8-2014

Assessing the Reporting of Reliability in Published Content Analyses: 1985–2010

Jennette Lovejoy
University of Portland, lovejoy@up.edu

Brendan R. Watson

Stephan Lacy

Daniel Riffe

Assessing the Reporting of Reliability in Published Content Analyses: 1985-2010

The reliability of a content analysis protocol is a precondition for its validity. Content analysis reliability testing, typically in the form of a reliability coefficient, is employed to provide an estimate of the error introduced by inconsistent coding. A protocol that is not 100% reliable generates data with some level of measurement error. Absent reliable and valid measurement of concepts, generation of cumulative knowledge based on those concepts is impossible. Thus, the most important part of conducting a content analysis is establishing the reliability of the coding protocol, or the decision rules, for assigning values to content.

But if scholars agree in principle that failure to assess reliability is a "fatal flaw," in practice many published content analyses have not reported an assessment of reliability. Riffe and Freitag (1997) found that about half (56%) of content analyses in *Journalism & Mass Communication Quarterly* between 1971 and 1995 reported reliability results. Pasadeos, Huhman, Standley, and Wilson (1995) found similar results (49%) in content analyses of news media content in four journals for 1988-1993. Lombard, Snyder-Duch, and Bracken (2002) and Snyder-Duch, Bracken, and Lombard (2001) examined 200 content analyses indexed in *Communication Abstracts* between 1994 and 1998, finding that 69% of the articles reported an assessment of reliability, though only 41% reported reliability for each variable. More recently, Riffe, Lacy, and Fico (2005) found 74% of 80 *Journalism & Mass Communication Quarterly* content analyses between 1998 and 2004 reported reliability.

While these studies indicate a growing trend in the reporting of reliability testing, whether their findings reflect a trend toward *more transparent* reporting of reliability is

an empirical question, and an important one. Method transparency is essential to the reproducibility of the reliability test which, again, is integral to assessing the overall quality of content analysis findings. Thus, the present study of reliability reporting is based on representative samples of issues of three flagship journals of major communication research associations: *Communication Monographs,* published by the National Communication Association; *Journal of Communication,* published by the International Communication Association; and *Journalism & Mass Communication Quarterly,* published by the Association for Education in Journalism and Mass Communication. The study covers a 26-year period (1985-2010) and examines not only changes over time in whether a reliability test was conducted but also whether a reliability coefficient correcting for chance was used and how well authors reported two key decisions in reliability testing: how content units were sampled for reliability testing and how many content units were sampled.

This study is founded on two assumptions. First, content analysis is a social science research method. Second, as such, the reporting of content analyses must be transparent enough to allow for replication and for other scholars to adequately evaluate data reliability. Just as the reliability of psychometric scales is reported and scale items made available for replication, the reliability of all content analysis variables should be reported and the protocols made available.

When and How to Select Content Units

After developing the coding protocol and selecting the content to be analyzed, the next step in the content analysis process is to decide how many units from the study to select for the reliability test and how to select those units. One problem facing researchers

in making these two decisions is that there is limited empirical justification presented by scholars who advocate various reliability sampling guidelines. Some scholars advocate that independent samples of non-study content be drawn for some tests; others specify coding of "overlap," i.e., a subset of the study content that is coded by all coders (Potter & Levine-Donnerstein, 1999). Wimmer and Dominick (2003) advise that between 10% and 25% of the population of content be tested, while Kaid and Wadsworth (1989) suggest 5%-7%. One online resource (Lombard, Snyder-Duch, & Bracken, 2010) prescribes that the reliability sample "should not be less than 50 units or 10% of the full sample, and it rarely will need to be greater than 300 units; larger reliability samples are required when the full sample is large and/or when the expected reliability level is low." Lombard et al. (2010) also suggest that reliability be tested informally during coder training with units that are *not* drawn from the study units, that this informal process should be repeated until results suggest "an adequate level of agreement," and that reliability be assessed formally during a pilot test, using a "representative sample" ("using a random or other justifiable procedure") of at least 30 units.

