CONTENT ANALYSIS is a research method that uses a set of procedures to make valid inferences from text. ... Compared with other data-generating and analysis techniques, content analysis has several advantages:

- Communication is a central aspect of social interaction. Content-analytic procedures operate directly on text or transcripts of human communications.
- The best content-analytic studies use both qualitative and quantitative operations on texts. Thus content analysis methods combine what are usually thought to be antithetical modes of analysis.
- Documents of various kinds exist over long periods of time. Culture indicators generated from such series of documents constitute reliable data that may span even centuries.
- In more recent times, when reliable data of other kinds exist, culture indicators can be used to assess quantitatively the relationships among economic, social, political, and cultural change.
- Compared with techniques such as interviews, content analysis usually yields unobtrusive measures in which neither the sender nor the receiver of the message is aware that it is being analyzed. Hence, there is little danger that the act of measurement itself will act as a force for change that confounds the data.

Two very different studies show some ways content analysis has been used.

Walker (1975) analyzed differences and similarities in American black and
white popular song lyrics, 1962–1973. Using computer-aided content analysis, Walker investigated differences in narrative form. He found that compared with popular white song lyrics, 'rhythm and blues' and 'soul' song lyrics showed greater emphasis on action in the objective world, less concern with time, and greater emphasis on what Walker calls 'toughmindedness' or 'existential concreteness'.

In another study, Aries (1973) . . . studied differences in female, male, and mixed-sex small groups. She found that differential sex-role socialization and sex-role stereotyping affect thematic content and social interaction. In female groups, women show much concern with interpersonal issues. Women discuss 'themselves, their homes and families, and their relationships, defining themselves by the way they relate to the significant others who surround them' (Aries 1973: 254).

In male groups, members do not address interpersonal matters directly. Instead, men indirectly relate personal experiences and feelings through stories and metaphors. Men 'achieve a closeness through the sharing of laughter and stories of activities, rather than the sharing of the understanding of those experiences' (Aries 1973: 254). Also, all-male groups manifest more themes involving aggression than do all-female groups.

In mixed groups, Aries found that women talked less of their homes and families. Women also spoke less of achievement and institutions. In short, women in these groups 'orient themselves around being women with men by assuming the traditional female role' (Aries 1973: 256). Men in mixed groups expressed their competitiveness less through storytelling than through assuming leadership roles in the group. Moreover, in the presence of women, men shift more toward reflection of themselves and their feelings.

A central idea in content analysis is that the many words of the text are classified into much fewer content categories. Each category may consist of one, several, or many words. Words, phrases, or other units of text classified in the same category are presumed to have similar meanings. Depending on the purposes of the investigator, this similarity may be based on the precise meaning of the words (such as grouping synonyms together), or may be based on words sharing similar connotations (such as grouping together several words implying a concern with a concept such as WEALTH or POWER). 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. Also, the classification procedure must generate variables that are valid. A variable is valid to the extent that it measures or represents what the investigator intends it to measure.

Content classification and interpretation

. . . In content analysis, reliability problems usually grow out of the ambiguity of word meanings, category definitions, or other coding rules. Classification by multiple human coders permits the quantitative assessment of achieved reliability. Classification by computer, however, leads to perfect coder reliability. . . . Once correctly defined for the computer, the coding rules are always applied in the same way.
A much more difficult set of problems concerns the validity of variables based on content classification. A content analysis variable is valid to the extent that it measures the construct the investigator intends it to measure. As happens with reliability, validity problems also grow out of the ambiguity of word meanings and category or variable definitions.

As an introduction to these problems, consider two sample texts and some simple coding rules. Using commonsense definitions, imagine that the coding instructions define five categories: CITIZENS’ RIGHTS, ECONOMIC, GOVERNMENT, POLITICAL DOCTRINE, and WELFARE. Imagine also that coders are instructed to classify each entire paragraph in one category only. Consider first a portion of the Carter 1980 Democratic Platform:

Our current economic situation is unique. In 1977, we inherited a severe recession from the Republicans. The Democratic Administration and the Democratic Congress acted quickly to reduce the unacceptably high levels of unemployment and to stimulate the economy. And we succeeded. We recovered from that deep recession and our economy was strengthened and revitalized. As that fight was won, the enormous increases in foreign oil prices – 120 percent last year – and declining productivity fueled an inflationary spiral that also had to be fought. The Democrats did that, and inflation has begun to recede. In working to combat these dual problems, significant economic actions have been taken.

(Johnson 1982: 38)

Now consider another paragraph from the Reagan 1980 Republican platform:

Through long association with government programs, the word ‘welfare’ has come to be perceived almost exclusively as tax-supported aid to the needy. But in its most inclusive sense – and as Americans understood it from the beginning of the Republic – such aid also encompasses those charitable works performed by private citizens, families, and social, ethnic, and religious organizations. Policies of the federal government leading to high taxes, rising inflation, and bureaucratic empire-building have made it difficult and often impossible for such individuals and groups to exercise their charitable instincts. We believe that government policies that fight inflation, reduce tax rates, and end bureaucratic excesses can help make private effort by the American people once again a major force in those works of charity which are the true signs of a progressive and humane society.

