Content Analysis in Mass Communication
Assessment and Reporting of Intercoder Reliability

MATTHEW LOMBARD
Temple University
JENNIFER SNYDER-DUCH
Carlow College
CHERYL CAMPANELLA BRACKEN
Cleveland State University

As a method specifically intended for the study of messages, content analysis is fundamental to mass communication research. Intercoder reliability, more specifically termed intercoder agreement, is a measure of the extent to which independent judges make the same coding decisions in evaluating the characteristics of messages, and is at the heart of this method. Yet there are few standard and accessible guidelines available regarding the appropriate procedures to use to assess and report intercoder reliability, or software tools to calculate it. As a result, it seems likely that there is little consistency in how this critical element of content analysis is assessed and reported in published mass communication studies. Following a review of relevant concepts, indices, and tools, a content analysis of 200 studies utilizing content analysis published in the communication literature between 1994 and 1998 is used to characterize practices in the field. The results demonstrate that mass communication researchers often fail to assess (or at least report) intercoder reliability and often rely on percent agreement, an overly liberal index. Based on the review and these results, concrete guidelines are offered regarding procedures for assessment and reporting of this important aspect of content analysis.

The study of communication is interdisciplinary, sharing topics, literatures, expertise, and research methods with many academic fields and disciplines. But one method, content analysis, is specifically appropriate and necessary for (arguably) the central work of communication scholars, in particular those who study mass communication: the analysis of messages. Given that content analysis is fundamental to communication research (and thus theory), it would be logical to expect researchers in communication to be among the most, if not the most, proficient and rigorous in their use of this method.

Matthew Lombard (Ph.D., Stanford University, 1994) is an associate professor in the Department of Broadcasting, Telecommunications and Mass Media as well as Director of the Mass Media and Communication doctoral program at Temple University, Philadelphia, PA. *Jennifer Snyder-Duch* (Ph.D., Temple University, 2000) is an assistant professor of communication studies at Carlow College, Pittsburgh, PA. *Cheryl Campanella Bracken* (Ph.D., Temple University, 2000) is an assistant professor in the Department of Communication at Cleveland State University, OH.

Human Communication Research, Vol. 28 No. 4, October 2002 587–604

Intercoder reliability (more specifically "intercoder agreement"; Tinsley & Weiss, 1975, 2000) is "near the heart of content analysis; if the coding is not reliable, the analysis cannot be trusted" (Singletary, 1993, p. 294). However, there are few standards or guidelines available concerning how to properly calculate and report intercoder reliability. Further, although a handful of tools are available to implement the sometimes complex formulae required, information about them is often difficult to find and they are often difficult to use. It therefore seems likely that many studies fail to adequately establish and report this critical component of the content analysis method.

This article reviews the importance of intercoder agreement for content analysis in mass communication research. It first describes several indices for calculating this type of reliability (varying in appropriateness, complexity, and apparent prevalence of use), and then presents a content analysis of content analyses reported in communication journals to establish how mass communication researchers have assessed and reported reliability, demonstrating the importance of the choices they make concerning it. The article concludes with a presentation of guidelines and recommendations for the calculation and reporting of intercoder reliability.

CONTENT ANALYSIS AND THE IMPORTANCE OF INTERCODER RELIABILITY

Berelson's (1952) often cited definition of content analysis as "a research technique for the objective, systematic, and quantitative description of the manifest content of communication" (p. 18) makes clear the technique's unique appropriateness for researchers in our field. This is reinforced by Kolbe and Burnett's (1991) definition which states that content analysis is "an observational research method that is used to systematically evaluate the symbolic content of all forms of recorded communication. These communications can also be analyzed at many levels (image, word, roles, etc.), thereby creating a realm of research opportunities" (p. 243). While content analysis can be applied to any message, the method is often used in research on mass mediated communication.

Riffe and Freitag (1997) note several studies that demonstrate the widespread and increasing use of content analysis in communication. The method has been well represented in graduate research methods courses, theses, dissertations, and journals. In their own study they report a statistically significant trend over 25 years (1971–1995) in the percentage of full research reports in *Journalism & Mass Communication Quarterly* that feature this method, and they note that improved access to media content through databases and archives, along with new tools for computerized content analysis, suggests the trend is likely to continue.