The sample of study units used to test reliability should be representative. Representativeness, though, could mean that the sample is representative of the population of content being studied (Riffe et al., 2005; Scott, 1955), and it could also mean that the sample adequately represents the full range of categories used in the coding protocol (Krippendorff, 2011). Although both types of representativeness are important, the latter is of particular concern with rarely occurring categories. If, for example, a researcher is coding the frequency with which TV news covers particular crimes, including incidents of domestic terrorism, a probability sample that is otherwise of

sufficient size might include only one such domestic terrorism incident because of the relative infrequency of such crimes in the population of newscasts. However, if coders have to make only one decision as to the presence/absence of a domestic terror incident, this sample of one does not adequately test the reliability of coders applying this infrequent category. Krippendorff (2011) suggests that in such instances, the researcher needs to draw a stratified random sample of units in each coding category to adequately test the reliability of the coding protocol. Still, ensuring representativeness, unless one uses all study units in the reliability check, requires some form of probability sample.

Reproducibility also requires that if the study uses something other than a simple-random sample, then the process for drawing that probability sample must be explained by the study's authors (e.g., how was a sample stratified).

Previous reviews of content analysis practices, however, have found that random selection for reliability testing is relatively rare. Riffe et al. (2005) reported that their study of 80 quantitative content analyses published between 1998 and 2004 found only one in three used random sampling for reliability testing. Further, only 16% met "standards" for reliability tests, which included random selection of units for reliability tests, using coefficients that consider chance agreement, and reporting coefficients for each focal variable.

In terms of the number of units sampled for the reliability test, two sets of guidelines have been suggested, both based on random selection of the reliability sample from the population of content units being studied. First, Krippendorff (2013) argues that the sample size "is related to the proportion of units in different categories" (p. 322). He also says that the sample size is related to the number of coders involved. Using the work

of Bloch and Kraemer (1989), Krippendorff (2013) presents a table (p. 323) that provides reliability sample size on the basis of the smallest value of Krippendorff's alpha that is acceptable, the *a priori* significance or *p*-value set by the researcher, the probability of the least frequent value among all the population values, and the number of coders.

Lacy and Riffe (1996) took a somewhat different approach in recommending sample sizes for reliability checks with nominal-level variables, adapting a formula developed by Schutz (1952) that allowed a scholar to compensate for chance agreement through sampling once an acceptable reliability level had been decided. Riffe et al. (2005, p. 146-147) provide tables (based on Lacy & Riffe, 1996) for selecting reliability samples on the basis of three factors that can vary from study to study: acceptable level of probability for inference, assumed level of simple agreement in the population, and population size.

While the above studies suggest some disagreements over the exact method of selecting a reliability sample, the scientific method requires transparency in reporting how study units were sampled and the number of units used. Replication can only occur when study procedures are explicitly provided.

Reporting of Reliability Coefficients

After selecting a sample of study units for the reliability test and using the protocol to code those units, the next step is to choose a reliability coefficient to indicate the reliability of the coding categories. Reliability coefficients summarize the extent to which multiple coding operations classify the same content units into the same categories. Krippendorff (1980) characterizes reliable measurement as a process involving stability (yielding consistent results across time), reproducibility (yielding the

same results across coders or raters), and accuracy (the process produces results that "conform to a known standard," p. 131). This study will examine reproducibility reliability, but it will not analyze which of the reliability coefficients available to scholars is reported. Rather it will examine whether the reliability coefficient reported in an article considers chance agreement and whether reliability coefficients are reported for each variable.

Coefficients Correcting for Chance Agreement

There are two primary types of reliability coefficients: those that report simple percentages of observed agreement (e.g., 85%) and those that take into account the possibility of chance agreements (e.g., Scott's pi, Cohen's kappa, Krippendorff's alpha, Zhao's $Alpha_i$). Early, "classic" content analysis texts primarily advocated using either percentage of agreement or correlation coefficients, such as Pearson's *r* (Berelson, 1952; Holsti, 1969; Stempel, 1955; Stempel, 2003). Most contemporary content analysis scholars, however, recognize that agreements can occur by chance, even between untrained coders not employing coding rules. Thus, contemporary standards for reporting coding reliability recognize that percentage agreement is insufficient and that one must report a coefficient that takes chance into consideration (Krippendorff, 2004b; Lombard et al., 2010; Neuendorf, 2002; and Riffe et al., 2005). These coefficients create a scale with zero representing chance agreement, which means no statistical relationship between the nature of content coded and values assigned by coders. Riffe et al. (2005) noted that only 46% of the 1998-2004 content analysis articles they examined reported coefficients that account for chance agreement.

