University of Portland
Pilot Scholars

Communication Studies Faculty Publications and Presentations

Communication Studies

2015

Issues and Best Practices in Content Analysis

Stephen Lacy

Brendan R. Watson

Daniel Riffe

Jennette Lovejoy
University of Portland, lovejoy@up.edu

Issues and Best Practices in Content Analysis

Content analysis has become a central method in communication research. Of the 2,534 articles Lovejoy, Watson, Lacy, and Riffe (2014) studied from *Journalism & Mass Communication Quarterly*, *Journal of Communication,* and *Communication Monographs* between 1985 and 2010, 23% involved content analysis. But use of content analysis is not limited to communication research (Krippendorff, 2013, pp. 11-12), and its expanded use has resulted in efforts to standardize the method. From Berelson (1952) and Holsti (1969), through Krippendorff (1980), Riffe, Lacy, and Fico (1998), and Neuendorf (2002), texts have aimed to explain and codify standards for content analysis data generation and reporting. Nonetheless, articles continue to be published that do not meet design standards, reporting standards, or both: Lovejoy et al. (2014) stated in their study of three flagship journals, "However, even in 2010, the final year of this study period, many articles did not meet reporting standards necessary for evaluation and replication" (p. 220).

The failure to meet standards may reflect a lack of knowledge, changes in content analysis methods as a result technological developments, or the fact that agreement on some standards is in flux. For example, there is disagreement in the literature about which coefficients should be used in evaluating reliability. At the same time, digital technology has increased the use of computers for accessing, storing, and coding content, but the best use of those approaches continues to be explored.

Though scholars continue to innovate data collection and content coding, the fundamental elements of content analysis are captured in the definition proffered by Riffe, Lacy, and Fico (2014): "the systematic and replicable examination of symbols of communication, which have been assigned numeric values according to valid measurement rules, and the analysis of relationships involving those values using statistical methods, to describe the communication, draw inferences about its meaning, or infer from the communication to its context, both of production and consumption" (p. 19).

This essay addresses some of the important issues concerning content analysis sampling, reliability, and computer coding. Sampling merits discussion because it is at the heart of the research process and determines the generalizability of results. Lovejoy et al. (2014) found: "A majority of the articles did not use a census or probability reliability sample and were not transparent about the sample selection process" (p. 220). Reliability is addressed because of continuing debate about the appropriate reliability coefficient (Feng, 2014; Gwet, 2008; Krippendorff, 2012; Potter & Levine- Donnerstein, 1999; Zhao, Liu, & Deng, 2012). Finally, the use of computer-based content analysis, which includes algorithmic coding, to save time has raised a number of issues.

Following the discussion, the essay summarizes current best practices for conducting and reporting content analysis in order to help scholars and students use the content analysis method, to help reviewers evaluate such research, and to stimulate further methodological research.

Sampling

Random Sampling

Today's content analysts face significant challenges analyzing digital media content in "Internet time" (Karpf, 2012). Traditional content analyses have long featured practical reliance on well-archived and indexed content, e.g., newspapers or broadcast news captured by Vanderbilt's television news archive service. Internet content, however, is more challenging given its sheer volume and the fact that its population is unknowable. It is ephemeral in nature, public data are limited, and there is "noise" introduced by spammers and fake social media accounts. As Riffe et al. (2014) noted, the universe of online posts or Tweets is "unlimited and unknowable and inherently unstable over time" (p. 168). Thus, it becomes difficult to construct scientific probability samples, which requires that every unit in the population has an equal chance of being included in the sample and that inclusion/exclusion of any particular unit be based on random selection and not any potential researcher selection bias.

A probability sample allows inferences about population statistics without observing every unit of that population. The extent to which the sample accurately "mirrors" the distribution of units in the population is the extent of the sample's representativeness (and external validity). Probability samples are the gold standard of social scientific methods, including content analysis, because the representativeness of a sample statistic (e.g., a percentage or a mean) can be easily measured using margins of error (e.g., +/- 5%) and confidence intervals.

On the other hand, units for a study may be selected on the basis of convenience (e.g., condom ads individuals have uploaded to YouTube, likely an incomplete subset of such ads, or all Tweets that can be collected because they share the common hashtag, #BlackLivesMatter) or purposively because they represent the "natural history" of an event (e.g., all newscasts aired from the first to the last day of the Democratic National Convention). Riffe et al. (2014) describe a convenience sample as "a census in which the population is defined by availability rather than research questions” (p. 75) while purposive sampling is based on a “logical or deductive reason dictated by the nature of the research project” (p. 76), such as an ongoing or continuing event. But even with such justification, a purposive sample’s generalizability is limited, and relationships found in the data cannot be extended to content outside the event. And while a data set made up of any Tweets with #BlackLivesMatter might contain important public discourse about relations between police and African Americans in Ferguson, Missouri, it is impossible to determine if those Tweets represent all such discourse, what universe of discourse they came from, or the nature of that universe. The requirement that all units in a population have equal odds of being selected obviously becomes problematic if it is impossible to identify what constitutes a population. The representativeness of non-random samples, drawn from an unknowable universe, is pure conjecture. Tests of statistical significance with non-probability samples, while they may be calculated or computed, are of dubious value.

Probability sampling is “conservative” in the sense of establishing the most rigorous conditions for testing a hypothesis or answering a research question. Allowing chance (a coin toss, a random number generator, a table of random numbers) to “decide” which units to include in the sample removes the potential for any conscious or unconscious researcher bias in selecting units (consider sending a male student onto the university green to conduct a survey, and the likelihood of his oversampling attractive female students). By contrast, selecting a sample *because* it is convenient, or purposively *because* of a relevant characteristic, is by this definition less conservative.

