Considerations and Recommendations

for the Annenberg Media Health Coding Project

by

Kimberly A. Neuendorf, Ph.D.
Director, Communication Research Center
School of Communication
Cleveland State University
Cleveland, OH 44115
k.neuendorf@csuohio.edu

November 2, 2006

The Annenberg Health Media Coding Project provides an ambitious forum for exploring notions of human coding validity, reliability, and replicability. Notably, the Project's goal of making coding materials readily available to other researchers, providing full replicability, is a welcome contribution to the content analysis community.

What follows is a series of considerations and recommendations aimed at assisting the Project researchers in making decisions about sampling, unitizing, and coding, with an eye to maximizing *reliability*. This set of guidelines assumes that the reader is familiar with my earlier works on this topic (Neuendorf, 2002; 2006), and attempts to extend past work via new thoughts and more proscriptive recommendations.

Unitizing

A major challenge to reliability assessment is the development of clearly defined units of data collection. There may be a definitive set of rules for the identification of units (e.g., all

characters listed in Halliwell's Film Guide, a nicely grounded choice employed by the Project for film coding). More often, though, coders may be unitizing as they code, in which case a separate layer of reliability assessment is in order—the reliability of unitizing (Krippendorff, 2004).

Often, multiple units of analysis are employed in a content analysis. For example, a recent data collection at my university involved the analysis of feature films (a) at the whole-film level, (b) with each lead, major, or medium character as the unit, and (c) with production techniques and motifs measured with a five-minute time interval as the unit of data collection (Janstova & Neuendorf, 2006). In such instances of multiple units of data collection, it's important to clearly separate the coding—this is in essence *three* different content analyses. Any muddling of the units of data collection will result in coder confusion and fatigue.

In general, it is recommended that unitizing be done in such a concrete fashion that coders do not have to make decisions on the fly. Reliability is compromised whenever coders have difficulty in identifying units. For example, coding each discrete instance of smoking in a linear narrative will surely be a less reliably process than coding each character's smoking behavior, or coding whether smoking occurred in a five-minute interval.

Reliability Sample(s)

Optimally, at least two reliability subsamples will be selected for a given content analysis. One will serve as the content for a pilot reliability test before final coding commences (this pilot provides one last chance to change the coding scheme to maximize reliability); another will provide material for the final reliability test, conducted during the process of final coding. A number of options exist for the selection of reliability subsamples. The most common technique is to randomly select a subset of the main content analysis sample, usually about 10-20% of the full sample. Just as the full sample typically is representative of a larger defined population of

interest, the reliability subsample is viewed as representative of the sample. However, another option exists, similar to the choice of testing hypotheses in experiments by using only the extreme high and low groups. This second option is to select a reliability sample that maximizes the variance on key dimensions of interest (e.g., Potter et al., 1998).

This second option is particularly appealing in cases where many of the variables under examination are "rare event" measures, in which the targeted activity occurs in only a small proportion of the cases. The option calls for, in essence, oversampling for these rare events, (a) providing more opportunity for coders to become skilled at identifying these instances[i], and (b) producing variables within the reliability data set that have greater variance, which often corresponds to a higher reliability figure.

Content Analysis Measures: Not So Different

In general, the measurement of content analytic variables should be executed and evaluated in much the same way as survey and experimental measures. That is, individual measures need to have categories or levels that are exhaustive and mutually exclusive. Measures should be attempted at the highest possible level of measurement (e.g., counts rather than presence/absence indicators of an activity). Attention should be paid to individual variables' variances and distributions, and "transforms" of the variables should be made as needed. And, variables should be combined into scales when it makes conceptual and empirical sense to do so. When scales are constructed, internal consistency reliability should be assessed (e.g., with Cronbach's alpha).[ii]

Measures with good characteristics are more likely to result in reliable and valid outcomes. Additionally, measures that are clear and easy to use get a better response from

coders—just as clearly worded questionnaire items make for better respondent reactions in a survey.

However, one unique characteristic of [human-coder] content analysis measures that should be mentioned is their reliance on trained coders as part of the coding protocol. I'll repeat one piece of advice from my 2002 book: For good coder preparation, "train, train, and train." Nothing can replace solid coder training in the contribution it can make to reliability.

