Cleveland State University
EngagedScholarship@CSU

MSL
ACADEMIC ENDEAVORS

Communication Faculty Publications

School of Communication

7-1-2004

A Call for Standardization in Content Analysis Reliability

Matthew Lombard

Jennifer Snyder-Duch

Cheryl C. Bracken
Cleveland State University, c.bracken@csuohio.edu

Recommended Citation

A Call for Standardization in Content Analysis Reliability

MATTHEW LOMBARD

JENNIFER SNYDER-DUCH

CHERYL CAMPANELLA BRACKEN

In Lombard, Snyder-Duch, and Bracken (2002) we reviewed the literature regarding a wide variety of issues related to intercoder reliability in content analysis, presented a study that demonstrated the continuing lack of careful reporting of reliability in mass communication, and concluded with guidelines designed to encourage and enable researchers to follow generally accepted systematic procedures for assessing reliability and clearly explaining what was done and found in all research reports. Our online supplement to the article (2003) offered additional detailed, practical guidance to content analysis researchers on this topic.

In his article in this issue of *HCR*, Krippendorff has provided a valuable extension to one aspect of the literature on reliability by questioning the received wisdom regarding the mathematical attributes of some commonly used reliability indices. We use the limited space allotted to us to clarify a few important issues he raised, as part of a larger effort to encourage the community of content analysis researchers in communication to continue to refine, and apply, appropriate and practical standards for the assessment and reporting of intercoder reliability.

Our work joins that of a growing number of communication researchers trying to improve the quality of content analyses and to make the

method more accessible. Scholars interested in conducting content analysis research can consult recent accessible and practical books by Neuendorf (2002) and Riffe, Lacy, and Fico (1998); articles by Potter and Levine-Donnerstein (1999) and Riffe and Freitag (1997); and online resources such as *The Content Analysis Guidebook Online* (2002) by Kimberly Neuendorf and the *Content Analysis Resources* (n.d.) website maintained by William Evans. These resources—and other communication scholars—have reached consensus on the importance of intercoder reliability and on many aspects of how researchers should assess and report it. Several issues, however, remain.

Which Index/Indices?

The key issue that Krippendorff raised concerns which index or indices researchers should use. While echoing criticism of percent agreement (despite its intuitive appeal and the ease of calculating it) and praising Krippendorff's alpha (despite its mathematical density and the difficulty of calculating it), we have recommended only that researchers "[c]hoose one or more appropriate indices of intercoder reliability based on the characteristics of the variables, including their level(s) of measurement, expected distributions across coding categories, and the number of coders" and "[b]e prepared to justify and explain the selection of the index or indices" (p. 600). Krippendorff argued that "when it comes to discussing mathematical objects such as agreement measures and their use as indices of the reliability of data, mathematical proofs and demonstrations should speak louder than majority opinions, even when published" (2004, p. 430); however, the majority of opinions on this topic in most cases come from mathematical experts like himself, leaving the researcher unsure what to do. The next step in refining guidance for content analysis researchers would be to produce an expert consensus regarding the precise contexts, if any, which call for the use of each available index.

How to Assess Low Variance Data?

Given the diversity of indices available and debate regarding their specific mathematical properties, we follow others in characterizing those that do not account for chance agreement and tend to produce inflated estimates of agreement as liberal, and those that do account for this agreement as conservative. Rather than recommending that researchers "compute several agreement coefficients and . . . find a balance between conservative and liberal coefficients" (2004, p. 422) as Krippendorff suggested, we have recommended they "[establish] a decision rule [for acceptable coefficient values] that takes into account the assumptions and/or weaknesses of each" index used and "be prepared to justify the criterion or criteria used" (p. 600). We selected the indices and decision rules for our

study because of the context of our particular dataset. Based on the literature and extensive pretesting to refine the coding instrument, we expected many variables to contain little variance, and the researchers (the second and third authors) served as the coders. We knew that our reliability and other coding data were not "obtained by broken instruments or by coders who fell asleep or agreed in advance of the coding effort to make their task easy" (Krippendorff, 2004, p. 425). We are confident in the limited conclusions we reached (i.e., that the characteristics examined in the low variance variables are in fact rare); however, researchers need more specific guidelines regarding how to assess reliability in such situations.