This is not to dismiss studies that use convenience or purposive sampling—the point is simply that selection criteria for both are exercised as the researcher's prerogative. In addition to representativeness, a sampling method should be evaluated in terms of whether or not it *removes* the potential for such selection bias. At minimum, researchers need to make it clear that they realize limitations of their samples, before reviewers bring those limitations to the authors' attention. Equally important, they should realize that estimating sampling error for a non-probability sample makes no sense.

Sampling with Keyword Searches

The use of electronic databases—whether commercial databases (e.g., PR Newswire, Factiva, America's Newspapers, or LexisNexis) or searchable databases collected or compiled by researchers (e.g., 60,000 Tweets sent during the Arab Spring protests [Lewis, Zamith, & Hermida, 2013])—and keyword searches used to compile samples from these databases also poses a significant challenge to drawing representative samples. Examining 198 content analyses published between 2000 and 2005 in six top communication research journals, Stryker, Wray, Hornik, and Yanovitsky (2006) found that 42% used databases and keywords. The appeal of such searches, according to the authors, is "the capability to retrieve a large quantity of relevant items with a single search term, thereby providing easier application of random sampling methods" (p. 414). Yet does random sampling from what might be a massive but non-representative collection of units yield valid results and inferences? Is an initial set of items with a common hashtag or that meets keyword definitions a population, a convenience sample, or a purposive sample? It is certainly not a probability sample, because of the inherent flaws in keyword searches and questions about the comprehensiveness and lack of comparability of databases (see Hansen, 2003; and Tankard, Hendrickson, & Lee, 1994).

Generating a complete collection of content with keyword searches depends on the terms used in the search. Of the 83 content analyses examined by Stryker et al. (2006), only 39% provided the search term and 6% discussed the search term's validity. Yet choosing search terms is often a subjective decision made by a researcher, and individual search terms in particular can lack precision in identifying "relevant" content. Sobel and Riffe (2015), for example, used LexisNexis to identify *New York Times* stories about Nigeria, Ethiopia, and Botswana to explore the importance of U.S. economic interests in of coverage of those nations. Using the country names as keywords, they found 7,454 news stories mentioning at least one of the three; however, screening revealed that 91% of the "hits" did not have one of the countries as "primary focus." In terms of tradeoffs among efficiency, validity, and context, this example points to the risk of using single keywords to locate content and to the need for precision in identifying the terms one selects in order to identify "relevant" content.

In order to limit the role of individual subjectivity, researchers should draw upon the literature and previous studies to assemble multiple keywords or keyword strings that offer more than face validity. Just as a good attitude or belief measure uses multiple items, a keyword search should have content validity (e.g., represent different facets of the same concept). Studying news coverage of cancer in 44 major U.S. newspapers (they planned to study 50 top papers, but Lexis-Nexis offered access to only 44, another problem with database content analyses), Stryker et al. (2006) developed a search string that named more than two dozen types of cancers or terms related to human cancer (they specifically excluded mentions of feline leukemia and the astrological sign, Cancer).

Additionally, researchers should conduct a formal test of a search string's "recall" and "precision." According to Stryker et al. (2006), recall is a measure of a search string's ability to retrieve relevant articles. Precision is a measure of whether the retrieved articles were relevant. There is some tradeoff between the two: the more precise a search string is, the more likely it will also fail to recall some relevant articles. Recall is measured by first establishing a broad search criterion that is likely to capture all relevant articles in a database. Those articles are then coded—using a set protocol, the reliability of which must be established—for whether they are relevant. Then a more precise search term is applied, and the researcher measures what proportion of relevant articles was retrieved: (relevant articles retrieved/relevant articles in database). Precision measures what proportion of all retrieved articles was relevant: (relevant articles retrieved/all articles retrieved).

Measuring recall and precision gives an estimate of sampling error associated with a search term and provides a correction coefficient (precision/recall) that can be used to "correct" sample statistics. For example, if one were trying to estimate the number of articles about violent crime in a local newspaper and the precision associated with a search string was .8 (that is, 80% of articles retrieved were relevant), and recall was .5 (50% of relevant articles in the database were retrieved), the correction coefficient is .8/.5=1.6. Thus, if the search string suggested that there were 55 articles about violent crime, a more accurate estimate correcting for the sampling error associated with the search string would be 55*1.6=88 articles (a correction coefficient of less than one would suggest that a given search string over-estimated the number of articles). Stryker et al. (2006) suggest that the correction coefficient produces inaccurate estimates over short periods of time (a day or week), but provides more accurate estimates over longer periods of time (a month, or a year of coverage).

Completeness and Comparability

Another challenge content analysts face is whether databases used in a study are comprehensive and comparable. Occasionally, researchers gather data from different databases for the same study. In their study of coverage of abortion protest before and after 1973's Roe v. Wade decision, Armstrong and Boyle (2011) used Pro-Quest Historical for 1960-1973, but Pro-Quest Historical *and* Pro-Quest National Newspaper for 1974-2006. Past use of databases has dealt primarily with text, but issues of compatibility can arise for any form of data because archiving software varies. Whether the two databases used similar indexing and archiving procedures is an empirical question.

Riffe et al. (2014) describe an elementary approach to answering that question: All but one major Ohio newspaper in a particular study were included in Lexis-Nexis, while the missing newspaper was included only in America's Newspapers. To test the compatibility of the databases, a newspaper that was in both was chosen and the agreement between both databases in providing environmental news coverage was assessed. NewsBank returned 179 articles and Lexis-Nexis returned 141. Wu (2015) studied news coverage of post-traumatic stress disorder in newspapers drawn from LexisNexis and America's News. To assess the ability of the two databases to yield similar results, she used a non-study health term ("Ebola") for a trial search over the same time period from the same newspaper in both databases, finding a matching rate of 94%.