Reliability Statistics

As I've noted elsewhere (Neuendorf, 2002; 2006), reliability stats may be categorized as indicators of (a) agreement, (b) agreement beyond chance, and (c) covariation. Generally, it is *not* acceptable to present only indicators of agreement with no correction for chance (i.e., percetn agreement).

Which reliability statistics are most appropriate, then? My own recent scholarship has made me aware of the complex issues involved in answering this question. Several months ago, with a small team of co-researchers, I began to assess extant reliability statistics with regard to their performance characteristics; quickly, this effort expanded into a consideration of the sheer definition of reliability.[iii] Efforts to compare and contrast available reliability statistics continue.

Hayes and Krippendorff (2006) have made an explicit claim for the superiority of a set of reliability statistics for different levels of measurement devised by Krippendorff (2004), termed Krippendorff's alpha. Although they identify the set of statistics as a "family," the coefficients' roots lie in a number of statistics not usually linked and based on diverse statistical assumptions—Cohen's kappa, Spearman rho, and the ICC (intra-class correlation coefficient).

As noted elsewhere (Neuendorf, 2006), these coefficients should join the more frequently used indicators as the object of inquiry in a battery of independently-conducted tests, involving

both statisticians and content analysis practitioners (to assure that emergent criteria for evaluation are grounded in experience with the challenges of actual content analysis research). Such evaluative processes should examine the robustness of the statistics to violations of assumptions, describe their properties, assess their incremental advantage over alternative statistics, and establish their responsivity to a host of variations in conditions.

More generally, all of us involved in content analytic research need to examine the assumptions of each test we use. For example, the ICC assumes a variance-partitioning model, rather than an explicit covariation model, which may or may not meet the needs of a given researcher (Shrout & Fleiss, 1979).

Additionally, the development of new reliability statistics might be considered. For example, problems with achieving an acceptable level of reliability with "rare event" variables have been noted. Such problems follow from existing nominal-type coefficients' reliance on marginal probabilities that may be imbalanced, and correlational statistics' sensitivity to low variance and truncated range.[iv]

For now, lacking a comprehensive set of assessments for reliability coefficients, it is recommended that researchers use some of the more widely-accepted reliability statistics (see Lombard, Snyder-Duch, and Bracken, 2002, for a systematic review of intercoder reliability in the communication literature). Those with a richer "track record" provide us with greater basis for comparison with past work, and allow a more standard shared statistical "language" for discussion among scholars. Most of the frequently-used statistics are calculable via PRAM, an "alpha" program available online (Skymeg Software, 2006 (http://www.geocities.com/skymegsoftware/pram.html)) with an update add-on available from this author (the add-on calculates Fleiss' multi-coder version of Cohen's kappa (Fleiss, 1981;

Fleiss, Levin, & Paik, 2003) and Krippendorff's alphas (the latter is not yet fully validated)). PRAM is the only program I'm aware of that handles multiple coders and multiple variables simultaneously, utilizing an Excel-type database format that is compatible with SPSS.

The stats that I currently use in my own research are as follows: For nominal data—Cohen's kappa or multi-coder kappa; for ordinal data—not satisfied with the assumptions of Spearman rho, I typically drop down to a nominal analysis (if I'm not able to create an interval/ratio measure, which of course is preferable); for interval/ratio data—Lin's concordance coefficient (a "relative" of the Pearson correlation coefficient that takes coder differences in level into account; Lin, 1989)[v].

It has been recommended that reliability statistics be used as diagnostics (Neuendorf, 2006), so as to identify problematic variables, problematic coders ("rogue" coders; Neuendorf, 2002), and problematic variable/coder interactions. Reliability assessment may also result in the collapsing of categories within a single variable, or the combining of multiple variables into scales. Again, having an initial pilot reliability test gives the researcher an opportunity to conduct any desired diagnostics, and change the coding scheme as needed.

Variance and Reliability: A Simple Rule

Regardless of the selection of a particular reliability statistic, one truism holds—there is greater opportunity for reliability figures to be high for a variable that has a good amount of variance. There are several ways this might be achieved:

1. Select variables that past work has indicated hold good variance in the population under examination;

2. If selecting a set of indicators that measure the same general construct, be prepared to pool these indicators in order to achieve good variance. For example, in a recent study, we

measured whether various colored and diffusion filters were used in films. Due to rare occurrence of each filter type, we pooled these measures (Janstova & Neuendorf, 2006). Doing this may result in a loss in precision of prediction, but a gain in reliability;

3. Be prepared to combine categories within individual variables in order to achieve a better distribution on that variable.