A number of chance-correcting coefficients are available. Proponents of particular coefficients base their positions on conceptual arguments, including how estimates of chance agreement are calculated (Gwet, 2008; Krippendorff, 2012; Krippendorff, 2004a; Lombard, Snyder-Duch, & Bracken, 2004; Zhao, 2012; Zhao, Liu, & Deng, 2012). This is not an arcane dispute; there are important consequences of how chance agreement is computed, particularly when a variable's distribution is highly "imbalanced" (e.g., 97% to 3%), as illustrated lucidly by Potter and Levine-Donnerstein (1999). This study does not explore which of the available coefficients is "best." It argues that at least some form of coefficient correcting for chance should be reported and examines if this has indeed been the case.

Number of Coefficients reported

In addition to using a coefficient that corrects for chance, this study takes the position that reliability should be reported for each variable coded in a study (Riffe, Lacy, & Fico, 2014). Yet published articles sometimes report what authors call an "overall" reliability score. The problem with an "overall" or average score is that content analyses often have "easy" variables that involve the administrative task of recording explicit information, rather than requiring coder judgment in applying the protocol. For example, recording the newspaper in which a given story ran, or the date a news segment was broadcast. These "easy" variables can inflate the average coefficient reported, masking the true reliability – or lack thereof – of problematic variables. Content analyses should always report a reliability coefficient for each variable, even if some values are less than acceptable and need to be dropped from the analysis; such reports allow readers to

evaluate the data and, perhaps, help improve weak measures in future studies (Riffe et al., 2014).

Previous studies of content analyses, however, have found that authors frequently have not reported reliability coefficients for each variable. Snyder-Duch et al. (2001) found that 69% of content analyses they studied reported results of a reliability test, but only 41% provided coefficients for *all* the individual variables. Four years later, Riffe et al. (2005) found 74% of their studied articles reported reliability, but only 54% reported on all the individual variables. Thus, this study examines whether the proportion of articles reporting reliability coefficients for each variable has continued to improve.

In sum, in order to allow for necessary judgment of the adequacy and reproducibility of a given content analysis protocol, studies should, at minimum: report the number of sample units used in the reliability test; use some form of probability sample to select those units; describe how the probability sample was drawn (e.g., every n^{th} unit, stratified sample, etc.); report reliability coefficients that account for chance agreement; and report such reliability coefficient for each variable in the protocol. This study examines how well published content analyses adhere to these suggested standards, and changes that have taken place in reliability reporting practices over time.

Method

The Sample

We systematically examined representative samples of issues from three major communication associations' flagship research journals drawn from a 26-year period (1985-2010): *Communication Monographs (CM)*, published on behalf of the National Communication Association; *Journal of Communication (JoC)*, published for the

International Communication Association; and *Journalism & Mass Communication Quarterly (JMCQ),* a journal of the Association for Education in Journalism and Mass Communication. There are other communication journals, of course, but as the flagship journals of the field's major communication associations, these three constitute a reasonable barometer of the field's research practices. All are published quarterly and all publish multiple articles in each issue.

The years 1985 to 2010 were selected to represent the time period when three major content analysis texts (Krippendorff, 1980, 2004b; Neuendorf, 2002; Riffe, Lacy, & Fico, 1998, 2005) encouraged the use of reliability coefficients that considered chance agreement. The first edition of Krippendorff's text was published in 1980, and 1985 was selected because of the lag time needed for its adoption and use in graduate study. The 2010 endpoint was selected because it was five years after the second edition of the Riffe et al. (2005) text was published. Krippendorff's second edition was published in 2004.

Preliminary examination of past journal issues indicated that *JMCQ* traditionally carried more content analysis articles than *CM* or *JoC*. We included all of the 104 issues for *CM* and *JoC* published during the period, but randomly selected two *JMCQ* issues per year, for a total of 52 issues. The resulting distribution of articles among the three journals confirmed the assumption that *JMCQ* would have more content analysis articles. Each issue was examined by two coders to identify content analysis articles that met three standards:

1. At least some of the data analyzed were obtained by examining existing content (mediated or interpersonal) or content created specifically in response to experimental stimuli.

2. The content must be divided into discrete measurement units in order to assign numbers to the units for quantitative analysis (i.e., a historical, legal, or qualitative study, or essay, based on a reading of all texts that include a key term, is not a content analysis, even if the population of texts was "filtered" for presence of that key term).