Intercoder reliability is the widely used term for the extent to which independent coders evaluate a characteristic of a message or artifact and reach the same conclusion. Although this term is appropriate and will be used here, Tinsley and Weiss (1975, 2000) note that the more specific term for the type of consistency required in content analysis is intercoder (or interrater) agreement. They write that while reliability could be based on correlational (or analysis of variance) indices that assess the degree to which "ratings of different judges are the same when expressed as deviations from their means," intercoder agreement is needed in content analysis because it measures only "the extent to which the different judges tend to assign exactly the same rating to each object" (Tinsley & Weiss, 2000, p. 98).[1]

It is widely acknowledged that intercoder reliability is a critical component of content analysis and (although it does not ensure validity) when it is not established, the data and interpretations of the data can never be considered valid. As Neuendorf (2002) notes, "given that a goal of content analysis is to identify and record relatively objective (or at least intersubjective) characteristics of messages, reliability is paramount. Without the establishment of reliability, content analysis measures are useless" (p. 141). Kolbe and Burnett (1991) write that "interjudge reliability is often perceived as the standard measure of research quality. High levels of disagreement among judges suggest weaknesses in research methods, including the possibility of poor operational definitions, categories, and judge training" (p. 248).

A distinction is often made between the coding of the manifest content, information "on the surface," and the latent content beneath these surface elements. Potter and Levine-Donnerstein (1999) note that for latent content the coders must provide subjective interpretations based on their own mental schema and that this "only increases the importance of making the case that the judgments of coders are intersubjective, that is, those judgments, while subjectively derived, are shared across coders, and the meaning therefore is also likely to reach out to readers of the research" (p. 266).

There are important practical reasons to establish intercoder reliability as well. Neuendorf (2002) argues that, in addition to being a necessary (although not sufficient) step in validating a coding scheme, establishing a high level of reliability also has the practical benefit of allowing the researcher to divide the coding work among many different coders. Rust and Cooil (1994) note that intercoder reliability is important to marketing researchers in part because "high reliability makes it less likely that bad managerial decisions will result from using the data" (p. 11). Potter and Levine-Donnerstein (1999) make a similar argument regarding applied work in public information campaigns.

MEASURING INTERCODER RELIABILITY

Intercoder reliability is assessed by having two or more coders categorize units (programs, scenes, articles, stories, words, etc.), and then using these categorizations to calculate a numerical index of the extent of agreement between or among the coders. There are many variations in how this process can and should be conducted, but at a minimum the researcher has to create a representative set of units for testing reliability and the coding decisions must be made independently under the same conditions. A separate pilot test is often used to assess reliability during coder training, with a final test to establish reliability levels for the coding of the full sample (or census) of units. Researchers themselves may serve as coders, a practice questioned by some (e.g., Kolbe & Burnett, 1991) because it weakens the argument that other independent judges can reliably apply the coding scheme. In some cases the coders evaluate different but overlapping units (e.g., coder 1 codes units 1–20, coder 2 codes units 11–30, etc.), but this technique has also been questioned (Neuendorf, 2002).

With the coding data in hand, the researcher calculates and reports one or more indices of reliability. Popping (1988) identified 39 different "agreement indices" for coding nominal categories, which excludes several techniques for ratio and interval level data, but only a handful of techniques are widely used.[2]

Percent Agreement

Percent agreement—also called simple agreement, percentage of agreement, raw percent agreement, or crude agreement—is the percentage of all coding decisions made by pairs of coders on which the coders agree. As with most indices, percent agreement takes values of .00 (no agreement) to 1.00 (perfect agreement). The obvious advantages of this index are that it is simple, intuitive, and easy to calculate. It also can accommodate any number of coders. However, this method also has major weaknesses, the most important of which involves its failure to account for agreement that would occur simply by chance. Consider this example: Two coders are given 100 units (news stories, words, etc.) to code as having or not having a given property. Without any instructions or training, without even knowing the property they are to identify, they will agree half of the time, and these random agreements will produce a percent agreement value of .50.

This problem is most severe when there are fewer categories in a coding scheme, but it remains in any case, making it difficult to judge and compare true reliability across variables (Perrault & Leigh, 1989). Seun and Lee (1985) reanalyzed data from a sample of published studies correcting for chance agreement and concluded that "between one-fourth and three-fourths of the reported observations

could be judged as unreliable against a lenient criterion and between one-half and three fourths could be judged as unreliable against a more stringent criterion" (p. 221).

Characteristics of the percent agreement index also allow researchers to artificially inflate reliability by adding categories they know will rarely be used or produce disagreement. Kolbe and Burnett (1991) note that, while this can be done with other indices, it is a particular problem with percent agreement.

Another limitation is that percent agreement records only agreements and disagreements—there is no "credit" for coders whose decisions are "close." Thus it only makes sense to use percent agreement with nominal level variables.[3]

Holsti's Method

Holsti (1969) proposed a variation on the percent agreement index; with two coders evaluating the same units for a reliability test it is identical to percent agreement, however, it also accounts for situations in which the coders evaluate different units. The result is often calculated not for a single variable but across a set of variables, a very poor practice which can hide variables with unacceptably low levels of reliability (Kolbe & Burnett, 1991; Neuendorf, 2002).

