Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 7, Number 17, June, 2001 ISSN=1531-7714

An Overview of Content Analysis

Steve Stemler
Yale University

Content analysis has been defined as a systematic, replicable technique for compressing many words of text into fewer content categories based on explicit rules of coding (Berelson, 1952; GAO, 1996; Krippendorff, 1980; and Weber, 1990). Holsti (1969) offers a broad definition of content analysis as, "any technique for making inferences by objectively and systematically identifying specified characteristics of messages" (p. 14). Under Holsti's definition, the technique of content analysis is not restricted to the domain of textual analysis, but may be applied to other areas such as coding student drawings (Wheelock, Haney, & Bebell, 2000), or coding of actions observed in videotaped studies (Stigler, Gonzales, Kawanaka, Knoll, & Serrano, 1999). In order to allow for replication, however, the technique can only be applied to data that are durable in nature.

Content analysis enables researchers to sift through large volumes of data with relative ease in a systematic fashion (GAO, 1996). It can be a useful technique for allowing us to discover and describe the focus of individual, group, institutional, or social attention (Weber, 1990). It also allows inferences to be made which can then be corroborated using other methods of data collection. Krippendorff (1980) notes that "[m]uch content analysis research is motivated by the search for techniques to infer from symbolic data what would be either too costly, no longer possible, or too obtrusive by the use of other techniques" (p. 51).

Practical Applications of Content Analysis

Content analysis can be a powerful tool for determining authorship. For instance, one technique for determining authorship is to compile a list of suspected authors, examine their prior writings, and correlate the frequency of nouns or function words to help build a case for the probability of each person's authorship of the data of interest. Mosteller and Wallace (1964) used Bayesian techniques based on word frequency to show that Madison was indeed the author of the Federalist papers; recently, Foster (1996) used a more holistic approach in order to determine the identity of the anonymous author of the 1992 book *Primary Colors*.

Content analysis is also useful for examining trends and patterns in documents. For example, Stemler and Bebell (1998) conducted a content analysis of school mission statements to make some inferences about what schools hold as their primary reasons for existence. One of the major research questions was whether the criteria being used to measure program effectiveness (e.g., academic test scores) were aligned with the overall program objectives or reason for existence.

Additionally, content analysis provides an empirical basis for monitoring shifts in public opinion. Data collected from the mission statements project in the late 1990s can be objectively compared to data collected at some point in the future to determine if policy changes related to standards-based reform have manifested themselves in school mission statements.

Conducting a Content Analysis

According to Krippendorff (1980), six questions must be addressed in every content analysis:

1) Which data are analyzed?
2) How are they defined?
3) What is the population from which they are drawn?
4) What is the context relative to which the data are analyzed?
5) What are the boundaries of the analysis?
6) What is the target of the inferences?

At least three problems can occur when documents are being assembled for content analysis. First, when a substantial number of documents from the population are missing, the content analysis must be abandoned. Second, inappropriate records (e.g., ones that do not match the definition of the document required for analysis) should be discarded, but a record should be kept of the reasons. Finally, some documents might match the requirements for analysis but just be uncodable because they contain missing passages or ambiguous content (GAO, 1996).

Analyzing the Data

Perhaps the most common notion in qualitative research is that a content analysis simply means doing a word-frequency count. The assumption made is that the words that are mentioned most often are the words that reflect the greatest concerns. While this may be true in some cases, there are several counterpoints to consider when using simple word frequency counts to make inferences about matters of importance.

One thing to consider is that synonyms may be used for stylistic reasons throughout a document and thus may lead the researchers to underestimate the importance of a concept (Weber, 1990). Also bear in mind that each word may not represent a category equally well. Unfortunately, there are no well-developed weighting procedures, so for now, using word counts requires the researcher to be aware of this limitation. Furthermore, Weber reminds us that, "not all issues are equally difficult to raise. In contemporary America it may well be easier for political parties to address economic issues such as trade and deficits than the history and current plight of Native American living precariously on reservations" (1990, p. 73). Finally, in performing word frequency counts, one should bear in mind that some words may have multiple meanings. For instance the word "state" could mean a political body, a situation, or a verb meaning "to speak."

A good rule of thumb to follow in the analysis is to use word frequency counts to identify words of potential interest, and then to use a Key Word In Context (KWIC) search to test for the consistency of usage of words. Most qualitative research software (e.g., NUD*IST, HyperRESEARCH) allows the researcher to pull up the sentence in which that word was used so that he or she can see the word in some context. This procedure will help to strengthen the validity of the inferences that are being made from the data. Certain software packages (e.g., the revised General Inquirer) are able to incorporate artificial intelligence systems that can differentiate between the same word used with two different meanings based on context (Rosenberg, Schnurr, & Oxman, 1990). There are a number of different software packages available that will help to facilitate content analyses (see further information at the end of this paper).