3. The content analysis data do not have to have been collected by author(s) for the article to count as a content analysis article. Secondary analysis of previously collected content analysis data would qualify the article as content analysis.

Reliability of identification and inclusion of relevant articles was tested. For example, the 52 issues of *JMCQ* included 919 total articles. Two trained graduate student coders identified 306 content analysis articles across the 52 *JMCQ* issues, Krippendorff's alpha equaled 0.90. Krippendorff's alpha equaled 0.89 for *CM'*s 130 content analyses (from 1,008 total articles), and 0.93 for *JoC'*s 145 content analyses (from 617 total articles).

Protocol Development and Reliability

Among the reasons for conducting this study are previous observations about the lack of clear and explicit information in reporting on reliability testing of coding protocols in published content analyses (Lombard et al., 2002; Pasadeos et al., 1995; Riffe et al., 2005; Riffe & Freitag, 1997; Snyder-Duch et al., 2001). Because of this tendency, it was essential that this study use formally trained, experienced content analysts as coders. While it may be preferable in many instances to establish reliability using independent coders and even multiple sets of coders each time significant revisions

are made to the coding protocol, this study's purpose, the technical nature of the content being coded, and the lack of significant funding rendered such independence impractical. Nonetheless, we took several steps in order to attain a degree of coder independence in testing the coding protocol: two of the study's authors were responsible for designing the protocol, and while one of those two was involved in coder training and pre-tests of the protocol, two additional authors who did not design the protocol tested its reliability and subsequently carried out the study's main coding.

The variable definitions and the coding protocol (available upon request) were refined through several rounds of training, practice sessions, and test coding by three of the authors on randomly selected articles *not* in the study sample but drawn from the three selected journals. These practice rounds resulted in protocol modifications and refinement of focal variables. After the training rounds, two of the same coders were involved in three pilot coding checks, again using articles *not* in the sample (n = 20, 17, and 10 articles, respectively). Simple agreement was greater than 82% for all variables for the three pilot tests, which was judged sufficient to begin coding.

The final sample of 581 articles (*JMCQ* = 306; *JoC* = 145; *CM* = 130) was assigned to two of the authors who had performed the pilot tests. These authors served as main coders, and each was randomly assigned half the sampled units from each of the three journals. To assess intercoder reliability of the protocol for the main coding of the study sample, a subset of 86 study articles was randomly chosen (*JMCQ* = 44; *JoC* = 23; *CM* = 19) and double-coded by both coders. The 86 articles were determined using the Lacy and Riffe (1996) formula, assuming 95% probability and a 90% agreement level in

the population. The 86 articles represented all values for all variables used in the study. Final intercoder reliabilities and focal variables are listed below:

Reliability Check Reported (Krippendorff's alpha = 0.84): Was a reliability check reported? Coders coded either "yes" or "no."

Sample Size (Krippendorff's alpha = 0.93): Was the number of units in the reliability sample explicitly stated in text or reported in tables or notes? If a reader cannot identify the number from text, tables, or notes, then the sample size was not reported. Sample size identification by percentage ("a 10% sample") or method ("every fifth article") was coded as not reported if the total number of units being sampled was not explicitly stated.

Reliability Sampling (Krippendorff's alpha = 0.88): Did the study authors report that a probability sample (or a census) was used to select content units for testing reliability, or was a non-probability, non-census sample used? A probability sample was any sample that involved random sampling and the probability of each unit being in the sample was known. A census involved the use of all units in the study for the reliability test. Coders examined text and notes to determine that key terms were explicitly reported ("random," "randomly," "probability," "systematic random," "stratified" or "constructed-week"). If an article included a reliability check but did not report a probability sample or census, the default value was "non-probability." Examples of non-probability samples are coding the first 10 articles of a total sample; selecting 10 articles, but not explicitly specifying that probabilistic sampling was used; or reporting reliability figures, but not specifying how the reliability sample was drawn.

Sampling Method (Krippendorff's alpha = 0.90): Coders judged whether, based on described procedures, they would be able to replicate the method used in selecting the sample of content units for reliability testing. The sampling method was either explained enough for replication or not. Criteria would include reports on the portion of the study sample to be used (including a "skip interval," for example, in a systematic sample), identification of clusters or strata, and specification of levels in multi-stage sampling.