Scott's Pi (π)

One index that accounts for chance agreement is Scott's pi (1955). Unlike percent agreement and Holsti's method, this index takes into account the number of categories as well as the distribution of values across them, but because it assumes that these proportions are the true proportions rather than the result of agreement among the coders, many consider the index too conservative. To illustrate this, consider the coding results in Table 1. Although according to those results agreement appears to be high, the Scott's pi index would be only .05.

This index also does not account for differences in how the individual coders distribute their values across the coding categories, a potential source of systematic bias; that is, it assumes the coders have distributed their values across the categories identically and if this is not the case, the formula fails to account for the reduced agreement (Craig, 1981; Hughes & Garrett, 1990; Neuendorf, 2002). Scott's pi is appropriate only for nominal level variables and two coders (although Craig, 1981, has suggested an extension for three or more coders).

Cohen's Kappa (κ)

Cohen's kappa (1960, 1968) index also accounts for chance agreement, using the same conceptual formula as Scott's pi. Expected agreement by

TABLE 1
Example Data for Illustration of Scott's Pi and Cohen's Kappa

Coder 1	*Coder 2*		
	Yes	*No*	*Total*
Yes	81	9	90
No	9	1	10
Total	90	10	100

chance in this case is calculated based on the "multiplicative marginals" rather than the additive ones, which has the effect of accounting for differences in the distribution of values across the categories for different coders. However, this, along with the fact that it still only "gives credit" for agreement beyond the distributions of values in the marginals, makes it another conservative measure (kappa in the example illustrated in Table 1 would be 0.00; Perrault & Leigh, 1989). It should be noted that Cohen recognized this limitation of his measure, but as Perreault and Leigh note, "he was concerned mainly with psychological applications for which there often would be clearly established prior knowledge of the likely distribution of observations across cells" (p. 139), which is not typically the case in communication research. Brennan and Prediger (1981) discuss this and other potential problems with Cohen's kappa, including cases where even with perfect agreement the index has a maximum value less than 1.00. The index has been adapted for multiple coders and cases in which different coders evaluate different units (Fleiss, 1971). Cohen (1968) proposed a weighted kappa to account for different types of disagreements, however, as with the other indices discussed so far, this measure is generally used only for nominal level variables.

Krippendorff's Alpha (α)

Krippendorff's alpha index (1980) is attractive for several reasons. It allows for any number of coders and is explicitly designed to be used for variables at different levels of measurement from nominal to ratio. It also accounts for chance agreements, using the same assumption as Scott's pi of equal marginal proportions for the coders. The biggest drawback to its use has been its complexity and the resulting difficulty of "by-hand" calculations, especially for interval and ratio level variables.

Despite all the effort that scholars, methodologists, and statisticians

have devoted to developing and testing indices of intercoder reliability, there is no consensus on a single, "best" index. There are several recommendations for Cohen's kappa (e.g., Dewey, 1983, argued that despite its drawbacks, kappa should still be "the measure of choice") and it appears to be commonly used in research that involves the coding of behavior (Bakeman, 2000); however, others favor a different index. There is general agreement that indices which do not account for chance agreement are too liberal while those that do are too conservative. There is also consensus that a few indices used are inappropriate measures of intercoder reliability: Cronbach's alpha was designed to only measure internal consistency via correlation, standardizing the means and variance of data from different coders and only measuring covariation (Hughes & Garrett, 1990), and chi-square produces high values for both agreement and disagreement deviating from agreement expected by chance (the "expected values" in the chi-square formula).

DETERMINING AN "ACCEPTABLE" LEVEL OF RELIABILITY

In addition to the choice of the appropriate index of intercoder reliability, another difficulty is determining what constitutes an acceptable level of reliability. Again, there are no established standards, but Neuendorf (2002) reviews "rules of thumb" set out by several methodologists (including Banerjee, Capozzoli, McSweeney, & Sinha, 1999; Ellis, 1994; Frey, Botan, & Kreps, 2000; Krippendorff, 1980; Popping, 1988; and Riffe, Lacy, & Fico, 1998) and concludes that "coefficients of .90 or greater would be acceptable to all, .80 or greater would be acceptable in most situations, and below that, there exists great disagreement" (p. 145). The criterion of .70 is often used for exploratory research. More liberal criteria are usually used for the indices known to be more conservative (i.e., Cohen's kappa and Scott's pi).

TOOLS FOR CALCULATING INTERCODER RELIABILITY

Researchers who need to calculate intercoder reliability have had few automated tools at their disposal and have usually had to do the calculations by hand. A few people (Berry & Mielke, 1997; Kang, Kara, Laskey, & Seaton, 1993) have written "macros," customized programming that can be used with existing software, to automate the calculations. Others have created stand-alone software (Krippendorff, 2001; Popping, 1984; see ProGAMMA, 2002; Skymeg Software, 2002). To date, however, none of these tools have been widely available or easy to use, and most have been neither.