Content analysis extends far beyond simple word counts, however. What makes the technique particularly rich and meaningful is its reliance on coding and categorizing of the data. The basics of categorizing can be summed up in these quotes: "A category is a group of words with similar meaning or connotations" (Weber, 1990, p. 37). "Categories must be mutually exclusive and exhaustive" (GAO, 1996, p. 20). Mutually exclusive categories exist when no unit falls between two data points, and each unit is represented by only one data point. The requirement of exhaustive categories is met when the data language represents all recording units without exception.

Emergent vs. *a priori* coding. There are two approaches to coding data that operate with slightly different rules. With *emergent coding*, categories are established following some preliminary examination of the data. The steps to follow are outlined in Haney, Russell, Gulek, & Fierros (1998) and will be summarized here. First, two people independently review the material and come up with a set of features that form a checklist. Second, the researchers compare notes and reconcile any differences that show up on their initial checklists. Third, the researchers use a consolidated checklist to independently apply coding. Fourth, the researchers check the reliability of the coding (a 95% agreement is suggested; .8 for Cohen's kappa). If the level of reliability is not acceptable, then the researchers repeat the previous steps. Once the reliability has been established, the coding is applied on a large-scale basis. The final stage is a periodic quality control check.

When dealing with *a priori* coding, the categories are established prior to the analysis based upon some theory. Professional colleagues agree on the categories, and the coding is applied to the data. Revisions are made as necessary, and the categories are tightened up to the point that maximizes mutual exclusivity and exhaustiveness (Weber, 1990).

Coding units. There are several different ways of defining coding units. The first way is to define them physically in terms of their natural or intuitive borders. For instance, newspaper articles, letters, or poems all have natural boundaries. The second way to define the recording units syntactically, that is, to use the separations created by the author, such as words, sentences, or paragraphs. A third way to define them is to use referential units. Referential units refer to the way a unit is represented. For example a paper might refer to George W. Bush as "President Bush," "the 43rd president of the United States," or "W." Referential units are useful when we are interested in making inferences about attitudes, values, or preferences. A fourth method of defining coding units is by using propositional units. Propositional units are perhaps the most complex method of defining coding units because they work by breaking down the text in order to examine underlying assumptions. For example, in a sentence that would read, "Investors took another hit as the stock market continued its descent," we would break it down to: The stock market has been performing poorly recently/Investors have been losing money (Krippendorff, 1980).

Typically, three kinds of units are employed in content analysis: sampling units, context units, and recording units.

- *Sampling units* will vary depending on how the researcher makes meaning; they could be words, sentences, or paragraphs. In the mission statements project, the sampling unit was the mission statement.

- *Context units* neither need be independent or separately describable. They may overlap and contain many recording units. Context units do, however, set physical limits on what kind of data you are trying to record. In the mission statements project, the context units are sentences. This was an arbitrary decision, and the context unit

could just as easily have been paragraphs or entire statements of purpose.

- *Recording units*, by contrast, are rarely defined in terms of physical boundaries. In the mission statements project, the recording unit was the idea(s) regarding the purpose of school found in the mission statements (e.g., develop responsible citizens or promote student self-worth). Thus a sentence that reads "The mission of Jason Lee school is to enhance students' social skills, develop responsible citizens, and foster emotional growth" could be coded in three separate recording units, with each idea belonging to only one category (Krippendorff, 1980).

Reliability. Weber (1990) notes: "To make valid inferences from the text, it is important that the classification procedure be reliable in the sense of being consistent: Different people should code the same text in the same way" (p. 12). As Weber further notes, "reliability problems usually grow out of the ambiguity of word meanings, category definitions, or other coding rules" (p. 15). Yet, it is important to recognize that the people who have developed the coding scheme have often been working so closely on the project that they have established shared and hidden meanings of the coding. The obvious result is that the reliability coefficient they report is artificially inflated (Krippendorff, 1980). In order to avoid this, one of the most critical steps in content analysis involves developing a set of explicit recording instructions. These instructions then allow outside coders to be trained until reliability requirements are met.

Reliability may be discussed in the following terms:

- *Stability*, or intra-rater reliability. Can the same coder get the same results try after try?
- *Reproducibility*, or inter-rater reliability. Do coding schemes lead to the same text being coded in the same category by different people?