Content Effects and Choice of Variables

The Annenberg Health Media Coding Project has already acknowledged the importance of examining the media effects literature as part of the process of identifying important variables. I heartily concur with this decision (see Neuendorf, 2002, chapter on an "integrative approach" to content analysis). The efforts to date to code for "modeling" potential are admirable, and should not pose a threat to reliable coding so long as concrete definitions are maintained. For example, smoking might be coded as shown performed by a "happy/content group" (indicating reinforcement and normative behavior), rather than coded as vaguely "high modeling potential."

The Medium

Although the Annenberg Health Media Coding Project has identified as a primary goal the development of measures that may be used cross-modally (i.e., in various media), it's worth considering the particular medium in several regards. First, there may be "critical" form variables that moderate the presentation of content in that medium. My own favorite example is the study of Music Television's portrayals of aggressive acts and cues that found that females were no more likely than were males to be the victims of aggression; however, when females were victims, they were significantly more likely to be shown in closeup, and were shown for a significantly longer length of time (Kalis & Neuendorf, 1989). The critical form variables of

shot type and shot length provided additional information about the presentation of the content (aggression) that was important to a full understanding of its reception by an audience.

In order to preserve the notion of cross-model measures, perhaps many of these "critical" variables may be framed as manifestations of "universal" variables that may be identified in any medium. For example, the visual emphasis provided by a closeup in film or television may be comparable to the emphasis given on a website via placement near the top of the page or with a larger font size. In both cases, the variable may be a measure of "intensity" (Marks, 1978; Neuendorf, 2002).

The particular medium will clearly also affect the sampling model (e.g., sampling frame, units of sampling) and even the definition of the population to which the researchers wish to extrapolate. Let us consider some choices with regard to population definition, and examine some issues relevant to some of the more problematic media being studied.

Defining the population. The Annenberg Health Media Coding Project is aimed at identifying content relevant to effects on youth. Thus, they have several choices for defining the populations of content to which they hope to generalize their findings. First, they may take a "message pool" approach, defining the population as the set of messages available via a given medium at a certain time (Kunkel et al., 1995, utilized a form of this approach (the "what's on" method) for their National Television Violence Study). For example, the population of television content may be defined as all programs airing/cablecasting on a wide set of broadcast and cable networks (note that even this broad-brush approach requires a selection and itemization of networks). Second, the researchers may take an "exposure-based" approach, defining the population as those messages most widely attended to by audience members. For example, a television program population may consist of the top-50 rated TV/cable programs.

Third, the researchers may take a "specific audience exposure-based" approach, in which the population becomes those messages most heavily attended to by a particular target audience. For example, the TV population may be defined as those TV/cable programs most heavily watched by youths aged 12-18. Each of these choices has some validity; what's important is that the decision process be clearly and fully reported.

Music videos. As with television in general, the delivery systems for this "medium" may vary quite widely. Music videos may be watched on broadcast TV, via cable TV, on video or DVD, or online, and this variety of delivery modes makes the definition of the population to which one wishes to generalize a more difficult task. Further, if coders are coding using different display or delivery systems (e.g., progressive scan DVD vs. online), variations in quality and look may jeopardize reliability assessment.

Music. Let's assume that the primary focus will be on music lyrics. However, it's worth considering some form variables such as music pacing and melodic complexity, elements that are sure to moderate the content's effects.

For content such as music lyrics that is entirely verbal (written or transcribed), researchers may wish to consider using CATA (computer-aided text analysis). A wide variety of programs now provide dictionaries intended to measure such constructs as optimism, aggression, and emotional tone (Neuendorf & Skalski, 2006b). Even if the provided dictionaries are found not to be useful, most CATA programs can operate as simple search tools, making sure, for example that no occurrence of a term such as "smoking" or "cigarettes" is missed. This can reduce coding and recording errors that can restrict reliability. At the same time, the obtuse and heavily symbolic nature of popular music lyrics will work against the validity of outcomes for any application of CATA analyses.

