, or Involuntary Mediacracy? (The English Origin of a Chinese Book Chapter). In X. Li & X. Zhao (Eds.), *The Power of the Media* (pp. 37–129). Guangzhou, China: Southern Daily Publishing House. 赵心树 (2002). 知理的民主, 还是盲情的媒主? (英文原稿). 中文改写的删节版载于: 李希光、赵心树编著《媒体的力量》第三章（27-129页）. 广州: 南方日报出版社.

Zhao, X. (2004). The ten-plus-one principles for concept explication and naming. *China Media Report*, *2004*(5), 113–118. 赵心树(2004). 细释冠名的十加一原则.《中国传媒报告》第5期, 113-118 页.

Zhao, X. (2005). The ten-plus-one principles for concept explication and naming. In Y. Luo, Z. Qin, Q. Xia, & H. Wang (Eds.), *Journalism and Communication Review， Vol. 2004* (pp. 49–56). Wuhan: Wuhan Publishing House. 赵心树(2005).细释冠名的十加一原则.罗以澄、秦志希、夏倩芳、王瀚（编）:《新闻与传播评论*２００４卷》49-56页. 武汉: 武汉出版社.

Zhao, X. (2007). Names, missions and constitution of journalism and mass communication -- a discussion with LI Xiguang and PAN Zhongdang. *Journal of Tsinghua University (Social Sciences Edition)*, *22*(5), 100–120. 赵心树 (2007).新闻学与传播学的命名、使命及构成—与李希光、潘忠党商榷.《清华大学学报（社会科学版）》.22(5)总93，100-120页.

Zhao, X. (2008). *Plight of Elections - A Critique of Election Systems and Constitutional Reforms, Expanded Edition*. Chengdu: Sichuan People's Publishing House. 赵心树 (2008).《选举的困境—民选制度及宪政改革批判, 增订版》.成都，四川人民出版社. http://www. chinaelections. org/uploadfile/201003/20100304175038608.pdf

Zhao, X. (2009a). *Plight of Elections - A Critique of*

Election Systems and Constitutional Reforms, Electronic Edition. Beijing: China Election and Governance. 赵心树(2009).《选举的困境－民选制度及宪政改革批判，电子版》. 北京：选举与治理网. http://www.chinaelections.org/uploadfile/201003/20100304175038608.pdf

Zhao, X. (2009b). Spiral of imbalance and national image. In L. Su & C. Chen (Eds.), *Thirty years of Humanities and Social Sciences in China* (pp. 407–412). Beijing: SDX Joint Publishing Company. 赵心树(2009). 失衡螺旋与国际形象.载于苏力、陈春声(编).《中国人文社会科学三十年》.407-412页 北京： 生活·读书·新知三联书店, ISBN: 978-7-108-03226-3

Zhao, X. (2011a). *When to use Cohen's κ, if ever?* Paper presented at the 61st annual conference of International Communication Association., Boston, USA. https://repository.hkbu.edu.hk/coms_conf/2/.

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, USA. https://repository.hkbu.edu.hk/coms_conf/3/.

Zhao, X. (2012a). *A Reliability Index (ai) that Assumes Honest Coders and Variable Randomness*. Paper presented at the annual conference of Association for Education in Journalism and Mass Communication, Chicago, USA, August. http://repository.hkbu.edu.hk/hkbu_staff_publication/6241/.

Zhao, X. (2012b). Spiral of Imbalance in 2008 Olympics Communication and the Growth Pain of China. In A. Shi, Y. Guo, & H. Li (Eds.), *Tsinghua Lecture Notes on Journalism and Communication， Volume II* (pp. 200–210). Beijing,: Tsinghua University Press. 赵心树(2012).中国奥运传播中的"失衡螺旋"与中华民族"成长的烦恼"，载于史安斌、郭云强、李宏刚 (编)：《清华新闻传播学前沿讲座录（续编）》200-210页. 北京： 清华大学出版社

Zhao, X. (2017). Spiral of Hostilities in Politics between China Mainland and Hong Kong (in lieu of foreword). In A. H. Chen, X. Zhao, & X. J. Zhang (Eds.), *Democracy and Election: A Retrospect and Prospect of Hong Kong Political Reform*. Hong Kong: Cosmos Books. 趙心樹(2017).螺旋恶化的中港矛盾（代前言）.陳弘毅、趙心樹著，張小佳編《民主與選舉－香港政改的回顧前瞻》香港：天地圖書

Zhao, X., Chen, Q., & Tong, B. (2011). *Does c' test help, anytime? – on communication fallacy of "effect to mediate."* Paper presented at the annual conference of Association for Education in Journalism and Mass Communication, St. Louis, USA.

Zhao, X., Deng, K., Feng, G. C., Zhu, L., & Chan, V. K. C. (2012). *Liberal-conservative hierarchies for indices of inter-coder reliability*. Paper presented at the 62nd Annual Conference of International Communication Association, Phoenix, Arizona, USA, May.

Zhao, X., Liu, J. S., & Deng, K. (2013). Assumptions behind Intercoder Reliability Indices. In C. T. Salmon (Ed.), *Communication Yearbook 36* (pp. 419–480). New York and London: Routledge. http://doi.org/10.1080/23808985.2013.11679142

(01-12-2018)

CORRECTION. — Since this article was published April, 2018, a correction has been made. In the eighth paragraph of Section 1, entitled "Intercoder Reliability = Intercoder Agreement," the earlier version read: "With identical agreement rates, a skewer distribution can make the three indices hundreds or even tens of thousands times higher than a more even distribution." An apostrophe and "chance estimates" have been inserted after "indices." The corrected sentence now reads: "With identical agreement rates, a skewer distribution can make the three indices' chance estimates hundreds or even tens of thousands times higher than a more even distribution." The correction was made 26 July, 2018 in the online version posted by Hong Kong Baptist University Institutional Repository, https://repository.hkbu.edu.hk/cgi/viewcontent.cgi?article=7826&context=hkbu_staff_publication.