One way to measure reliability is to measure the percent of agreement between raters. This involves simply adding up the number of cases that were coded the same way by the two raters and dividing by the total number of cases. The problem with a percent agreement approach, however, is that it does not account for the fact that raters are expected to agree with each other a certain percentage of the time simply based on chance (Cohen, 1960). In order to combat this shortfall, reliability may be calculated by using Cohen's Kappa, which approaches 1 as coding is perfectly reliable and goes to 0 when there is no agreement other than what would be expected by chance (Haney et al., 1998). Kappa is computed as:

$$\kappa = \frac{P_A - P_c}{1 - P_c}$$

where:

P_A = proportion of units on which the raters agree

P_c = the proportion of units for which agreement is expected by chance.

Table 1 – Example Agreement Matrix

		Rater 1			Marginal Totals
		Academic	Emotional	Physical	
Rater 2	Academic	.42 (.29)*	.10 (.21)	.05 (.07)	**.57**
	Emotional	.07 (.18)	.25 (.13)	.03 (.05)	**.35**
	Physical	.01 (.04)	.02 (.03)	.05 (.01)	**.08**
		.50	**.37**	**.13**	**1.00**

*Values in parentheses represent the expected proportions on the basis of chance associations, i.e. the joint probabilities of the marginal proportions.

Given the data in Table 1, a percent agreement calculation can be derived by summing the values found in the diagonals (i.e., the proportion of times that the two raters agreed):

$$P_A = .42 + .25 + .05 = .72$$

By multiplying the marginal values, we can arrive at an expected proportion for each cell (reported in parentheses in the table). Summing the product of the marginal values in the diagonal we find that on the basis of chance alone, we expect an observed agreement value of:

$$P_C = .29 + .13 + .01 = .43$$

Kappa provides an adjustment for this chance agreement factor. Thus, for the data in Table 1, kappa would be calculated as:

$$kappa = \frac{.72 - .43}{1 - .43} = .51$$

In practice, this value may be interpreted as the proportion of agreement between raters after accounting for chance (Cohen, 1960). Crocker & Algina (1986) point out that a value of $\kappa = 0$ does not mean that the coding decisions are so inconsistent as to be worthless, rather, $\kappa = 0$ may be interpreted to mean that the decisions are no more consistent than we would expect based on chance, and a negative value of kappa reveals that the observed agreement is worse than expected on the basis of chance alone. "In his methodological note on kappa in *Psychological Reports*, Kvalseth (1989) suggests that a kappa coefficient of 0.61 represents reasonably good overall agreement." (Wheelock et al., 2000). In addition, Landis & Koch (1977, p.165) have suggested the following benchmarks for interpreting kappa:

Kappa Statistic	Strength of Agreement
<0.00	Poor
0.00– 0.20	Slight
0.21– 0.40	Fair
0.41– 0.60	Moderate
0.61– 0.80	Substantial
0.81– 1.00	Almost Perfect

Cohen (1960) notes that there are three assumptions to attend to in using this measure. First, the units of analysis must be independent. For example, each mission statement that was coded was independent of all others. This assumption would be violated if in attempting to look at school mission statements, the same district level mission statement was coded for two different schools within the same district in the sample.

Second, the categories of the nominal scale must be independent, mutually exclusive, and exhaustive. Suppose the goal of an analysis was to code the kinds of courses offered at a particular school. Now suppose that a coding scheme was devised that had five classification groups: mathematics, science, literature, biology, and calculus. The categories on the scale would no longer be independent or mutually exclusive because whenever a biology course is encountered it also would be coded as a science course. Similarly, a calculus would always be coded into two categories as well, calculus and mathematics. Finally, the five categories listed are not mutually exhaustive of all of the different types of courses that

are likely to be offered at a school. For example, a foreign language course could not be adequately described by any of the five categories.

The third assumption when using kappa is that the raters are operating independently. In other words, two raters should not be working together to come to a consensus about what rating they will give.

Validity. It is important to recognize that a methodology is always employed in the service of a research question. As such, validation of the inferences made on the basis of data from one analytic approach demands the use of multiple sources of information. If at all possible, the researcher should try to have some sort of validation study built into the design. In qualitative research, validation takes the form of triangulation. Triangulation lends credibility to the findings by incorporating multiple sources of data, methods, investigators, or theories (Erlandson, Harris, Skipper, & Allen, 1993).

For example, in the mission statements project, the research question was aimed at discovering the purpose of school from the perspective of the institution. In order to cross-validate the findings from a content analysis, schoolmasters and those making hiring decisions could be interviewed about the emphasis placed upon the school's mission statement when hiring prospective teachers to get a sense of the extent to which a school's values are truly reflected by mission statements. Another way to validate the inferences would be to survey students and teachers regarding the mission statement to see the level of awareness of the aims of the school. A third option would be to take a look at the degree to which the ideals mentioned in the mission statement are being implemented in the classrooms.