A list of CATA programs may be found at the website in support of *The Content Analysis Guidebook* (Neuendorf, 2002; Neuendorf & Skalski, 2006a; http://academic.csuohio.edu/kneuendorf/content).

Internet. The analysis of websites has perhaps posed the greatest recent challenge to content analysts. The fluid nature of the medium and the complex structure of the content frame have posed difficulties in unitizing and coding. And, defining the population is notably problematic for web studies. Ha and James' (1998) study of business websites used the archives of the Web Digest for Marketers as their population; however, this was in 1995, and a list of corporate websites was still conceivable. A more recent approach is Salinas' (2006) decision to define her population as the top 100 sites obtained in searches using the three search tools Google, Clusty, and Yahoo!

Some researchers have limited their analysis to just portions of websites—e.g., homepage only or just banner ads (An & Wachanga, 2005; Pashupati & Lee, 2003) or mission statements or corporate responsibility statements contained in corporate sites (Kemp & Dwyer, 2003; Penev, 2006). Others have extended the analysis to the entire site (Curtin & Gaither, 2003) or the homepage plus secondary pages (Salinas, 2006).

For content analyses of websites, it seems imperative that a "snapshot" approach be used for collecting the sample (Norris, 2003). For example, Curtin and Gaither (2003) downloaded entire websites, collecting their content twice, one month apart, in order to capture the "dynamic nature of the web" (p. 12). This freezing of the content is essential to reliability.

Future Initiatives

A number of future scholarly endeavors would help provide all content analysts with more guidance in the selection of their "tools" for reliability assessment. The aforementioned

set of tests of reliability statistics' characteristics, including Monte Carlo tests (Mooney, 1997) and/or bootstrapping techniques (Hayes & Krippendorff, 2006), could also provide new information on the statistics' sampling distributions (e.g., Petersson, Gill, & Ahlfeldt, 2002) and viable methods for establishing confidence intervals and tests of statistical significance for reliability stats.[vi] Additionally, it is hoped that the available statistics be examined and compared with regard to their response to changes in such conditions as: number of coders, number of cases, level of measurement, precision of measurement, presence of missing data, and distributional characteristics of a variable (variance, skew, etc.).

Additionally, there is need of a readily available software package that allows for multiple coders and multiple variables, and provides reliability diagnostics as well. With such a facility, reliability assessment may be more clearly viewed as a *process* of improving the content analysis coding scheme rather than a rigid indicator of success or failure.

References

An, D., & Wachanga, D. (2005). *An exploratory investigation into the role of ad visuals in multinational brands' local website advertising*. Paper presented at the annual conference of the International Communication Association, New York, NY.

Curtin, P., & Gaither, K. (2003). *Public relations and propaganda in cyberspace: A quantitative content analysis of Middle Eastern government websites*. Paper presented at the annual meeting of the International Communication Association, San Diego, CA.

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

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). Hoboken, NJ: Wiley-Interscience.

Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Physical Education and Exercise Science, 5*(1), 13-34.

Gray, J. H., & Densten, I. L. (1998). Integrating quantitative and qualitative analysis using latent and manifest variables. *Quality & Quantity, 32*, 419-431.

Ha, L., & James, E. L. (1998). Interactivity reexamined: A baseline analysis of early business web sites. *Journal of Broadcasting & Electronic Media, 42*, 457+.

Hayes, A. F., & Krippendorff, K. (2006, in press). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*.

Hubbell, A. P., & Dearing, J. W. (2003). Local newspapers, community partnerships, and health improvement projects: Their roles in a comprehensive community initiative. *Journal of Community Health, 28*, 363-376.

Janstova, P., & Neuendorf, K. A. (2006, in progress). *Empirical testing of auteur theory via content analysis: The case of Jane Campion.* School of Communication, Cleveland State University, Unpublished manuscript.

Kalis, P., & Neuendorf, K. A. (1989). Aggressive cue prominence and gender participation in MTV. *Journalism Quarterly, 66*, 148-154, 229.

Kemp, S., & Dwyer, L. (2003). Mission statements of international airlines: A content analysis. *Tourism Management, 24*, 635-653.

