Commentary

Do Not Force Agreement

A Response to Krippendorff (2016)

Guangchao Charles Feng[1] and Xinshu Zhao[2]

[1]School of Journalism and Communication, Jinan University, Guangzhou, China
[2]Hong Kong Baptist University, Hong Kong, China

In an earlier issue of *Methodology*, Feng (2015) surveyed communication scholars' practices of measuring intercoder reliability and recommended guidelines for avoiding mistakes. In the process, Feng (2015) critically reviewed existing agreement indices including Krippendorff's α. Krippendorff commented on Feng (2015) in another contribution to this issue. We welcome the opportunity to rebut on selected issues. Readers may peruse other papers we cite, which provide the context of the debate or the evidences on which our views are based.

Krippendorff's Arbitrary Redefinition of Intercoder Reliability

Krippendorff disagreed with Feng's (2015, p. 13) statement that "intercoder reliability assesses the degree of agreement… in the ratings given by judges…" Feng (2015) did not invent this view of reliability. Many scholars (e.g., Freelon, 2010; Hruschka et al., 2004; Zhao, Liu, & Deng, 2013) share our views. Krippendorff once considered *intercoder reliability* and *intercoder agreement* synonyms. Feng (2015) never said "reliability is a measure of agreement among coders," which Krippendorff (2016) refutes at length. What Feng (2015) meant by "assess" is that intercoder reliability coefficients indicate agreement between coders. Interestingly, it was Krippendorff (2011b) who used the word "measure" to define α as "a reliability coefficient developed to measure the agreement among observers, coders, judges, and raters." To be precise, we prefer "assess" and "evaluate" over "measure" in this context.

Krippendorff frequently and arbitrarily changed the definition of reliability. Krippendorff (2011a) defined reliability to be "the extent to which different methods, research results, or people arrive at the same interpretations or fact." Krippendorff (2016, p. X), however, redefined reliability as "the ability to rely on data that are generated to analyze phenomena a researcher intends to study, theorize, or use in pursuit of practical decision." Krippendorff (2016, p. Y) further said "the ability to rely on data is independent of whether the data result from mechanical measuring devices, or generated by human coders." Krippendorff does not say whose "ability" he refers, which makes the statements confusing. Literally, reliability is the ability to be relied on (Oxford University Press, 2016), not "the ability to rely on" which Krippendorff asserts. What we care about is specifically coders' ability to give consistent results, instead of researchers' ability to rely on coded data in general, as inter*coder* reliability is mainly about coders, not about researchers.

For instance, $\alpha = 0.9$ may indicate high reliability; an experienced researcher would still be more capable of utilizing the data than an inexperienced one. Moreover, in content analysis, coders' ability to give consistent results is affected by contextual, sociocultural, and linguistic factors, making some tasks more difficult than others. Furthermore, coders' abilities vary significantly when coding task is culturally bound or otherwise complex, even with the best training and instruction. When a task is extremely difficult, almost every coder is forced to guess randomly, producing high chance agreement; when a task is extremely easy, almost every coder codes correctly, producing high true agreement. In either scenario it would be difficult to differentiate individuals' abilities. While Item Response Theory (IRT) attempts to address these issues, it also challenges the assumption of interchangeable coders underlying Krippendorff's α.

Krippendorff's Denial of Influence of Coding Difficulty on Chance Agreement

Contrary to Krippendorff (2016) claim, Feng (2015) did not equate reliability or lack thereof with coders' ease or

Methodology (2016), *12*(4), 145–148
DOI: 10.1027/1614-2241/a000120

difficulty of categorizing units of analysis. We discovered that the chance agreement of Krippendorff's α is affected by the difficulty level of coding tasks (Feng, 2013; Zhao, 2012; Zhao et al., 2013).

Krippendorff (2016) denies that coding difficulty has anything to do with the (researcher's) ability to rely on data. Nevertheless, we are discussing intercoder reliability. Krippendorff (2011a) asserts that the number of values or categories available for coding implies the difficulty level, which contradicts findings from a controlled experiment and a simulation study, that the number of categories is uncorrelated with difficulty (Feng, 2013; Zhao, 2012; Zhao et al., 2013). Both studies found negative correlations between difficulty level and α-estimated chance agreement, indicating that α-estimated chance agreement is not really chance agreement, but more distribution skewness.