Other problems include using different search engines to examine what is purportedly the same content. Weaver and Bimber (2008) compared LexisNexis and Google News searches of *New York Times* coverage of nanotechnology, finding only "modest" agreement (71%). They note also that wire service stories are often removed from newspapers before the stories are archived in the newspaper database, which might pose a particular challenge for studying newspaper coverage of foreign issues.

These are some of the concerns editors, reviewers, and scholars need to consider as they design studies involving archives, databases, and text searches. Methodologists should continue to scrutinize issues in accessing and managing content data. With the continued developments in storage and custom computer programming options, more innovations in content analysis are likely. Nonetheless, "content analysts need to subject such data to scrutiny and 'traditional' standards for sampling and validity" (Riffe et al., 2014, p. 167). Again, researchers need to recognize and disclose limitations of their samples and data.

Reliability

Reliability is a necessary, but not sufficient, condition for content analysis data to be valid. Reliability is not a "new" issue in content analysis, but in practice, questions about what reliability is supposed to represent and standards for reporting, including what coefficients to report and what levels of reliability are considered "acceptable," continue to vex novice and senior scholars alike (Feng, 2014; Gwet, 2008; Zhao, Liu, & Deng, 2012). Reliability takes two forms: intracoder reliability, which involves a coder's consistency across time, and intercoder reliability, which involves consistency across coders. Minimally, content analysis requires that intercoder reliability must be tested and reported. Intracoder reliability should also be established when the coding process will run for an extended time period.

Although it is easy to think of reliability as a property between or within coders, it is important to remember that the primary aim of inter- or intracoder reliability checks is to test the reliability of *the coding protocol*, and the protocol's ability to result in consistent categorization of content. This goal reflects the need to replicate research with a range of relevant content, irrespective of particular groups of coders and their idiosyncrasies. Social science is probabilistic. It aims to establish a high probability that relationships exist in the relevant populations. The probability increases as scholars replicate research by applying the same design and measures to the same and related content.

Furthermore, the objective must not be to establish a "minimum" level of acceptable reliability for a protocol. Rather, the protocol should be designed with the objective of producing the most reliable (and valid) data possible. In our experience, if time is spent on accomplishing the latter objective, reaching minimal reliability levels will usually take care of itself.

The question of what constitutes acceptable minimum levels of reliability has no definitive answer. It is an area ripe for empirical research about the relationship between reliability levels and valid conclusions from data. Krippendorff (2004a) suggests that scholars rely on variables with alphas above .8 and use variables with alphas between

.667 and .8 for tentative conclusions. Riffe et al. (2014) suggested that if the coefficient does not exceed .8, the author should provide a detailed argument about why the variables are reliable and valid.

Variable complexity can affect the difficulty of achieving a minimum level of reliability. This complexity often represents the degree to which understanding symbols depends on connotative versus denotative meaning (Riffe et al, 2014). For example, coding news story topic is easier than coding the valence (positive or negative leaning toward an object, person, or issue). However, difficulty of coding is not a reason to lower levels of acceptable reliability. The authors have conducted or supervised multiple studies using valence variables that have exceeded reliability coefficients of .80. As Krippendorff (2004) wrote: “Even a cutoff point of α = .80—meaning only 80% of the data are coded or transcribed to a degree better than chance—is a pretty low standard by comparison to standards used in engineering, architecture, and medical research” (p. 242).

With regard to what coefficients should be reported, simple percentage of agreement, the number of agreements among coders divided by the number of decisions, was initially used as a reliability measure (Holsti, 1969). Additional coefficients were developed because simple agreement might contain agreements among coders that occur for some reason other than the protocol.

Consider the most commonly used reliability coefficients in content analysis: Scott’s Pi, Cohen’s Kappa, and Krippendorff’s Alpha. All three share the characteristic of estimating error *attributable* to chance agreement, based on the *measured* agreement. The debate about which reliability coefficient to use centers on the best way to calculate error when coders actually agree. (The mathematical implications of the assumptions underlying each coefficient and how they treat “agreement error” are more complex than can be discussed here. Readers are referred to the various citations below.) Agreement error can result from systematic problems in protocols, poor training, coders’ failure to understand their role, or guessing by coders (chance error). Because there is no way to measure these false agreements, reliability coefficients use calculations of chance agreement as an estimate of the agreement error.

Among the three coefficients commonly used for content analysis, Krippendorff proposes that Alpha should become the accepted coefficient (Hayes & Krippendorff, 2007; Krippendorff, 2004b). He argues (Krippendorff, 2011) that it is superior to Kappa because it treats coders as independent, and (Krippendorff, 2004b) that it is superior to Pi because it adjusts for small sample sizes and can be used with multiple coders and all levels of data (nominal, ordinal, interval, and ratio). With nominal-level variables, two coders, and a large sample, Pi and Alpha provide the same values. Although Alpha and Kappa are often close in value (Gwet, 2014), under some conditions, Kappa will exceed Alpha, which means that if the Alpha coefficient exceeds an acceptable level, so will Kappa.

Scholars have criticized Alpha, Pi, and Kappa for a variety of reasons (Feng, 2014; Gwet, 2008; Zhao et al., 2012). Perhaps the most common criticism is that they can yield lower coefficients even when the levels of simple agreement are high (Feng, 2014; Gwet, 2008; Zhao et al., 2012), as can happen with skewed distributions (e.g., most of the coded units are in one category; see Riffe et al., 2014, pp. 119-120). Krippendorff has acknowledged that chance-corrected agreement coefficients can be more sensitive to rare cases than to frequent ones (Krippendorff, 2012), has suggested that variables with little variance may not be very important (perhaps reflecting the researchers' inadequate familiarity with the content to be coded), and has suggested that the high agreement/low reliability phenomenon might also represent insufficient sampling for testing reliability (Krippendorff, 2011, 2012). That is, he suggests that studies with infrequently occurring categories use stratified sampling to ensure that all categories are adequately represented in reliability tests.