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

Kunkel, D., Wilson, B., Donnerstein, E., Linz, D., Smith, S., Gray, T., Blumenthal, E., & Potter, W. J. (1995). Measuring television violence: The importance of context. *Journal of Broadcasting & Electronic Media, 39,* 284-291.
Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics, 45,* 255-268.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research, 28,* 587-604.

Marks, L. E. (1978). *The unity of the senses: Interrelations among the modalities*. New York: Academic Press.

Mooney, C. Z. (1997). *Monte Carlo simulation.* Thousands Oaks, CA: Sage.

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

Neuendorf, K. A. (2006, in press). Reliability. In D. Kunkel, A. Jordan, J. Manganello, & M. Fishbein (Eds.), *Media messages and public health: A decisions approach to content analysis*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Neuendorf, K. A., & Skalski, P. D. (2006a). *The content analysis guidebook online*. Retrieved on February 13, 2006, from: http://academic.csuohio.edu/kneuendorf/content

Neuendorf, K. A., & Skalski, P. D. (2006b). Quantitative content analysis and the measurement of collective identity. In R. Abdelal, Y. M. Herrera, A. I. Johnston, & R. McDermott (Eds.), *Identity as a variable*. Cambridge, MA: Harvard Identity Project, under review, Cambridge University Press.

Norris, P. (2003). Preaching to the converted? Pluralism, participation and party websites. *Party Politics, 9*(1), 21-45.

Pashupati, K., & Lee, J. H. (2003). Web banner ads in online newspapers: A cross-national comparison of India and Korea. *International Journal of Advertising, 22,* 531-564.

Penev, E. (2006, expected). *Corporate social responsibility in websites: A content analysis.* School of Communication, Cleveland State University, Masters Thesis.

Petersson, H., Gill, H., & Ahlfeldt, H. (2002). A variance-based measure of inter-rater agreement in medical databases. *Journal of Biomedical Informatics, 35,* 331-342.

Potter, J., Linz, D., Wilson, B. J., Kunkel, D., Donnerstein, E., Smith, S. L., Blumenthal, E., & Gray, T. (1998). Content analysis of entertainment television: New methodological developments. In J. T. Hamilton (Ed.), *Television violence and public policy* (pp. 55-103). Ann Arbor: The University of Michigan Press.

Salinas, R. (2006). A content analysis of Latina web content. *Library & Information Science Research, 28*, 297-324.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428.

Endnotes

[i] Such "range" sampling is also suited to the selection of a *training* subsample (e.g., Hubbell & Dearing, 2003).

[ii] We might draw a comparison between this notion of multiple discrete measures and the concept of manifest and latent variables (Gray & Densten, 1998). Individual content analytic measures may be thought of as items in a scale; the individual items may be quite manifest, while the overall scale is seen as measuring a latent construct.

[iii] We have begun to carve out the different assumptions of an "intercoder reliability" approach vs. the "interrater reliability" approach more commonly found in clinical applications. The latter treats the raters more as experts, and acknowledges and allows for disagreements among them—indeed, their differences are sometimes valued and closely examined (Goodwin, 2001).

[iv] In a recent study of film content, this author encountered a number of "rare event" variables for which reliability was, typically, compromised. For one variable, only one coder recorded any instances of the target behavior, an explicit declaration of love by one character for another character. The variable obtained an unacceptable multicoder kappa of -.02, while percent agreement across the eight coders averaged 98%. For another variable, an unacceptable kappa of .22 was obtained, with a percent agreement of 91%. The unacceptable kappa was due not to disagreement as to whether the rather rarely-seen behavior occurred or not (hand-to-hand combat), but due to disagreement on the precise coded values for the behavior when it occurred (i.e., the target(s) of the combat).

[v] The Lin's concordance coefficient is designed to emulate the Pearson correlation coefficient, but with the correlation line forced to extend through the origin, and having a slope of 1.

[vi] Some efforts to establish confidence intervals and/or tests of statistical significance for reliability statistics have been reported. For example, Shrout and Fleiss (1979) have presented confidence intervals for six different forms of the ICC, and Hayes and Krippendorff (2006) provide a demonstration of the construction of a confidence interval via bootstrapping for one version of Krippendorff's alpha.