One possible solution is to replace the classic test theory (CTT), upon which Krippendorff's α is built, with IRT, which is a model for the association between a subject's response and the underlying latent ability, and can explicitly take into account coding difficulty (Feng, 2014; Reeve, 2002). Some scholars have explicitly taken difficulty into account, using either the indexing or modeling approach to estimate intercoder agreement (e.g., Aickin, 1990; Gwet, 2010; Schuster & Smith, 2002).

Krippendorff (2016) accused Feng (2015) for not providing evidence for the correlation between AC_1's chance agreement and coding difficulty, and not revealing what is meant by "normal" relationship between difficulty and chance agreement. In fact, Feng (2015) clearly referenced Feng (2013), which discussed the relationship extensively and illustrated it with several tables. The chance agreement of a normally distributed agreement index should be positively correlated with coding difficulty, which indicates exactly what "normal" means.

The chance agreement of AC_1 has a strong positive correlation with coding difficulty, whereas chance agreement of Krippendorff's α has a strong and negative correlation with coding difficulty. That is, Krippendorff's α assumes that the more difficult the coding task is, the less likely people code randomly (i.e., by guessing). Coding difficulty is the underlying factor, which causes the paradox of "high agreement, but low α."

Krippendorff's Comments on Other Indices

Krippendorff (2016) refutes Feng's (2015) defense for Gwet's AC_1. Krippendorff (2016) mentioned that AC_1 is not a chance-corrected agreement coefficient unless $D_e = P_e$. Gwet (2008, 2010), however, argued that the chance agreement of Scott's π (1955), which is defended by Krippendorff, represents chance agreement only under the unrealistic assumption of independence of all ratings. This is in fact a direct application of the well-known special multiplication rule of probability. In case of independence of the ratings, chance agreement is given by

$$P_e = 1 - 2\pi_1(1-\pi_1) = 0.5 = 1 - P_e = D_e, \quad (1)$$

where π_1 is the mean value of the two marginal proportions associated with category 1, for two categories and two raters. The chance agreement of AC_1 is,

$$\begin{aligned} P_e(AC_1) &= 1 - P_e(\text{Scott's } \pi) \\ &= 1 - (1 - 2\pi_1(1-\pi_1)) = 2\pi_1(1-\pi_1) \\ &= \left(\frac{1}{2}\right) \times \left(\frac{\pi_1(1-\pi_1)}{\frac{1}{4}}\right), \end{aligned} \quad (2)$$

whether there is independence of ratings (Gwet, 2008, 2010). In the first pair of parentheses in the last part of Equation 2, 1/2 represents the probability that the two raters agree of either one of the two categories 1 or 2 they score the subjects randomly (i.e., 1/2 = 1/4 + 1/4). In the second pair, it measures the likelihood for a rater to perform a random rating based on observed data; It is the ratio of the variance of the dichotomous category-1 membership variable to its maximum value of 1/4 (Gwet, 2008, 2010). The assumption here is that the more difficult the rating, the higher this variance, and the more likely the raters to perform random rating (Gwet, 2016). Gwet's explanation coupled with empirical evidences (Feng, 2013) are sufficient to dispel Krippendorff's query.

Krippendorff (2016) also accuses Feng (2015) of not "offering quantitative criteria" when recommending percent agreement. In fact, Feng (2013) gives a detailed rationale. In Table 4 of Feng (2013), when coding tasks are easy, the percent agreement has correlations of more than 0.9 with S, I^r, and AC_1, which are so-called chance-corrected agreement coefficients. That is, there is very little chance agreement removed from percent agreement for easy coding tasks. Therefore, percent agreement is better than it is thought to be, especially when coding tasks are easy.

Krippendorff (2016) objects to most of the results in Table 2 of Feng (2015). Feng never claimed they are the same indices, but only reported some indices which often produce very similar reliability scores. Therefore, although Krippendorff's α differs from Scott's π or Fleiss' π in terms of mathematical formula, the three give almost the same estimations for two coders, nominal data, and a sample size above 20, according to Zhao et al. (2013).