However, there are populations of content whose skewed category representation is meaningful. For example, over the years, representation of minority groups and the elderly in movies and television has been so small as to be almost non-existent (Mastro & Greenberg, 2000; Signorielli, 2001). Clearly, coefficients that do not accurately measure the reliability of some forms of data (because of skewed distributions) make it difficult for scholars to publish studies about the antecedents of content with skewed distributions.

In response to this concern, two options have been proposed. Potter and Levine-Donnerstein (1999) suggested that expected agreement should be calculated by using the normal approximation to the binomial distribution (rather than calculating it based on the measured agreement). This involves calculating the probability n coders would agree on a decision when facing k options. In addition, Gwet (2008, 2014) suggested coefficient AC_1 for inter-rater reliability specifically to deal with this phenomenon. Gwet (2104) developed AC_1 for nominal-level data based on dividing scoring decisions into hard-to-score and easy-to-score. He based his coefficient " . . . on the more realistic assumption that only a portion of the observed ratings will potentially lead to agreement by chance" (2014, p.103). In a series of comparisons with other coefficients, including Alpha, Pi, and Kappa, AC_1 values were lower than simple agreement but higher than the three usual coefficients. Gwet (2014) has extended AC_1 to ordinal and interval data and called it AC_2.

Gwet's coefficients have been adopted extensively in health fields, but the process of medical diagnosis differs from coding media content because content analysis deals with symbolic meaning rather than physical manifestations and because content analysis involves a written protocol used by coders. Krippendorff argued that AC_1 is inadequate because of its "odd behavior" and because "its zero value occurs when all coincidences are equal" (Krippendorff, 2014, p. 490).

It is possible that Gwet's coefficients might be acceptable replacements for Alpha in situations other than high agreement and low reliability that comes with skewed data, but that possibility cannot be determined here. One study in the health field compared Kappa and AC_1 for use with assortative transmission of infectious diseases (Ejima, Aihara, & Nishiura, 2013) and found that AC_1 was superior with regards to the skewed data problem, but they concluded that AC_1 was harder to interpret than Kappa. The debate about whether AC_1 could replace Alpha needs a more extensive discussion of the mathematics underlying the two coefficients, empirical research about coder behavior, and meta-analyses of reliability levels for the coefficients and implications for validity. However, researchers must have some way of establishing and reporting reliability as that debate continues, which will be addressed in the best practices section.

The Distinction Between Human and "Algorithmic" Coders

Debates over reliability could be rendered moot by future applications of the algorithmic coder, a computer application that assigns numeric values to attributes of media content based on a set of programmed rules. An algorithmic coder has two principal advantages over its human counterpart: computers are 100% reliable and more efficient than human coders. Thus, the algorithmic coder reduces the costs in time and money of using human coders, and facilitates analyses of “big data” sets. To the extent that continued use of content analysis methods is slowed by the method’s inherently resource-intensive nature (Conway, 2006), the algorithmic coder is heralded as a significant advancement.

We will not restate Zamith and Lewis’s (2015) detailed comparison of the algorithmic and human coder. However, because the promise of 100% reliability and increased efficiency has an understandably strong allure, we will discuss methodological concerns associated with using computers to code content. Additionally, because of unique methodological processes and challenges associated with the algorithmic coder, we argue that these concerns are distinct from the processes associated with content analysis. Thus, use of an algorithmic coder should be considered a unique research method: *algorithmic text analysis (ATA).* (That is not to dismiss a “hybrid approach,” which leverages digital tools to collect, sift, and organize content in order to improve the efficiency and reliability of the human coder [Lewis et al., 2013; Zamith & Lewis, 2015]).

Algorithmic text analysis’ dominant concern is validating computerized coding of complex human language (Grimmer & Stewart, 2013). Improvements in natural language processing allow the algorithmic coder to perform complex tasks such as opinion mining and sentiment analysis, identifying, for example, irony and humor based on context. However, these nascent techniques carry computational challenges involved in training computers to have the human coder’s complex, contextual understanding of language’s nuances (Pang & Lee, 2008). The algorithmic coder can also examine the network structure of content, such as network analyses of hyperlinks (Park & Thelwall, 2003) or sources named in a news article (Morgan, 2015).

Algorithmic text analysis remains best suited to analyses of particularly manifest variables of digitally well-archived/indexed material (Zamith & Lewis, 2015). For example, the most common application of algorithmic text analysis is to count the presence of words in a pre-determined dictionary. Qin (2015) compared how frequently different hashtags were used on social media with how frequently different words were used in traditional media to describe Edward Snowden, the former government contractor turned leaker of government secrets, as a hero or traitor. More sophisticated analyses attempt to understand more complex meaning units—for example, news media frames—based on *co-occurrence* of words in a given unit of analysis. For example, Luther and Miller (2005) used a cluster analysis, which identifies sets of words to identify unique frames that pro- and anti-war groups used during demonstrations during the 2003 U.S.-Iraq war.

Reducing nuanced, complex human language to particularly manifest variables, such as word counts or clusters that can be quantified, poses a particular challenge for making valid inferences. Because use of the algorithmic coder has been interpreted as representing the same content analysis method, one way to validate the algorithmic coder's data would seem to be to compare them to data generated by human coders.

In reality, though, the processes associated with each coding approach are unique. For example, a human coder can recognize that "walk," "walking," and "walkable" are all words that reference the same activity, whereas texts need to be "stemmed" before the algorithmic coder starts its task, lest these words are confused as having fundamentally distinct meaning. It is also necessary to remove “function words” that serve grammatical purposes rather than conveying meaning, as well as rarely occurring words unlikely to meaningfully distinguish one unit of analysis from another (Grimmer & Stewart, 2013). These data cleaning processes are unique to use of an algorithmic coder; hence our contention that use of an algorithmic coder represents a distinct research method.