INTERCODER RELIABILITY IN RESEARCH REPORTS

Given its importance to content analysis, several researchers have examined the calculation and reporting of intercoder reliability in a variety of literatures over a variety of time spans. Perrault and Leigh (1989) note that in the marketing research literature there is "no accepted standard for evaluating or reporting the reliability of coded data" and that "the most commonly used measure . . . is the simple percentage of agreement"; they call Cohen's kappa "the most widely used measure of interjudge reliability across the behavioral science literature" (p. 137).

Hughes and Garrett (1990) coded 68 articles in *Journal of Marketing Research*, *Journal of Marketing*, and *Journal of Consumer Research* during 1984–1987 that contained reports of intercoder reliability and found 65% used percent agreement. Kolbe and Burnett (1991) coded 128 articles from consumer behavior research in 28 journals, three proceedings and one anthology between 1978 and 1989. Most of the authors were in marketing departments (only 12.2% were from communication, advertising, and journalism schools or departments). Percent agreement was reported in 32% of the studies, followed by Krippendorff's alpha (7%), and Holsti's method (4%); often the calculation method wasn't specified, and in 31% of the articles no reliability was reported. Also, 36% of the studies reported only an overall reliability, which can hide variables with unacceptably low agreement. Consistent with these findings, Kang et al. (1993) reviewed the 22 articles published in the *Journal of Advertising* between 1981 and 1990 that employed content analysis and found that 78% "used percentage agreement or some other inappropriate measure" (p. 18).

Pasadeos, Huhman, Standley, and Wilson (1995) coded 163 content analyses of news-media messages in four journals (*Journalism & Mass Communication Quarterly*, *Newspaper Research Journal*, *Journal of Broadcasting and Electronic Media*, and *Journal of Communication*) for the 6-year period of 1988–1993. They wrote that "we were not able to ascertain who specifically had done the coding in approximately 55% of the studies; a similar number had not reported on whether coding was done independently or by consensus; and more than 80% made no mention of coder training" (p. 8). In their study 51% of the articles did not address reliability at all, 31% used percent agreement, 10% used Scott's pi, and 6% used Holsti's method. Only 19% gave reliability figures for all variables while 20% gave only an overall figure.

In a study of content analyses published in *Journalism & Mass Communication Quarterly* between 1971 and 1995, Riffe and Freitag (1997) found that out of 486 articles, only 56% reported intercoder reliability and of those most only reported an overall figure, while only 10% "explicitly specified random sampling in reliability tests" (p. 877). But an en-

couraging result was a near-monotonic rise in the percentage of articles reporting intercoder reliability from 50% in 1971–1975 to 71.7% in 1991–1995.

RESEARCH QUESTION: INTERCODER RELIABILITY IN RECENT COMMUNICATION RESEARCH REPORTS

These reviews focus on subsets of communication scholarship (in terms of topic area or journals), and focus primarily on the adequacy of assessment rather than the reporting of intercoder agreement, with the most recent publications examined from the mid-1990s. They provide reason for both optimism and pessimism regarding the use of intercoder reliability in content analyses in mass communication and suggest an important research question: How adequately and consistently is intercoder reliability currently assessed and reported in published mass communication research?

METHOD

To answer the research question, a content analysis was conducted of research reports in communication in which content analysis was the primary research method.

Sample

All articles indexed in *Communication Abstracts* for the years 1994 through 1998 for which one of the keywords was "content analysis" were selected for coding. *Communication Abstracts* is a comprehensive bimonthly index of the communication literature published in over 75 journals. The final sample (considered a census) consisted of 200 articles.[4]

Variables Coded

The variables coded for each article are presented in Table 2. The complete coding instrument is available from the authors.

Instrument Development, Coder Training, and Intercoder Reliability

The authors tested an initial draft of the coding instrument informally by independently coding six articles, some from within the sample and some that were published in years prior to those in the sample. Based on this test, coding problems and disagreements were discussed and the instrument was revised. This process was repeated several times until it

was believed the instrument would permit reliable coding by any competent and trained set of coders, at which time a pilot test of reliability was conducted formally using the indices discussed below.

To establish intercoder reliability, the second and third authors both coded 128 (64%) of the articles. They later each coded half of the remaining 72 articles. To create the final dataset, the articles used in the reliability analysis were divided randomly into two groups and the coding decisions of each coder were randomly selected to be used for each group of articles.