(Johnson 1982: 179)

Most people would code the first excerpt in the economic category, but the proper coding of the second is less obvious. This paragraph could be taken to be mainly about the rights of citizens, the desirability of restricting the government’s role, the welfare state, or to be the espousal of a political doctrine. In fact, it occurs at the end of a section titled Improving the Welfare System.

The difficulty of classifying the second excerpt is contrived partly by the present
author, because it results from the lack of clear and detailed coding rules for each category and from the variety of the subject matter. Large portions of text, such as paragraphs and complete texts, usually are more difficult to code as a unit than smaller portions, such as words and phrases, because large units typically contain more information and a greater diversity of topics. Hence they are more likely to present coders with conflicting cues.

These examples show the kind of difficulties investigators face with coding text. The next two sections look more systematically at coding problems, first from the perspective of reliability assessment and then from the perspective of validity assessment.

**Reliability**

Three types of reliability are pertinent to content analysis: stability, reproducibility, and accuracy (Krippendorff 1980: 130–154). Stability refers to the extent to which the results of content classification are invariant over time. Stability can be determined when the same content is coded more than once by the same coder. Inconsistencies in coding constitute unreliability. These inconsistencies may stem from a variety of factors, including ambiguities in the coding rules, ambiguities in the text, cognitive changes within the coder, or simple errors, such as recording the wrong numeric code for a category. Because only one person is coding, stability is the weakest form of reliability.

Reproducibility, sometimes called intercoder reliability, refers to the extent to which content classification produces the same results when the same text is coded by more than one coder. Conflicting codings usually result from cognitive differences among the coders, ambiguous coding instructions, or from random recording errors. High reproducibility is a minimum standard for content analysis. This is because stability measures the consistency of the individual coder’s private understandings, whereas reproducibility measures the consistency of shared understandings (or meaning) held by two or more coders.

Accuracy refers to the extent to which the classification of text corresponds to a standard or norm. It is the strongest form of reliability. As Krippendorff notes (1980: 131), it has sometimes been used to test the performance of human coders where a standard coding for some text has already been established. Except for training purposes, standard codings are established infrequently for texts. Consequently, researchers seldom use accuracy in reliability assessment.

Krippendorff (1980: 132) also points out that many investigators fail totally to assess the reliability of their coding. Even when reliability is assessed, some investigators engage in practices that often make data seem more reliable than they actually are. In particular, where coders have disagreed, investigators have resolved these disagreements by negotiations or by invoking the authority of the principal investigator or senior graduate assistant. Resolving these disagreements may produce judgments biased toward the opinions of the most verbal or more senior of the coders. Consequently, the reliability of the coding should be calculated before these disagreements are resolved. . . .
Validity

Perhaps the weakest form of validity is face validity, which consists of the correspondence between investigators' definitions of concepts and their definitions of the categories that measured them. A category has face validity to the extent that it appears to measure the construct it is intended to measure. Even if several expert judges agree, face validity is still a weak claim because it rests on a single variable. Stronger forms of validity involve more than one variable. Unfortunately, content analysts often have relied heavily on face validity; consequently, some other social scientists have viewed their results skeptically.

Much stronger validity is obtained by comparing content-analytic data with some external criterion. Four types of external validity are pertinent. [For example,] a measure has predictive validity to the extent that forecasts about events or conditions external to the study are shown to correspond to actual events or conditions. These predictions may concern future, past (postdict), or concurrent events. Predictive validity is powerful because the inferences from data are generalized successfully beyond the study to situations not under the direct control of the investigator.

Semantic validity exists when persons familiar with the language and texts examine lists of words (or other units) placed in the same category and agree that these words have similar meanings or connotations. Although this seems an obvious requirement for valid content analysis, many difficulties arise because words and category definitions are sometimes ambiguous. For example, some systems for computer-aided content analysis cannot distinguish among the various senses of words with more than one meaning, such as mine. Does this refer to a hole in the ground, the process of extraction, or a possessive pronoun? Because of this failure, word counts including the frequency of mine lack semantic validity.