Furthermore, the distinct processes associated with the distinct methods produce distinct data. Despite the assumption that a human coder represents the “gold standard” against which the algorithmic coder can be validated (e.g., Conway, 2006), the algorithmic coder is capable of analyzing texts at a level of granularity—i.e., counts of individual words—that humans cannot reliably replicate, particularly over a large number of “big data” study units (Grimmer & Stewart, 2013). The two distinct methods of coding media texts cannot be presumed to produce identical data, and validating the algorithmic coder against the human coder may be a straw man. Because the methods do not produce identical data does not mean that the human coder’s data are more valid. Both methods have unique sources of measurement error. To establish external validity, content data gathered via both human and algorithmic coders should be compared to external, theoretically relevant variables not gathered via either method (Short, Broberg, Cogliser, & Brigham, 2010).

Algorithmic text analysis is also distinct from content analysis in that the former does not have a process analogous to establishing reliability of the content analysis coding protocol. That is not to say, however, that human subjectivity and error are insignificant factors in algorithmic text analysis. An algorithm is a set of steps, necessarily programmed by a *human*, which the computer follows. Given the exact same set of coding instructions, a computer will execute those commands with perfect reliability. That said, the *human process* of generating the algorithm can be subjective (Grimmer & Stewart, 2013), and oftentimes a single missing or misplaced character—human error—can significantly alter the meaning of a computer command. However, and perhaps because of the often repeated refrain that the computer is 100% reliable, algorithmic text analysis has yet to establish parallel procedures for estimating error associated with individual subjectivity and error. Because prior labels for this method (i.e., computer assisted text analysis (CATA) [Popping, 2000]), fail to capture the subjective *human* processes of generating the algorithm, we prefer “algorithmic text analysis (ATA).”

The algorithmic coder also does not double-check data for researcher error. Studies that rely on keyword searches lack perfect precision; articles that use an off-topic keyword in an unrelated context are included in the sample (Stryker et al., 2006). Human coders can be directed to set these off-topic articles aside. The algorithmic coder, however, classifies all articles in the sample, introducing error. Thus, while formal estimates of article recall and precision should be part of any study that analyzes media content gathered by keyword searches, estimating precision is particularly important in studies that use algorithmic text analysis. It is also important to recognize that the “big data” social media-based studies to which the algorithmic coder is often applied have other data quality issues, such as spammers and fake online profiles that have the potential to skew research data (Karpf, 2012; Zamith & Lewis, 2015). Even in algorithmic text analysis, there is “no substitute for careful thought and close reading” and the method “require[s] extensive problem-specific validation” (Grimmer & Stewart, p. 267).

Close reading is also essential because “training” the algorithmic coder requires that every single detail of the coding rules be *explicitly written* out in the algorithm. In content analysis, coding training is done orally, and written coding instructions are shaped by back-and-forth discussions between coders that may or may not be documented in the coding protocol (Hak & Bernts, 1996; Zamith & Lewis, 2015). Because the coding rules for algorithmic text analysis must be written out in full detail, the method is potentially more transparent and replicable, but only if one uses open source tools. The use of proprietary, commercial software that requires secrecy for competitive advantage, such as Crimson Hexagon—used by Ceron, Curini, and Lacus (2013) to study the political preferences of residents of Italy and France—actually decreases transparency and replicability. In order to increase transparency and aid replication, the full algorithm must be available to other researchers.

Unfortunately, advances in digital, Internet based tools for conducting analyses of media content have not necessarily translated into better practices for sharing related research materials. While individual researchers may share software and data sets on popular code-sharing websites like GitHub (https://github.com/) and university data repositories, both algorithmic text analysis and content analysis would be enhanced by a standard scholarly repository for sharing open source software, coding protocols, and code sheets.

Our use of “text analysis” differs from “textual analysis,” itself a misused label that encompasses rhetorical analysis, narrative analysis, discourse analysis, semiotic analysis, critical analysis, etc. (Neuendorf, 2002, p. 5-8). Additionally, our focus on “text” is deliberate: while one strength of content analysis is that it can be applied to text, photographic, and audio-visual content, computer software used to analyze text cannot also be applied to audio-visual content. Although an algorithmic coder could be used to analyze visual content—Zhu, Luo, You, and Smith (2013) used face recognition software to analyze social media images of Barack Obama and Mitt Romney during the 2012 presidential election cycle—the algorithmic coder has primarily been applied to *text-based* media content.

Having computers able to read and meaningfully classify complex human language and visual content could still be decades, years, or weeks away. Though algorithmic text analysis continues to develop, content analysis using a human coder will remain a mainstay of social science methods for the foreseeable future.

Best Practices

The best practices discussed below were developed through research and experience and have been codified in articles and texts. All are based on the application of content analysis as a social science method, which assumes that empirical results will be replicated and that replication requires transparency of reporting. Articles must contain enough detail to allow replication and to allow social scientists to improve the method through extension.

This section will address both the standards for conducting and reporting content analysis. The last subsection will address the area of algorithm text analysis. Some standards may appear obvious and, perhaps, simplistic to some scholars, but research (Lovejoy et al., 2014) indicates that some researchers are unfamiliar with these standards.

Standards for Conducting Content Analysis

Study Sampling and Design

A1. Develop an explicit written protocol that can be shared with other researchers.

The replication of results requires that a set of guidelines, called a protocol, be written to instruct coders how to assign values to content units. The reliability of a data set should be conceived as resting with the protocol and not with the coders because replication will occur with a different set of coders. Replication also requires that scholars be willing to share their protocol with others who are studying similar content.

A2. When using a search program to identify content, do the following: (1) Consult the literature and previous research to ensure that the search term/string addresses as many aspects of the focal concept and captures as many relevant articles as possible. (2) Test the search term, using the precision/recall criteria described by Stryker, et al. (2006). (3) If multiple databases are used in the same study, find a way to test their comparability and completeness, as described in Riffe et al. (2014) and Wu (2015).