Percent agreement, Scott's pi, Cohen's kappa, and Krippendorff's alpha were all used to assess intercoder reliability for each variable coded. A beta version of the software package PRAM (Program for Reliability Assessment with Multiple-coders, Skymeg Software, 2002) was used to calculate the first three of these. A beta version of a separate program, Krippendorff's Alpha 3.12, was used to calculate the fourth. Holsti's (1969) method was not calculated because, in the case of two coders who evaluate the same reliability sample, the results are identical to those for percent agreement. For the coding of a variable to be considered reliable it was required that Krippendorff's alpha (an index that accounts for level of measurement and agreement expected by chance and is known to be conservative) be .70 or higher, or if this was not the case, percent agreement (a liberal index) be .90 or higher. The reliability results are reported in Table 2.

RESULTS

The results for each variable are presented in Table 2. Only 69% of the research reports (n = 137) contained any report of intercoder reliability. Of that subset, the mean number of sentences in the text and footnotes that were devoted to discussion and reporting of reliability was 4.5 (SD = 4), only 6% of these articles included a table that contained reliability results, and less than half (45%) of the articles included a citation related to intercoder reliability.

Usually the specific index used to calculate reliability was not given; when an index was reported the most frequently mentioned were Holsti's method (15%), Scott's pi (10%), percent agreement (9%), Cohen's kappa (7%), and Krippendorff's alpha (3%; percent agreement is most likely underrepresented in the results here because only use of the specific terms "percent agreement" and "simple agreement" were coded as representing use of this index). The reporting of which index or indices were used was often ambiguous, with labels such as "intercoder reliability," "interrater reliability," "intercoder agreement," and just "reliability" common. Among this same subset of articles, only 2% (n = 3) indicated

TABLE 2
Intercoder Reliability and Percentages and Means for All Variables

Variable	*Percent agreement*	*Scott's pi*	*Cohen's kappa*	*Krippendorff's alpha*	*% (n) or mean (SD)*
Name for study of interest's method in title, abstract, or text?					
"Content analysis"[a]	.90	.70	.72	.72	74% (147)
Is method of study of interest only method used in text?[a]	.91	.64	.66	.66	80% (160)
Is method of study of interest quantitative in nature?[a]	.90	.63	.71	.72	
No					8% (16)
Some quantitative, some not					12% (24)
Yes, all quantitiative					80% (160)
What medium is analyzed?[a]					
Newspapers	.95	.92	.90	.90	42% (83)
Magazines	.96	.88	.87	.88	30% (60)
Television	.98	1.00	.94	.95	18% (36)
Radio	1.00	1.00	1.00	.67	3% (5)
Film	1.00	1.00	1.00	.85	2% (3)
Data from respondents	1.00	.66	1.00	.56	2% (4)
Other	.94	.89	.75	.76	16% (32)
What type of content is analyzed?[a]					
Advertising	.97	.89	.92	.92	49% (97)
News	.95	.90	.89	.89	22% (44)
Entertainment	.97	.86	.88	.88	13% (25)
Number of coders who participated in coding the actual sample?[a]	.84	.74	.74	.75	
One coder					17% (34)
More than one coder					49% (96)
Coded by a computer system					2% (3)
Not reported					33% (64)
Number of multiple coders who participated in coding the actual sample?[ab]	.84	.89	.75	.89	2.47 (1.16) (min = 1; max = 40)
Was the amount of training reported?[a]	.94	.69	.66	.66	9% (19)
Reliability discussed?	.97	.91	.93	.93	69% (137)
Number of sentences about reliability in text and footnotes?[b]	.67	.41	.41	.86	4.45 (4.0)
Citation(s) about reliability?	.92	.87	.84	.84	45% (61)
Table(s) with reliability information?	.97	.61	.69	.69	6% (8)

TABLE 2 Continued
Intercoder Reliability and Percentages and Means for All Variables

Variable	*Percent agreement*	*Scott's pi*	*Cohen's kappa*	*Krippendorff's alpha*	*% (n) or mean (SD)*
Name of reliability method?					
Krippendorff's alpha	1.00	1.00	1.00	1.00	3% (4)
Scott's pi	1.00	1.00	1.00	1.00	10% (14)
Cohen's kappa	1.00	1.00	1.00	1.00	7% (10)
Holsti's method	.95	.81	.86	.86	15% (21)
"simple agreement" only	.97	.65	.65	-.006	2% (2)
"percentage agreement" only	.95	.65	.65	.65	7% (10)
"intercoder reliability" only	.93	.71	.74	.75	16% (22)
Lowest accepted reliability criterion reported?[b]	.85	.97	.85	.97	.75 (.26)
Is the specific reliability for one or more variables reported?	.87	.81	.74	.75	41% (56)
Specific formula(e) reprinted in text, table, or footnotes?	.97	.38	.65	.65	4% (5)
Computing method reported?	.99	.65	.66	.67	2% (3)
Reliability sample size?[b]	.60	.49	.55	.74	341 (1307) (n = 85; min = 1; max = 1,300 median = 79)
Does the text state that during coding of the reliability sample (not during coder training or coding of the actual sample) coders discussed specific units and how to code them?	.92	.47	.33	.33	11% (15)
Does the text state how discrepancies were resolved?	.87	.64	.72	.72	26% (36)
Number of coders who participated in reliability coding?	.92	.64	.55	.76	
More than one coder					91% (124)
Coded by a computer system					1% (1)
Not reported					9% (12)
Number of multiple coders who participated in reliability coding?[b]	.93	.76	.55	.79	2.34 (1.05) (min = 2, max = 40)