Creating and testing a coding scheme

Many studies require investigators to design and implement coding schemes. Whether the coding is to be done by humans or by computer, the process of creating and applying a coding scheme consists of several basic steps. If investigators have identified the substantive questions to be investigated, relevant theories, previous research, and the texts to be classified, they next proceed with the following necessary steps:

1. **Define the recording units.** One of the most fundamental and important decisions concerns the definition of the basic unit of text to be classified. There are six commonly used options:

   - **Word** – One choice is to code each word. As noted, some computer software for text analysis cannot distinguish among the various senses of words with more than one meaning, and hence may produce erroneous conclusions.
   - **Word sense** – Other computer programs are able to code the different senses of words with multiple meanings and to code phrases that constitute a semantic unit, such as idioms (e.g., taken for granted) or proper nouns (e.g., the Empire State Building).
Sentence — An entire sentence is often the recording unit when the investigator is interested in words or phrases that occur closely together. For example, coders may be instructed to count sentences in which either positive, negative, or affectively neutral references are made to the Soviet Union. A sentence with the phrase evil empire would be counted as NEGATIVE EVALUATION, whereas Talks with the Soviet Union continue would be coded NEUTRAL EVALUATION, and The President supports recent efforts to extend economic and political rights in the Soviet Union would be coded POSITIVE EVALUATION.

Theme — Holsti (1963: 136, emphasis in the original) defines a theme as a unit of text ‘having no more than one each of the following elements: (1) the perceiver, (2) the perceived or agent of action, (3) the action, (4) the target of the action.’ For example, the sentence The President hates Communists would be divided as shown. Numeric or other codes often are inserted in the text to represent subject/verb/object. This form of coding preserves important information and provides a means of distinguishing between the sentence above and the assertion that Communists hate the President.

Sometimes long, complex sentences must be broken down into shorter thematic units or segments. Here, parts of speech shared between themes must be repeated. Also, ambiguous phrases and pronouns must be identified manually. These steps are taken before coding for the content. Holsti (1963: 136–137) gives the following example of editing more complex sentences before coding for themes and content.

The sentence, ‘The American imperialists have perverted the peace and are preparing to attack the Socialist Camp,’ must be edited to read: The American imperialists have perverted the peace + (the Americans) are preparing to attack the Socialist Camp.

This form of coding is labor-intensive, but leads to much more detailed and sophisticated comparisons. . . .

Paragraph — When computer assistance is not feasible and when resources for human coding are limited, investigators sometimes code entire paragraphs to reduce the effort required. Evidence discussed later in this chapter shows that it is more difficult to achieve high reliability when coding large units, such as paragraphs, than when coding smaller units, such as words.

Whole text — Unless the entire text is short — like newspaper headlines, editorials, or stories — it is difficult to achieve high reliability when coding complete texts.

2. Define the categories. In creating category definitions, investigators must make two basic decisions. The first is whether the categories are to be mutually exclusive. Most statistical procedures require variables that are not confounded. If a recording unit can be classified simultaneously in two or more categories and if both categories (variables) are included in the same statistical analysis, then it is possible that,
because the basic statistical assumptions of the analysis are violated, the results are
dubious. This is likely to be the case when using common multivariate procedures
such as factor analysis, analysis of variance, and multiple regression.

The second choice concerns how narrow or broad the categories are to be.
Some categories are limited because of language. For example, a category indicating
self-references defined as first person singular pronouns will have only a few words
or entries. A category defined as concern with ECONOMIC matters may have many
entries. For some purposes, however, it may make sense to use much more narrow
or specific categories, such as INFLATION, TAXES, BUDGET, TRADE, AGRICUL-
TURE, and so on.

3. Text coding on sample of text. The best test of the clarity of category definitions
is to code a small sample of the text. Testing not only reveals ambiguities in the
rules, but also often leads to insights suggesting revisions of the classification
scheme.

4. Assess accuracy or reliability. Accuracy in this sense means the text is coded
correctly by the computer, not in the sense of the type of reliability that was
discussed earlier. If human coders are used, the reliability of the coding process
should be estimated before resolving disputes among the coders.

5. Revise the coding rules. If the reliability is low, or if errors in computer
procedures are discovered, the coding rules must be revised or the software
corrected.

6. Return to Step 3. This cycle will continue until the coders achieve sufficient
reliability or until the computer procedures work correctly.

7. Code all the text. When high coder reliability has been achieved or when the
computer programs are functioning correctly, the coding rules can then be applied
to all the text.

8. Assess achieved reliability or accuracy. The reliability of human coders should be
assessed after the text is classified. Never assume that if samples of text were coded
reliably then the entire corpus of text will also be coded reliably. Human coders are
subject to fatigue and are likely to make more mistakes as the coding proceeds. Also,
as the text is coded, their understanding of the coding rules may change in subtle
ways that lead to greater unreliability.

If the coding was done by computer, the output should be checked carefully to
insure that the coding rules were applied correctly. Text not in the sample(s) used
for testing may present novel combinations of words that were not anticipated or
encountered earlier, and these may be misclassified.

References

Aries, E. (1973) 'Interaction patterns and themes of male, female, and mixed groups'.

Holsti, O.R. (1963) 'Computer content analysis,' in R.C. North, O.R. Holsti, M.G. Zani-
novich, and D.A. Zinner, Content Analysis: A Handbook with Applications for the Study of
International Crisis, Evanston: Northwestern University Press.