Evaluating the contribution of any research article requires addressing data validity and generalizability. Do the data accurately reflect the range of content being studied? To what larger group, if any, do the results apply? Answering these questions requires careful planning of the search process and then appraising its outcome. Any study using a search program should report the full details of the process.

A3. The decision to select a probability, purposive, or convenience sample for a study should reflect the nature of the project—whether it is exploratory or builds on existing scholarship and whether it tests theory or not. The nature and selection process of the study sample should be clearly described. If a non-probability sample is used, it should be justified and its limitations specified. Identify, when possible, the universe, population, and sampling frame used in your study.

Because the goal of social science is to build generalizable theory, scholars should use as representative a sample from as large a population of content as the study will allow. Relationships that apply to a large number of outlets are generally more useful for prediction and explanation. However, there are situations when non-probability samples are necessary and situations when they are used because of resource limitations. Whatever the situation, the reasons behind the sampling process and the impact on the generalizability of the data should be explicitly addressed.

Coding and Reliability

B1. Two or more coders (three or more is better) should code content units independently of each other. At least one of the coders should NOT have developed the protocol.

The need for independence of coding should be obvious. Coding as a group can result in some individuals in the group having more influence than others; reliability will no longer be vested in the protocol and replicability using the protocol becomes impossible. The suggestion that coding should involve three or more coders may raise concern because this increases time needed to code and to train. However, differentiating the sources of error in the protocol becomes easier when more than two coders are involved. Using three or more coders allow a larger number of coder pairs to help analyze the sources of coder disagreements.

Because content analysis coding is time consuming and expensive there is a tendency for the protocol developer also to code. This can be problematic because the developer, in effect, has more "training" and may apply the protocol differently than other coders. Similarly, having all the coders participate in the protocol development could reduce the independence of coders. However, if one or more coders did not develop the protocol, the presence of developer bias in coding can be evaluated.

B2. When coders disagree on the values for content variables during the reliability check, the variable values should be randomly assigned to the cases on which there is disagreement. Articles should report how coding reliability problems and disagreement were resolved (retraining and re-testing, coder consensus, or dropping the variables, etc.).

Because the reliability check involves two or more coders coding the same content, there will be some content units with more than one value assigned by coders (disagreements). Previously used processes for resolving disagreements include the creator of the protocol deciding the final values, a discussion among the coders aimed at reaching consensus, and the random assignment of values. The first two introduce some unknown level of personal bias into the data. The third approach introduces random bias that is less likely to influence conclusions from the data. Of course, the higher the value for the reliability coefficients, the less bias will be introduced through the disagreement resolution process. Any adjustments should be reported in full within the text or footnotes of the article.

B3. Coders should practice with similar non-study content until intercoder reliability is established before they begin coding study content. A brief mention of this process should be included in the article.

The intercoder reliability check should occur with study content as it is coded. If reliability is not achieved, the coders have to be replaced in the recoding process because recoding content violates independence across coders and time. As a result, coders need to practice with content that will not be in the study but is similar to the study content in complexity. Practice with non-study content should continue until the protocol and coders can produce reliable data. At that point, the coding of study content can begin.

B4. Always conduct an intercoder reliability check for each variable using one or more reliability coefficients that takes chance into consideration.

The need for replication requires that all variables be checked individually for reliability. As discussed above, the reliability check requires a coefficient that takes error (chance agreement) into consideration. Establishing reliability for all variables also increases the possibility that protocols can be used across research projects and that measures of variables used in multiple studies can be standardized.

B5. Given the controversy over which reliability coefficient is appropriate, we suggest that the authors calculate at least two measures of reliability—simple agreement and Krippendorff's Alpha. Gwet's AC_1 (or AC_2) should also be calculated when the data have high levels of simple agreement but a low Alpha.

We suggest that simple agreement be reported in a footnote for all variables, even though it should not be used for determining the reliability of the variables. Simple agreement is used in calculating reliability coefficients, and if one knows the value of the coefficient and simple agreement, expected agreement can be calculated. Simple agreement also might be useful in discerning variables with high agreement and low reliability. Second, Krippendorff's Alpha should be reported if the data do not exhibit the high agreement and low reliability phenomenon. Alpha continues to be the most versatile of the commonly used coefficients. Third, if simple agreement is high and Alpha is low, Gwet's AC_1 or AC_2, whichever is appropriate, should be reported in lieu of Alpha. As in all studies and with all methods, authors should justify why they use any reliability coefficient.

B6. The intercoder reliability check should be based on a probability sample from the population or on a census of content units used in the study. Every value for every variable should be present in the reliability sample, which may require additional probability sampling. Report the selection method and the number (not percentage) of units used in the reliability check. The process for determining the number of reliability cases should be explained and justified with a citation.

Reliability coefficients should be calculated on a representative sample of the study content population. Otherwise, the protocol cannot be evaluated in a way that reflects the entire range of content. As an extension of this reasoning, the reliability sample should have content that represents all categories for every variable in the protocol (Krippendorff, 2004a). The researcher should evaluate the sample for this requirement, and if the sample content does not include all variables and categories, additional content units will need to be selected randomly until the standard is met. Exactly how many additional units are needed will vary with the population distribution, and there is no way to know this in advance.

Similar to describing the full sample, the reliability sample should be clearly described so that it may easily be replicated. The reliability sampling description should include how units were chosen (e.g., random, census, convenience, etc.) as well as details such as identification of clusters in the case of cluster reliability sampling, skip intervals in the case of systematic random reliability sampling, or how content was stratified during selection. The total number of reliability units as well as the specific number of units from each different cluster or outlet, if appropriate, should be provided.