Type of Reliability Coefficient (Krippendorff's alpha = 0.87): Were coefficients that correct for chance reported in the articles (e.g., Scott's pi, Krippendorff's alpha, Cohen's Kappa, Gwet's Gamma, Benini's Beta, or Guttman's Rho), or was a measure of simple agreement used, in the "final" (non-training, non-pilot) reliability test? Coefficients had to be identified explicitly or by how they were calculated. Coefficients other than those named above were coded as "other" and correlation measures, including Pearson's *r*, were coded as "correlations."

Number of Reliability Coefficients Reported (Krippendorff's alpha = 0.86): Did articles report a coefficient that corrects for chance for each variable in the study? Studies were originally coded into three levels: 1) Studies that only reported a single or "overall" average reliability coefficient, even though there was more than one variable used in the study, 2) studies that reported more than one reliability coefficient, but not one for every variable (e.g., a range of coefficients) and 3) studies that reported at least one reliability coefficient for *every* variable either in text or an endnote. For the purposes of analyses, articles that only reported an "overall" coefficient were further recoded as not reporting reliability coefficients for all variables.

Statistical Analyses

Although the primary predictor variable was change over time, we also examined differences in reliability test reporting between journals. A series of binary logistic regression models examined the effect of publication year and journal (*CM* vs. *JoC* vs. *JMCQ*) for each of six binary dependent variables: (1) number of units in the reliability sample was reported vs. number of units was not reported (n = 441; excludes 140 articles that did not conduct a reliability check), (2) sampling process was explained vs. sampling process was not explained (n = 441), (3) probability or census used in reliability check vs. non-probability sample used (n = 438; excludes 143 articles that did not conduct a reliability check or no reliability figures were reported), (4) reliability coefficient used that considers chance vs. no coefficient reported or correlation used (n = 581), (5) reliability coefficients reported but not for each variable vs. no reliability coefficient reported (n = 464; excludes 117 articles that reported a reliability coefficient for *every* variable), and (6) reliability coefficient reported for every variable vs. reliability coefficient reported for some or no variables (n = 581).

All models treated year as a continuous variable and journal as a nominal variable. We dummy coded journal, first treating *CM* as the reference category, then re-examined all models treating *JMCQ* as the reference category in order to examine the *JMCQ* vs. *JoC* comparison. The "year-by-journal" interaction examined differential changes in reporting practices over time by journal. Preliminary visual examination of the data indicated the possible presence of curvilinear relationships between year and several dependent variables. We thus rescaled year to range from 1 (1985) to 26 (2010) to reduce rounding error when computing estimated probabilities of the dependent variables and tested for curvilinear relationships following procedures described by Aiken and West

(1991). Specifically, we examined the quadratic and cubic effects of publication year (i.e., year^2 and year^3, respectively), as well as the year^2-by-journal and year^3-by-journal interactions. Additionally, we log-transformed year using a natural log function and evaluated models that included ln(year) and the ln(year)-by-journal interaction. We selected the most parsimonious model using likelihood ratio tests for nested models and the corrected Akaike's Information Criterion and Bayesian Information Criterion for non-nested models, selecting the model with the lowest information criteria. Effect sizes are represented by odds ratios (OR) with 95% confidence intervals. Omnibus models are evaluated with Likelihood Ratio chi-square statistic, and individual model predictors are evaluated using the Wald chi-square statistic. Two-tailed tests of significance and an alpha-level of 0.05 were used for all analyses.

Results

Sample Characteristics

During the 26 years included in this study, a quarter (24%, n = 140) of the content analysis articles did not conduct a reliability check. Of those that did, 64% (n = 283) reported the number of content units used in the reliability check, 37% (n = 163) used a probability or census reliability sample, and 34% (n = 149) explicitly described the reliability sampling process. Forty-nine percent (n = 215) computed a reliability coefficient that considers chance. Twenty-seven percent (n = 117) included reliability coefficients for every study variable, 30% (n = 130) included reliability coefficients for some but not all variables, while 44% (n = 194) either reported just simple agreement only or an overall (i.e., average) reliability coefficient.

Descriptively, all three journals improved across time in reporting the number of content units in the sample, in reporting reliability coefficients that consider chance, in reporting such coefficients for some variables, and in reporting such coefficients for all variables. Two variables that showed little or no improvement were clarity or transparency in explaining how the reliability sample was selected and the use of a census or probability sample for reliability testing. The degree of change or improvement across some variables varied by journal, and these findings are explicated in the following two sections.