NOTE: Holsti's method is not reported because it is identical to Scott's pi in the case of two coders evaluating the same units. [a] Based on full sample (N = 200). [b] = ratio; all other variables are nominal.

the computing tools used (e.g., SPSS, "by hand"). The lowest reliability level reported was .40, while the mean minimum accepted reliability level was .75 (*SD* = .26).

When reliability was addressed, many articles still excluded important information, including the size of the reliability sample (missing in 38% of the articles), the number of reliability coders (9%), the reliability for specific variables (rather than an overall average or range figure, 59%), the amount of training that had been required to reach the reliability levels reported (86%), and whether or how discrepancies among coders had been resolved (74%).

Some of the research reports contained thorough and yet concise reports of intercoder reliability (e.g., Lichter, Lichter, & Amundsom, 1997). Other authors provided much less, and in some cases ambiguous or inappropriate, information. Of course these articles did at least contain some report of information regarding intercoder reliability.

CONCLUSION

This content analysis has demonstrated that substantial problems remain in the assessment and reporting of intercoder reliability which cast doubt on the validity of much of the work of mass communication researchers. Pasadeos et al. (1995) found that only 49% of 163 content analyses of news media in four major communication journals between 1988 and 1993 reported reliability. Riffe and Freitag (1997) found that only 72% of articles in *Journalism & Mass Communication Quarterly* from 1991 to 1995 did so. The comparable figure in this study was 69%, or just over two thirds. Further, the results indicate that most studies that do report reliability devote little space to reliability procedures and results. In addition, reliability for individual variables is reported less than half the time. It was also shown that researchers often either don't identify the index used to calculate reliability or rely on indices that don't adequately account for the role of agreement expected by chance. The importance of the decision regarding the choice of index or indices of intercoder reliability is demonstrated by the wide variation in reliability levels presented in Table 2. Of course the assessment of intercoder reliability in many of the studies may have been adequate or even exemplary, but incomplete or ambiguous reporting of the procedures and results prevents readers from reaching this conclusion.

These results are not offered as an indictment of mass communication scholars or their work; rather, they can be seen as the consequence of a lack of detailed and practical guidelines and tools available to research-

ers regarding reliability. Therefore, based on the review of literature and the results of this study, the following standards and guidelines for the calculation and reporting of intercoder reliability are proposed.

1. *Calculate and report intercoder reliability.* All content analysis projects should be designed to include (a) multiple coders of the content and (b) assessment and reporting of intercoder reliability among them. Reliability is a necessary (although not sufficient) criterion for validity in the study and without it, all results and conclusions in the research project may justifiably be doubted or even considered meaningless.

2. *Select one or more appropriate indices.* Choose 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. If percent agreement is selected, use a second index that accounts for agreement expected by chance. Be prepared to justify and explain the selection of the index or indices.

3. *Obtain the necessary tools to calculate the index or indices selected.* Some of the indices can be calculated "by hand" (although this may be quite tedious) while others require automated calculation. For researchers proficient with their use, macros for some indices for the software packages SAS and SPSS are available from various sources (consult the authors for details). Popping's (1984) AGREE specialty software is available (see ProGAMMA, 2002) for two indices appropriate for nominal data, and Krippendorff (2001) and Neuendorf (2002; see Skymeg Software, 2002) have announced forthcoming software for Krippendorff's alpha and several indices, respectively.

4. *Select an appropriate minimum acceptable level of reliability for the index or indices to be used.* Coefficients of .90 or greater are nearly always acceptable, .80 or greater is acceptable in most situations, and .70 may be appropriate in some exploratory studies for some indices. Higher criteria should be used for indices known to be liberal (i.e., percent agreement) and lower criteria can be used for indices known to be more conservative (Cohen's kappa, Scott's pi, and Krippendorff's alpha). The preferred approach is to calculate and report two (or more) indices, establishing a decision rule that takes into account the assumptions and weaknesses of each (e.g., to be considered reliable, a variable may be at or above a moderate level for a conservative index, or at or above a high level for a liberal index). In any case the researcher should be prepared to justify the criterion/a used.