B7. The number of units used in the reliability sample should <u>*NOT*</u> *be based on a percentage of population because this will usually generate a sample too large or too small.*

Some general research methods books have recommended that a certain percentage of the study content should be selected (Kaid & Wadsworth, 1989; Wimmer & Dominick, 2003) for the reliability check. However, probability theory suggests the result will be fewer or more units than necessary. A sample's representativeness is based on three elements: 1. The size of the sample, 2. the homogeneity of the study content (population for the reliability sample), and 3. the percentage of the population in the sample. The last element is relatively unimportant until the sample becomes a large proportion of the population. If one takes 7% of a small population (e.g., 500), the number of units in the reliability sample would equal 35, which might not meet the requirements discussed in B6. On the other hand, 7% of 10,000 would yield 700 units, which is likely more than needed for a representative sample. At least two approaches are available for selecting a representative sample that does not depend on percentage of population. Lacy and Riffe (1996) developed an equation for selecting a sample for nominal variables based on four factors—total units to be coded, the desired confidence level, the degree of precision in the reliability assessment, and an estimate of the actual agreement had all units been included in the reliability check. Krippendorff (2013, p. 321-324) suggests a process based on the number of coders, the number of categories (values) used in the coding, the lowest acceptable level of Alpha, and the acceptable level of statistical significance. Evidence as to which is the "better" approach needs to be developed. Researchers should access these sources for more detail.

B8. If the coding process takes an extended period of time, the researcher should conduct more than one intercoder reliability check and at least one intracoder check.

If the content analysis extends over a long period of time, the reliability of coding could increase, decrease, or remain about the same level. The only way to establish that the coding remains reliable is to check it at more than one point in time. The test of reliability across time requires a test of both intracoder reliability and intercoder reliability. Of course, there is no standard for what constitutes an "extended period." There are at least two factors to consider. First, the longer the coding period, the greater the need for an intracoder and multiple intercoder reliability checks. Second is the "regularity" of coding. If coders are coding every day, the likelihood of reliability deterioration is less. The authors' experiences suggest that less than regular coding (at least every other day) would require an intracoder reliability check after a month, but even regular coding over more than two months should include an intracoder reliability check and a second intercoder reliability check.

B9. Reliability coefficients should exceed .8, unless the area is truly exploratory, and variables with coefficients lower than .7 should be dropped from the study.

These reliability standards are somewhat consistent with those recommended in the leading texts, except that Krippendorff (2004a) suggests that reliabilities as low as .67 could be acceptable in exploratory studies. All of these levels are somewhat arbitrary, but the higher the reliability the better. One concern is what counts as "exploratory." It is possible that an exploratory study could use content measures that have been previously used in studies. In this case, the relationship might be exploratory, but the content protocol is not. In order to provide consistent evaluations across time, authors need to include detail about their content measures and reviewers need to hold them to high standards if communication research is to advance.

B10. Articles should be transparent about study variables that ultimately failed to reach acceptable levels and were thus dropped. These dropped variables should be reported in a footnote.

Failure to reach acceptable reliability could result from a variety of causes, but this does not necessarily mean the variable or the protocol for the variable is useless. Science is cumulative, and reporting on efforts that are unsuccessful can nonetheless help advance communication research both theoretically and methodologically, and allow other scholars to learn from such experiences.

B11. Authors should report the number of coders employed, who supervised the coding, and how the coding work was distributed (what percentage of it was done by the PI or the coder[s], whether sets of coding units/assignments were assigned randomly to avoid systematic error, etc.).

Such information allows for better replication, extension, and evaluation of the data by readers.

Algorithmic Text Analysis as a Complementary Method

Although algorithmic text analysis is in its early stages of development, it can be an important source of complementary data, particularly for "big data" sets. Research has yet to develop widely accepted standards, but we will offer some suggestions that might be useful until more research sheds light on the best practices.

Currently, algorithmic text analysis (ATA) is best suited for well archived/indexed digitized data and for studies concerned with especially manifest variables (i.e., word counts). Researchers using ATA need to pay particular attention to establishing (and making explicit) the validity of the algorithms used in their analyses; human coders do not necessarily represent a valid "gold standard." As part of this process, algorithmic texts analysis should consider error associated with poor data quality (e.g., spammers and fake social media accounts). Data preparation and cleaning procedures (e.g., stemming), must be explicit in the write-up of the research. To aid replication, algorithmic texts analyses should use open-source software; algorithms should be made available online for other scholars' inspection and use.

When reporting ATA, researchers should explicitly report efforts to validate ATA measures and any data quality issues, as well as report any data processing/cleaning procedures. As with all research, methods decisions involving ATA should be reported in detail. The full algorithm (and any other essential details) should be reported or made available to allow for replication and for measurement standardization. As with content analysis protocols, sharing algorithms will allow researchers to built up the work of others and across time work toward standardize commonly used variables.

Conclusion

If science is to advance, all research methods must evolve and improve. Because all human verbal and mediated exchanges involve messages (content), content analysis is particularly important for the study of communication. Moreover, content analysis complements studies of the antecedents and effects of communication in a variety of fields. This essay addresses a few, but certainly not all, of the current issues engaging the content analysis research community. These issues require research and discussion to further the method's development. What reliability coefficient is best under what conditions? How can we establish the reliability of samples selected with digital searches? How can the validity of algorithmic textual analysis data best be established? We hope scholars will take up this calls to action.

The essay also presents our perspective on best practices. Because of space constraints, this list is incomplete. In addition, not all content analysis methodologists will agree with all the points, and the standards for algorithmic text analysis are just being developed. Scholars wishing to conduct content analyses should read the latest versions of the standard texts in the field (Krippendorff, 2013; Neuendorf, 2002; Riffe, et al., 2014), as well as methodological articles. It would be useful, however, if content analysts could come to a consensus on a set of best practices and standards for conducting and reporting research. Minimally, researchers should provide explicit detail on all methods decisions. We hope this essay makes a contribution toward helping with that process.