Reliability Sampling Procedures

Three study dependent variables characterized reliability sampling procedures: (1) whether the article made clear exactly how many units were selected for the reliability check, (2) whether the sampling procedure was sufficiently detailed and transparent, and (3) use of a probability or census sample. Figure 1 depicts the estimated probability of each of the three reliability sampling procedure dependent variables by journal over time. Probability estimates are based on models containing the year-by-journal interaction unless otherwise specified. For number of sampling units reported, the most parsimonious model contained journal and a linear effect of publication year (Likelihood Ratio $\chi^2[3] = 13.77$, $p < 0.01$; Cox & Snell $R^2 = 0.03$). As shown in Figure 1A, across all journals, the likelihood of clearly reporting the number of units used in reliability samples increased by 4% each year (OR = 1.04 [95% CI = 1.01-1.07]). Across the 26-year study interval, *CM* was more likely than *JMCQ* to publish content analyses that reported the number of sample units (OR = 1.74 [1.07-2.85]). No other journal differences were observed (*CM* vs. *JoC*: OR =1.52 [0.87-2.65]; *JoC* vs. *JMCQ*: OR = 1.15 [072-1.84). By

2010, the probability of published articles reporting the number of sampling units was similar across the three journals.

For both sampling method transparency and use of a probability or census sample, the most parsimonious models contained only journal (Likelihood Ratio $\chi^2[2] = 7.12$, $p = 0.03$, Cox & Snell $R^2 = 0.02$ and Likelihood Ratio $\chi^2[2] = 7.13$, $p = 0.03$, Cox & Snell $R^2 = 0.02$, respectively). No effect of time was observed for either of these reliability sampling variables. As shown in Figures 1B and 1C, when averaged across the entire study period, *CM* was more likely than *JMCQ* to publish content analyses that reported the sampling process (OR = 1.83 [1.14-2.92]) and used probability/census reliability samples (OR = 1.72 [1.08-2.72]). *CM* was also more likely than *JoC* to publish content analyses that reported an explicit sampling process (OR = 1.80 [1.04-3.11]) and used probability/census reliability samples (OR = 1.93 [1.12-3.32]), but did not differ from *JoC* on reporting the number of units in the sample (OR = 1.52 [0.87-2.65]). *JoC* and *JMCQ* did not differ on reporting of the sampling process (OR = 1.02 [0.62-1.67]) or use of a probability/census sample (OR = 0.89 [0.54-1.45]).

Reporting of Reliability Coefficients

Three dependent variables characterized the reporting of reliability coefficients: (1) whether the article reported use of a reliability coefficient that considers chance, (2) reporting of a reliability coefficient for at least one but not all study variables, and (3) reporting of a reliability coefficient for every study variable. For reporting of a reliability coefficient that considers chance and reporting of a reliability coefficient for at least one but not all study variables, the most parsimonious models contained the year^2-by-journal interaction (Likelihood Ratio $\chi^2[8] = 184.21$, $p < 0.001$, Cox & Snell $R^2 = 0.23$ and

Likelihood Ratio $\chi^2[8] = 113.06$, $p < 0.001$, Cox & Snell $R^2 = 0.17$, respectively). As shown in Figures 2A and 2B, *JoC* reporting of reliability coefficients that consider chance and reporting of reliability coefficients for some but not all variables followed curvilinear trends that were characterized by a slight worsening from 1985 to 1993 and 1990, respectively, followed by improvements through 2010. *JMCQ* evidenced increasingly pronounced improvements for both variables across the entire study period. *CM* demonstrated improvements in reporting of reliability coefficients that consider chance through 1997 and reporting of reliability coefficients for some but not all variables through 1999, followed by general declines over the subsequent publication intervals. The probabilities at the end of the time period studied were almost equal to those at the beginning of the period.

For reporting of a reliability coefficient for every study variable, the most parsimonious model contained the linear year-by-journal interaction (Likelihood Ratio $\chi^2[5] = 119.6$, $p < 0.001$, Cox & Snell $R^2 = 0.18$). As shown in Figure 2C, while all three journals improved reporting of reliability coefficients for all variables over time, the improvement was most pronounced in *JoC*.

Discussion

When a researcher creates a content analysis protocol, the goal is to generate reliable and valid data in order to draw valid conclusions about patterns in the data. Reliability is a prerequisite for validity. The reliability reporting process must be explicitly explained in order for other researchers to adequately evaluate the data used in the project and in order to replicate the results in future studies. This particular study

examines content analysis in three communication journals over a 26-year period to determine if the reliability reporting has been transparent and consistent.