Shapiro & Markoff (1997) assert that content analysis itself is only valid and meaningful to the extent that the results are related to other measures. From this perspective, an exploration of the relationship between average student achievement on cognitive measures and the emphasis on cognitive outcomes stated across school mission statements would enhance the validity of the findings. For further discussions related to the validity of content analysis see Roberts (1997), Erlandson et al. (1993), and Denzin & Lincoln (1994).

Conclusion

When used properly, content analysis is a powerful data reduction technique. Its major benefit comes from the fact that it is a systematic, replicable technique for compressing many words of text into fewer content categories based on explicit rules of coding. It has the attractive features of being unobtrusive, and being useful in dealing with large volumes of data. The technique of content analysis extends far beyond simple word frequency counts. Many limitations of word counts have been discussed and methods of extending content analysis to enhance the utility of the analysis have been addressed. Two fatal flaws that destroy the utility of a content analysis are faulty definitions of categories and non-mutually exclusive and exhaustive categories.

Further information

For links, articles, software and resources see

http://writing.colostate.edu/references/research/content/

http://www.gsu.edu/~wwwcom/.

References

Berelson, B. (1952). *Content Analysis in Communication Research*. Glencoe, Ill: Free Press.

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

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Orlando, FL: Harcourt Brace Jovanovich.

Denzin, N.K., & Lincoln, Y.S. (Eds.). (1994). *Handbook of Qualitative Research*. Thousand Oaks, CA: Sage Publications.

Erlandson, D.A., Harris, E.L., Skipper, B.L., & Allen, S.D. (1993). *Doing Naturalistic Inquiry: A Guide to Methods*. Newbury Park, CA: Sage Publications.

Foster, D. (1996, February 26). Primary culprit. *New York*, 50– 57.

Haney, W., Russell, M., Gulek, C., and Fierros, E. (Jan-Feb, 1998). Drawing on education: Using student drawings to promote middle school improvement. *Schools in the Middle, 7(3)*, 38– 43.

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

Krippendorff, K. (1980). *Content Analysis: An Introduction to Its Methodology*. Newbury Park, CA: Sage.

Kvalseth, T. O. (1989). Note on Cohen's kappa. *Psychological reports*, 65, 223– 26.

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, pp. 159–174.

Mosteller, F. and D.L. Wallace (1964). *Inference and Disputed Authorship: The Federalist*. Reading, Massachusetts: Addison-Wesley.

Nitko, A.J. (1983). *Educational Tests and Measurement: An Introduction*. New York, NY: Harcourt Brace Jovanovich.

Roberts, C.W. (Ed.) (1997). *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Mahwah, NJ: Lawrence Erlbaum Associates.

Rosenberg, S.D., Schnurr, P.P., & Oxman, T.E. (1990). Content analysis: A comparison of manual and computerized systems. *Journal of Personality Assessment, 54* (1 & 2), 298– 310.

Shapiro, G., & Markoff, J. (1997). 'A Matter of Definition' in C.W. Roberts (Ed.). *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Mahwah, NJ: Lawrence Erlbaum Associates.

Stemler, S., and Bebell, D. (1998). *An Empirical Approach to Understanding and Analyzing the Mission Statements of Selected Educational Institutions*. Paper presented at the annual meeting of the New England Educational Research Organization. Portsmouth, New Hampshire. Available: ERIC Doc No. ED 442 202.

Stigler, J.W., Gonzales, P., Kawanaka, T., Knoll, S. & Serrano, A. (1999). *The TIMSS Videotape Classroom Study: Methods and Findings from an Exploratory Research Project on Eighth-Grade Mathematics Instruction in Germany, Japan, and the United States*. U.S. Department of Education National Center for Educational Statistics: NCES 99-074. Washington, D.C.: Government Printing Office.

U.S. General Accounting Office (1996). *Content Analysis: A Methodology for Structuring and Analyzing Written Material*. GAO/PEMD-10.3.1. Washington, D.C. (This book can be ordered free from the GAO).

Weber, R. P. (1990). *Basic Content Analysis*, 2nd ed. Newbury Park, CA.

Wheelock, A., Haney, W., & Bebell, D. (2000). What can student drawings tell us about high-stakes testing in Massachusetts? *TCRecord.org*. Available: http://www.tcrecord.org/Content.asp?ContentID=10634.

Please address all correspondence regarding this article to:

Steve Stemler, Ph.D.
Wesleyan University
207 High Street
Middletown, CT 06459

E-Mail: sstemler [at] wesleyan.edu

Updated 3/7/2012

Descriptors: Content analysis; NUD*IST; Research Methods; Qualitative Analysis

Citation: Stemler, Steve (2001). An overview of content analysis. *Practical Assessment, Research & Evaluation*, 7(17). Available online: http://PAREonline.net/getvn.asp?v=7&n=17.