5. *Assess reliability informally during coder training.* Following instrument design and preliminary coder training, assess reliability informally with a small number of units which ideally are not part of the full sample (or census) to be coded, and refine the instrument and coding instructions until the informal assessment suggests an adequate level of agreement.

6. *Assess reliability formally in a pilot test.* Using a random or other justi-

fiable procedure, select a representative sample for a pilot test of intercoder reliability. The size of this sample can vary depending on the project but a good rule of thumb is 30 units (for more guidance see Lacy and Riffe, 1996). If at all possible, select a separate representative sample for use in pilot testing of reliability. Coding must be done independently and without consultation or guidance. If possible, the researcher should not be a coder. If reliability levels in the pilot test are adequate, proceed to the full sample. If they are not adequate, conduct additional training, refine the coding instrument and procedures, and only in extreme cases, replace one or more coders.

7. *Assess reliability formally during coding of the full sample.* When confident that reliability levels will be adequate (based on the results of the pilot test of reliability), use a representative sample from the full sample to be coded to assess reliability (the reliability levels obtained in this test are the ones to be presented in all reports of the project). This sample must also be selected using a random or other justifiable procedure. The appropriate size of the sample depends on many factors and should not be less than 50 units or 10% of the full sample, but it rarely will need to be greater than 300 units. Larger reliability samples are required when the full sample is large or when the expected reliability level is low (see Lacy & Riffe, 1996 for a discussion; Neuendorf, 2002). The units from the pilot test of reliability can be included in this reliability sample only if the reliability levels obtained in the pilot test were adequate. As with the pilot test, this coding must be done independently, without consultation or guidance.

8. *Select and follow an appropriate procedure for incorporating the coding of the reliability sample into the coding of the full sample.* Unless reliability is perfect, there will be coding disagreements for some units in the reliability sample. Although an adequate level of intercoder agreement suggests that the decisions of each of the coders could reasonably be included in the final data, and although it can only address the subset of potential coder disagreements that are discovered in the process of assessing reliability, the researcher must decide how to handle these coding disagreements. Depending on the characteristics of the data and the coders, the disagreements can be resolved by randomly selecting the decisions of the different coders, using a "majority" decision rule (when there are an odd number of coders), having the researcher or other expert serve as tie-breaker, or discussing and resolving the disagreements. The researcher should be able to justify whichever procedure is selected.

9. *Do not do any of the following:*

- Use only percent agreement to calculate reliability.
- Use Cronbach's alpha, Pearson's *r*, or other correlation-based indices that standardize coder values and only measure covariation. While these indices may be used as a measure of reliability in other contexts,

reliability in content analysis requires an assessment of intercoder agreement (i.e., the extent to which coders make the identical coding decisions) rather than covariation.

- Use chi-square to calculate reliability.
- Use overall reliability across variables (rather than reliability levels for each variable) as a standard for evaluating the reliability of the instrument.
- Use overlapping reliability coding, in which judges code overlapping sets of units.

10. *Report intercoder reliability in a careful, clear, and detailed manner in all research reports.* Even if the assessment of intercoder reliability is adequate, readers can only evaluate a study based on the information provided, which must be both complete and clear. Provide this minimum information:

- The size of and the method used to create the reliability sample, along with a justification of that method.
- The relationship of the reliability sample to the full sample (i.e., whether the reliability sample is the same as the full sample, a subset of the full sample, or a separate sample).
- The number of reliability coders (which must be two or more) and whether or not they include the researchers.
- The amount of coding conducted by each reliability and nonreliability coder.
- The index or indices selected to calculate reliability and a justification of these selections.
- The intercoder reliability level for each variable, for each index selected.
- The approximate amount of training (in hours) required to reach the reliability levels reported.
- How disagreements in the reliability coding were resolved in the full sample.
- Where and how the reader can obtain detailed information regarding the coding instrument, procedures, and instructions (e.g., from the authors).

Given the central role of intercoder reliability in content analysis and the fundamental and increasingly prominent role of this research method in communication, we hope that these guidelines, as well as the growing availability of the needed calculation tools, will help improve the quality of research in our field.

NOTES

1. Even when intercoder agreement is used for variables at the interval or ratio levels of measurement, actual agreement on the coded values (even if similar rather than identical values "count") is the basis for assessment.

2. Perreault and Leigh's (1989) I_r measure; Tinsley and Weiss's (1975) *T* index; Bennett, Alpert, and Goldstein's (1954) *S* index; Lin's (1989) concordance coefficient; Hughes and Garrett's (1990) approach based on generalizability theory; and Rust and Cooil's (1994) approach based on proportional reduction in loss are just some of the indices proposed, and in some cases widely used, in other fields.