The results provide good news and bad news. The good news is that articles in these three flagship communication journals improved in the transparency of reporting reliability for all key variables during the time period. *JoC* and *JMCQ* additionally showed improvement in reporting reliability coefficients that consider chance and reporting reliability for some key variables. As illustrated in Figures 2A and 2B, *CM* showed initial improvements in these reporting domains followed by a return to near baseline reporting trends, although it should be noted that *CM* also had the greatest likelihood of publishing content analysis articles that that included these reliability reporting practices at the beginning of the study time period. Despite the improvement in reporting reliability coefficients across these three journals, there remains room for improvement, as *all* content analysis articles should report rigorous reliability statistics for all study variables.

Although the overall improvement trends were generally positive, journals varied in the rate and magnitude of change. However, these data cannot reveal the reasons for these patterns of improvement. General improvements in reporting reliability likely represent growing acceptance of standards for reporting of reliability of data from a protocol. The acceptance of such standards may be related to the growth in number of content analysis texts during this period. Starting with the seminal text by Krippendorff (1980, 2004b, 2013), the field added two additional texts (Neuendorf, 2002; and Riffe et al., 1998, 2005, 2014) during this period. All three are consistent in calling for a reliability check and the reporting of coefficients that include correction for chance. As

noted in the literature review above, this time period also experienced a growth in the number of articles about reporting reliability coefficients and in the number of empirical articles about reliability reporting. Finally, one cannot dismiss the contribution of readily available statistical software and online resources for calculating a range of reliability coefficients.

Equally important in evaluating and replicating research is the nature and selection of the reliability sample. Even if the same protocol is used in a replication, selecting a different type of sample could lead to different results. This study evaluated articles using three measures related to sampling transparency: type of sampling procedure (probability, non-probability, or no reliability check was reported); sampling unit selection process; and number of sample units. Improvements in reporting were less positive on these three measures than for reporting reliability coefficients themselves.

While reporting the number of units used in the reliability tests showed an improvement over time, the other two reliability sampling variables showed no significant temporal changes. In addition, the likelihood of *CM* and *JoC* articles explaining the sampling method and using a probability or census sample actually declined over time, albeit to a non-significant degree. Under some conditions, a sample used for a reliability test might not need to be a census or a probability sample. For example, a simple random sample might not yield enough content units so that all values within a variable would be coded. However, when such a non-representative reliability sample is necessary, it should be explained. The majority of these articles were not clear in reporting how the reliability sample was selected.

Again, these data cannot explain the difference between the improvements in reliability coefficient reporting and the general lack of improvement in reporting the reliability sampling process and in using census or probability samples to test reliability. One possible explanation, however, is that content analysis articles and general research texts address reliability sampling, but they often provide less guidance on the sampling process than advice about which reliability coefficients to use (Kaid & Wadsworth, 1989; Lombard et al., 2010; Wimmer & Dominick, 2003).

As with all studies, this one has limitations. We acknowledge that word-count restrictions of journals, editor and reviewer preferences, and author inclination can affect how detailed and elaborate discussion of these procedures can be. This study does not allow for unique circumstances or changes to a manuscript that may have arisen during the peer-review publication process. Only three communication journals were included in the study, although these are the flagship journals for the three largest communication associations. It would be useful to extend this analysis to other communication journals and even to content analysis in non-communication journals, such as those addressing political science, economics, and sociology. Given that content analysis is a method related to communication, however, it is unlikely that scholars in other fields will perform better than those publishing in communication journals.

Conclusion

The improvement of research methods can be a slow process, but these data indicate that reporting about the reliability process has improved since 1985 in content analysis published in these journals. These data show that the reporting of reliability for content analysis studies in these three communication journals improved from 1985 to

2010 for four of the six content variables. This suggests an increasing standardization of reliability reporting. Still, the reporting process leaves much to be desired. Reporting reliability coefficients that take chance into consideration for each variable is essential for proper evaluation of data validity and for study replication. Equally important for evaluation and replication is a detailed discussion of the sampling process and the resulting type of sample. However, even in 2010, the final year of this study period, many articles did not meet reporting standards necessary for evaluation and replication. A majority of the articles did not use a census or probability reliability sample and were not transparent about the sample selection process. The important question is: How can the improvement process be accelerated? First, standards need to be adopted by the research community. We suggest the following standards for reporting the reliability process:

- Always conduct and report a reliability check.
- Report reliability for each and every variable using coefficients that take chance into consideration.
- Use a probability sample or a census of study units to establish data reliability. If some of the variables have skewed distributions, probability sampling should be used to guarantee adequate representation of the distribution.
- Explicitly report the process by which the reliability sample was selected.
- Select the number of units based on recommendations for the number needed to create a representative sample. Check Krippendorff (2013) and Riffe et al. (2014) for processes. Report that number explicitly.

Accepting these standards is the first step, and teaching of these standards to graduate students and enforcement of the standards by journal editors and journal reviewers would further the science of content analysis.

References

Aiken, L. S., & West, S. G. (1991). Multiple regression: Testing and interpreting interactions. Newbury Park: Sage.

Berelson, B. (1952). *Content Analysis in Communication Research*. New York: Hafner Publishing Company.

Bloch, D. A., & Kraemer, H. C. (1989). 2 x 2 Kappa Coefficients: Measures for Agreement or Association. *Biometrics*, *45*, 269-287.

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.

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

Kaid L. L., & Wadsworth, A. J. (1989). Content analysis. In P. E. Barker & L. L. Barker (Eds.). *Measurement of Communication Behavior*. New York: Longman.

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology.* Beverly Hills, CA: Sage.

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

Krippendorff, K. (2004b). *Content analysis: An introduction to its methodology*, 2nd ed. Thousand Oaks, CA: Sage.

Krippendorff, K. (2011). Agreement and information in the reliability of coding. *Communication Methods and Measures, 5*(2), 93-112.

Krippendorff, K. (2012). Commentary: A dissenting view on so-called paradoxes of

reliability coefficients. In C. T. Salmon (ed.), *Communication Yearbook* 36 (pp. 481-499). New York: Routledge.

Krippendorff, K. (2013). *Content analysis: An introduction to its methodology*, 3rd ed. Los Angeles, CA: Sage.

Lacy, S., & Riffe, D. (1996). Sampling error and selecting intercoder reliability samples for nominal content categories. *Journalism & Mass Communication Quarterly*, *73*, 963-973.

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. (2004). A call for standardization in content analysis reliability. *Human Communication Research, 30,* 434-437.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2010). Practical resources for assessing and reporting intercoder reliability in content analysis research projects. Accessed February 2, 2013, at http:/matthewlombard.com/reliability/index_print.html.

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

Pasadeos, Y., Huhman, B., Standley, T., & Wilson, G. (1995). Applications of content analysis in news research: A critical examination. Paper presented at the annual conference of the Association or Education in Journalism and Mass Communication, Washington, D.C.

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. (1997). A content analysis of content analyses: 25 ears of *Journalism Quarterly*. *Journalism & Mass Communication Quarterly, 74,* 873-882.

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

Riffe, D., Lacy, S., & Fico, F. G. (2005). *Analyzing media messages: Using quantitative content analysis in research* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Riffe, D., Lacy, S., & Fico, F. G. (2014). *Analyzing media messages: Using quantitative content analysis in research* (3rd ed.). New York: Routledge.

Schutz, W. C. (1952). Reliability, ambiguity and content analysis. *Psychological Review, 59,* 119-129.

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

Snyder-Duch, J., Bracken, C. C., & Lombard, M. (2001). Content analysis in communication: Assessment and reporting of intercoder reliability. Paper presented at the annual conference of the International Communication Association, Washington, D.C.

Stempel, G. H., III. (1955). Increasing reliability in content analysis. *Journalism Quarterly, 32,* 449-455.

Stempel, G. H., III. (2003). Content Analysis. In G. H. Stempel, III, D. H. Weaver, & G. C. Wilhoit (Eds.). *Mass Communication Research and Theory* (pp. 209-219). Boston, MA: Allyn and Bacon.

Wimmer, R. D., & Dominick, J. R. (2003). *Mass media research: An introduction* (7th ed.). Belmont, CA: Wadsworth/Thomson.

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

Zhao, X., Liu, J. S., & Deng, K. (2012). Assumptions behind intercoder reliability indices. In C. T. Salmon (ed.), *Communication yearbook 36* (pp. 419-480). New York: Routledge.