We mentioned that Krippendorff's α is equivalent to the intra-class correlation (ICC 1) for continuous data, rather than interclass correlation (e.g., Pearson correlation) as Krippendorff (2016) mentioned. Krippendorff (2016) said α is "incidentally" equal to CCC for interval data, which is not true. CCC is a complex index. Krippendorff's α sometimes equals CCC for interval data but always equals ICC 1. Table 2 summarizes the equivalence among existing indices of agreement. If there are no comparable indices at a certain measurement level, they are not listed. As a result, Feng (2015) used *continuous scale* to represent both interval and ratio scales, while Krippendorff differentiates interval scales from ratio scales when estimating agreement.

Krippendorff's Information Adequacy Theory Is Inadequate

Krippendorff (2016) underscores the importance of sufficient variance (or information) in the data to estimate reliability using his index. Contrary to Krippendorff claim, Feng never assumed that adequate sample sizes would eliminate the high-agreement-low-α paradox. Krippendorff (2011a) proposed a procedure to estimate the information adequacy as a prerequisite to further test intercoder reliability. Feng (2013, pp. 2–3) demonstrated that the paradox persists, when information is adequate by Krippendoff's criteria.

Why α Punishing Larger Sample and Rewarding Smaller Sample?

Krippendorff correctly noted two numerical errors in Table 3 of Feng (2015). Cohen's κ and Scott's π for *study types* should be 0.44 and 0.43, respectively. We take this opportunity to offer a corrigendum. The errors result from a bug in the third-party package, *rmac* of R language, which had produced correct results in the past but now we found it errs when the number of categories is high and class attribute is factor. We thank Krippendorff for catching the errors.

When pointing out the technical errors, Krippendorff (2016) noted $\alpha > \pi$ when sample is small. Gwet (2016) found that the difference between Krippendorff's α and Scott's π was $(1 - \pi)/(2n)$, with n being the sample size. If π is 0.85 or higher, the two coefficients would be almost identical even for very small samples. But even for small values of Scott's π, a sample of moderate size such as 25 will lead to similar values for both coefficients (Gwet, 2016).

In addition, both Gwet (2016) and Zhao et al. (2013) found that α, and only α, varies drastically with the variation in sample size. That is, everything else being equal, increasing sample size decreases α, which Scott's π does not do. This is important, because the main justification that Krippendorff (1970a, 1970b) offered for α, which is mathematically a small variation of Scott's π, was that α adjusts for sample size while π does not. Forty plus years later, we discovered that so-called "adjust" actually means punishing larger samples and rewarding smaller samples (Zhao et al., 2013).

Concluding Remarks

Krippendorff (2016) opened his critique with the statement "I was surprised that *Methodology* published Feng's paper." We have seen similar reactions from him (see Krippendorff, 2004, 2013; Krippendorff, 2016) when other groups discussed or demonstrated paradoxes, abnormalities, or imperfections of α (Lombard, Snyder Duch, & Bracken, 2002; Zhao et al., 2013). We have seen stronger reactions in reviews for publication.

The reactions are understandable, given that Krippendorff and colleagues argued that α is superior to all competing indices of intercoder reliability, and α is the only standard measure that meets all needs of assessing reliability (Krippendorff, 1970a, 1970b, 1980, 2004; c.f. Hayes & Krippendorff, 2007). We take no pleasure in encountering evidences contradicting an esteemed scholar who did so much to introduce content analysis and intercoder reliability to generations of students, including us.

Nevertheless, we plead with supporters of α to see that their actions foster a spiral of inertia in the developments of reliability concepts and techniques (Noelle-Neumann, 1974; Zhao et al., 2016). The selective barriers to publication discourage dissents, creating a perception of perfection of the status quo among researchers, reviewers, and editors, especially in some disciplines that rely on reliability (Zhao & Fung, 2014). The spiraled perception is a misperception, according to the empirical evidences that we will continue to push to publish (Zhao, 2012).

We hope Krippendorff will see that the legacy of Krippendorff is not dependent on the perfection of α. In the decades-long fight for α, he successfully persuaded generations of researchers, including us, of the imperativeness of measuring intercoder reliability and the importance of removing chance agreement. These are broad and fundamental contributions that no one should or can take away regardless of the fate of α. In that sense, any index of intercoder reliability that will be invented or adopted will have a chunk of Krippendorff in it.

References

Aickin, M. (1990). Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's kappa. *Biometrics, 46*, 293–302.