References

Armstrong, C. L., & Boyle, M. P. (2011). Views from the margins: News coverage of women in abortion protests, 1960-2006. *Mass Communication and Society, 14*(2), 153-177.

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

Ceron, A., Curini, L., Iacus, S. M. (2013). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society*, *16*, 340-358. doi: 10.1177/1461444813480466

Conway, M. (2006). The subjective precision of computers: A methodological comparison with human coding in content analysis. *Journalism & Mass Communication Quarterly*, *83*, 186-200. doi: 10.1177/107769900608300112.

Ejima, K., Aihara, K., & Nishiura, H. (2013). On the use of chance-adjusted agreement statistic to measure the assortative transmission of infectious diseases. *Computational and Applied Mathematics*, *32*(2), 303-313.

Feng, G. C. (2015). Mistakes and How to Avoid Mistakes in Using Intercoder Reliability Indices. *Methodology*, 11(1), 13-22.

Grimmer, J. & Stewart, B.M. (2013). Text as data: The Promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*, 267-297.

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.

Gwet, K. L. (2014). Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. *Advanced Analytics*, LLC.

Hak, T. & Bernts, T. (1996). Coder training: Theoretical training or practical socialization? *Qualitative Sociology*, *19*, 235-257. doi: 10.1007/BF02393420

Hansen, K. A. (2003). Using databases for content analysis. In G. H. Stempel III, D. H. Weaver, & G. C. Wilhoit (Eds.), *Mass communication research and theory* (pp. 220-230). Boston: Allyn & Bacon.

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

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

Kaid, L. L., & Wadsworth, A.J. (1989). Content analysis. In P. Emmert & L.L. Barker (Eds.), *Measurement of communication behavior* (pp. 197-217). New York: Longman.

Karpf, D. (2012). Social science research methods in Internet time. *Information, Communication, & Society*, *15*, 639-661. doi: 10.1080/1369118X.2012.665468

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

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

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

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 (eds.), *Communication Yearbook 36* (pp. 481-499). New York: Routledge.

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

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

Lewis, S. C., Zamith, R., Hermida, A. (2013). Content analysis in an era of big data: A hybrid approach to computational and manual methods. *Journal of Broadcasting and Electronic Media*, *57*, 34-52. doi: 10.1080/08838151.2012.761702

Lovejoy, J., Watson, B. R., Lacy, S., & Riffe, D. (2014). Assessing the reporting of reliability in published content analyses: 1985–2010. *Communication Methods and Measures, 8*(3), 207-221.

Luther, C. A. and Miller, M.M. (2005). Framing of the 2003 U.S.-Iraq war demonstrations: An analysis of news and partisan texts. *Journalism & Mass Communication Quarterly*, *82*, 78-96. doi: 10.1177/107769900508200106

Mastro, D. E., & Greenberg, B.S. (2000). The portrayal of racial minorities on prime time television. *Journal of Broadcasting & Electronic Media*, *44*(4), 690-703.

Morgan, J. (2015). Using technology and attribution in articles to examine ties between reporters and sources. Preliminary paper. Unpublished manuscript, Michigan State University, East Lansing, MI.

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

Pang, B. & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*, 1-135. doi: 10.1561/1500000011

Popping, R. (2000). *Computer-assisted Text Analysis*. Thousand Oaks, CA: Sage.

Park, H. W. & Thelwall, M. (2003). Hyperlink analyses of the world wide web: A review. *Journal of Computer-Mediated Communication, 8*, 0. doi: 10.1111/j.1083-6101.2003.tb00223.x

Potter, W. J., & Levine- Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research*, 27(3), 258-284.

Qin, J. (2015). Hero on Twitter, traitor on news: How social media and legacy news frame Snowden. *The International Journal of Press/Politics*, *20*, 166-184. doi: 10.1177/1940161214566709

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

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

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

Signorielli, N. (2001). Age on television: The picture in the nineties. *Generations*, *25*(3), 34-38.

Short, J. C., Broberg, J. C., Cogliser, C. C., & Brigham, K. H. (2010). Construct validation using computer-aided text analysis (CATA): An illustration using entrepreneurial orientation. *Organizational Research Methods*, 13, 320-347.

Sobel, M., & Riffe, D. (2015). “U.S. linkages in New York Times coverage of Nigeria, Ethiopia and Botswana (2004-13): Economic and strategic bases for news.” *International Communication Research Journal* (in press).

Stryker, J. E., Wray, R. J., Hornik, R. C., & Yanovitzky, I. (2006). Validation of database search terms for content analysis: The case of cancer news coverage. *Journalism & Mass Communication Quarterly, 83*(2), 413-430.

Tankard, J. W., Jr., Hendrickson, L. J., & Lee, D. G. (1994, August). *Using Lexis/Nexis and other databases for content analysis: Opportunities and risk.* Paper presented to the annual convention of the Association for Education in Journalism and Mass Communication, Atlanta, GA.

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

Weaver, D. A., & Bimber, B. (2008). Finding news stories: A comparison of searches using LexisNexis and Google News. *Journalism & Mass Communication Quarterly, 85*(3), 515-530.

Wu, L. (2015). “How do national and regional newspapers cover post-traumatic stress disorder? A content analysis.” Paper presented at Annual Convention, AEJMC, San Francisco.

Zamith, R. & Lewis, S. C. (2015). Content analysis and the algorithmic coder: What computational social science means for traditional modes of media analysis. *The Annals of the American Academy of Political and Social Science, 659,* 307-318. doi: 10.1177/0002716215570576

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

Zhu, J., Luo, J., You, Q., Smith, J. R. (2013, Dec.). Towards understanding the effectiveness of election related images in social media. Paper presented at the 2013 IEEE 13th International Conference on Data Mining Workshops (ICDMW), Dallas, TX. doi: 10.1109/ICDMW.2013.112