3. Lawlis and Lu (1972) and others have adapted percent agreement for ordinal, interval, and ratio scales by defining agreement as "within x values" on a scale, but these adaptations appear to be rarely used by researchers.

4. Twenty-two articles listed in *Communication Abstracts* were excluded from the sample for the following reasons: (a) the article was about the method of content analysis, but did not report a study that used the method; (b) the article was misidentified as being a report of a content analysis; or (c) the article could not be located in the libraries at Temple University and the University of Pennsylvania or through interlibrary loan.

REFERENCES

Bakeman, R. (2000). Behavioral observation and coding. In H. T. Reis & C. M. Judge (Eds.), *Handbook of research methods in social and personality psychology* (pp. 138–159). New York: Cambridge University Press.

Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *The Canadian Journal of Statistics, 27*, 3–23.

Bennett, E. M., Alpert, R., & Goldstein, A. C. (1954). Communications through limited response questioning. *Public Opinion Quarterly, 18,* 303–308.

Berelson, B. (1952). *Content analysis in communication research.* Glencoe, IL: Free Press.

Berry, K. J., & Mielke, P. W., Jr. (1997). Measuring the joint agreement between multiple raters and a standard. *Educational and Psychological Measurement, 57*, 527–530.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41,* 687–699.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement of partial credit. *Psychological Bulletin, 70*, 213–220.

Craig, R. T. (1981). Generalization of Scott's index of intercoder agreement. *Public Opinion Quarterly, 45*, 260–264.

Dewey, M. E. (1983). Coefficients of agreement. *British Journal of Psychiatry, 143,* 487–489.

Ellis, L. (1994). *Research methods in the social sciences.* Madison, WI: Brown & Benchmark.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76,* 378–382.

Frey, L. R., Botan, C. H., & Kreps, G. L. (2000). *Investigating communication: An introduction to research methods* (2nd ed.). Boston: Allyn & Bacon.

Holsti, O. R. (1969). *Content analysis for the social sciences and humanities.* Reading, MA: Addison-Wesley.

Hughes, M. A., & Garrett, D. E. (1990). Intercoder reliability estimation—approaches in marketing: A generalizability theory framework for quantitative data. *Journal of Marketing Research, 27*, 185–195.

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*, 17–28.

Kolbe, R. H., & Burnett, M. S. (1991). Content-analysis research: An examination of applications with directives for improving research reliability and objectivity. *Journal of Consumer Research, 18,* 243–250.

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

Krippendorff, K. (2001, August 22). Content #724. Message posted to the Content electronic mailing list, available from http://www.content-analysis.de/contpub.htm

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.

Lawlis, G. F., & Lu, E. (1972). Judgment of counseling process: Reliability, agreement, and error. *Psychological Bulletin, 78,* 17–20.

Lichter, S. R., Lichter, L. S., & Amundsom, D. (1997). Does Hollywood hate business or money? *Journal of Communication, 47*, 68–84.

Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics, 45,* 255–268.

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

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

Perreault, W. D., & Leigh, L. E. (1989). Reliability of nominal data based on qualitative judgments. *Journal of Marketing Research, 26*, 135–148.

Popping, R. (1984). AGREE, a package for computing nominal scale agreement. *Computational Statistics and Data Analysis, 2*, 182–185.

Popping, R. (1988). On agreement indices for nominal data. In W. E. Saris & I. N. Gallhofer (Eds.), *Sociometric research: Data collection and scaling* (Vol. 1, pp. 90–105). New York: St. Martin's Press.

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

ProGAMMA. (2002, July 7). AGREE [Computer software]. Available from http://www.gamma.rug.nl/

Riffe, D., & Freitag, A. A. (1997). A content analysis of content analyses: Twenty-five years 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: Erlbaum.

Rust, R., & Cooil, B. (1994). Reliability measures for qualitative data: Theory and implications. *Journal of Marketing Research, 31*, 1–14.

Scott, W. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly, 17,* 321–325.

Seun, H. K., & Lee, P. S. C. (1985). Effects of the use of percentage agreement on behavioral observation reliabilities: A reassessment. *Journal of Psychopathology and Behavioral Assessment, 7,* 221–234.

Singletary, M. W. (1993). *Mass communication research: Contemporary methods and applications.* Boston: Addison-Wesley.

Skymeg Software (2002, July 7). Program for reliability assessment with multiple coders (PRAM) [Computer software]. Available from http://www.geocities.com/skymegsoftware/pram.html

Tinsley, H. E. A., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgements. *Journal of Counseling Psychology, 22,* 358–376.

Tinsley, H. E. A., & Weiss, D. J. (2000). Interrater reliability and agreement. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 95–124). San Diego, CA: Academic Press.