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

Feng, G. C. (2014). Estimating intercoder reliability: A structural equation modeling approach. *Quality & Quantity, 48*, 2355–2369. doi: 10.1007/s11135-014-0034-7

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

Freelon, D. G. (2010). Recal: Intercoder reliability calculation as a web service. *International Journal of Internet Science*, *5*, 20–33. Retrieved from http://www.ijis.net/ijis5_1/ijis5_1_freelon.pdf

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology, 61*, 29–48. doi: 10.1348/000711006X126600

Gwet, K. L. (2010). *Handbook of inter-rater reliability-a definitive guide to measuring the extent of agreement among multiple raters*. Gaithersburg, MD: Advanced Analytics, LLC.

Gwet, K. L. (2016). On Krippendorff's comments to Feng's article entitled "mistakes and how to avoid mistakes in using intercoder reliability indices". Personal communication.

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

Hruschka, D. J., Schwartz, D., John, D. C. S., Picone-Decaro, E., Jenkins, R. A., & Carey, J. W. (2004). Reliability in coding open-ended data: Lessons learned from HIV behavioral research. *Field Methods, 16*, 307–331. doi: 10.1177/1525822X04266540

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

Krippendorff, K. (1970b). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement, 30*, 61–70. doi: 10.1177/001316447003000105

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. London, UK: Sage Publications.

Krippendorff, K. (2004). Reliability in content analysis. Some common misconceptions and recommendations. *Human Communication Research, 30*, 411–433. doi: 10.1111/j.1468-2958.2004.tb00738.x

Krippendorff, K. (2011a). Agreement and information in the reliability of coding. *Communication Methods and Measures, 5*, 93–112. doi: 10.1080/19312458.2011.568376

Krippendorff, K. (2011b). Computing Krippendorff's alpha reliability. *Departmental Papers (ASC)* 43. Retrieved from http://repository.upenn.edu/cgi/viewcontent.cgi?article=1043&context=asc_papers

Krippendorff, K. (2013). Commentary: A dissenting view on so-called paradoxes of reliability coefficients. *Annals of the International Communication Association, 36*, 481–499. doi: 10.1080/23808985.2013.11679143

Krippendorff, K. (2016). Misunderstanding reliability. *Methodology. European Journal of Research Methods for the Behavioral and Social Sciences, 12*, 139–144. doi: 10.1027/1614-2241/a000119

Lombard, M., Snyder Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research, 28*, 587–604. doi: 10.1111/j.1468-2958.2002.tb00826.x

Noelle-Neumann, E. (1974). The spiral of silence a theory of public opinion. *Journal of Communication, 24*, 43–51. doi: 10.1111/j.1460-2466.1974.tb00367.x

Oxford University Press. (Eds.). (2016). *Reliability*. Oxford, UK: Oxford University Press.

Reeve, B. B. (2002). An introduction to modern measurement theory. *National Cancer Institute*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.207.3244&rep=rep1&type=pdf

Schuster, C., & Smith, D. A. (2002). Indexing systematic rater agreement with a latent-class model. *Psychological Methods, 7*, 384–395. doi: 10.1037/1082-989X.7.3.384

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly, 19*, 321–325. doi: 10.1086/266577

Zhao, X. (2012, August). *A reliability index (AI) that assumes honest coders and variable randomness*. Paper presented at the Association for Education in Journalism and Mass Communication, Chicago, IL.

Zhao, X., Liu, J. S., & Deng, K. (2013). Assumptions behind intercoder reliability indices. In C. T. Salmon (Ed.), *Communication yearbook* (Vol. 36, pp. 419–480). New York, NY: Routledge.

Zhao, X., Lu, H. H., Huang, D., Liao, S., Liao, Z. V., Ng, Y. L., … Zhao, E. Z. S. (2016, July). *Spiral of just enough information-inverted u effect of title length on online read and relay*. Paper presented at the IAMCR Conference, University of Leicester, Leicester, UK.

Zhao, X., & Fung, T. (2014). Development in research on political communication. In J. Hong (Ed.), *New trends in communication studies* (pp. 485–507). Beijing, China: Tsinghua University Press.

Guangchao Charles Feng
School of Journalism and Communication
Jinan University
Guangzhou
China
fffchao@gmail.com

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