Effects of Sample Size?

In our reliability dataset, differences among reliability coefficients (excluding cases where the coefficient would normally be inappropriate because of the variable's level of measurement) were very minor (in the third decimal place). Krippendorff cautioned that smaller samples could result in larger differences. Researchers should always follow recommendations (e.g., Lacy & Rife, 1996) to obtain appropriately large reliability samples; however, it would help researchers to know just how big the differences in coefficients are likely to be with different sample sizes.

Overlapping Coders?

Some experts (e.g., Potter & Levine-Donnerstein, 1999, and Neuendorf, 2002) have cautioned against having different coders code different subsets of a reliability dataset. This was the basis of our recommendation to avoid that approach when possible, although we noted that many indices are designed to accommodate the missing data it produces. On the other hand, Krippendorff encouraged the use of overlapping coding "provided there is enough duplication or overlap" (2004, p. 428). Researchers need to resolve the disagreement regarding the appropriateness of this approach and, if it is endorsed, provide guidance regarding the minimum acceptable level of duplication.

Software Solutions?

Aside from standardized, practical guidelines regarding reliability, researchers need easy-to-get, easy-to-use software tools to calculate it. We hope more researchers follow the lead of Kang, Kara, Laskey, and Seaton (1993) and Skymeg Software in developing the specialized freeware software tools content analysts need. Unfortunately, we know of no publicly available program that calculates the family of Krippendorff's alpha indices, but as we note in the online supplement to our article, the latest version (0.4.5) of PRAM (Skymeg Software, 2004), far from being "faulty,"

calculates several indices and provides useful coder-pair and average index values for diagnosing reliability problems (not, as Krippendorff suggested, for reporting purposes).

CONCLUSION

We hope unresolved or nonexistent disagreements do not distract readers from the larger issues here—the critical importance of intercoder reliability and the need to assess and report it properly. Further, we encourage researchers and other experts interested in content analysis to work together to refine the guidelines we proposed and develop new software tools so that we all can produce the best possible communication research. We will continue to report on developments on this topic at http://www.temple.edu/mmc/reliability (Lombard, Snyder-Duch, & Bracken, 2003).

REFERENCES

Evans, W. (*n.d.*). *Content analysis resources.* Retrieved from http://www.car.ua.edu/

Kang, N., Kara, A., Laskey, H. A., & Seaton, F. B. (1993). A SAS MACRO for calculating intercoder agreement in content analysis. *Journal of Advertising, 23*(2), 17–28.

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

Lacy, S., & Riffe, D. (1996). Sampling error and selecting intercoder reliability samples for nominal content categories: Sins of omission and commission in mass communication quantitative research. *Journalism & Mass Communication Quarterly, 73,* 969–973.

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.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2003). *Practical resources for assessing and reporting intercoder reliability in content analysis research projects.* Retrieved from http://www.temple.edu/mmc/reliability/

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

Neuendorf, K. A. (2002). *The content analysis guidebook online: An accompaniment to* The Content Analysis Guidebook *by Kimberly A. Neuendorf.* Retrieved from http://academic.csuohio.edu/kneuendorf/content/

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

Riffe, D., & Freitag, A. A. (1997). A content analysis of content analyses: Twenty-five years of *Journalism & Mass Communication Quarterly. Journalism & Mass Communication Quarterly, 74*(4), 873–882.

Riffe, D., Lacy, S., & Fico, F. G. (1998). *Analyzing media messages: Using quantitative content analysis in research.* Mahwah, NJ: Erlbaum.

Skymeg Software (2004). Program for reliability assessment with multiple coders (PRAM). [Computer software]. Retrieved from http://www.geocities.com/skymegsoftware/pram.html