where. The goal is to produce a more interesting and useful book for those contemplating or actually doing research using content analysis. During the past 10 years, the introduction of relatively inexpensive microcomputers, the introduction of cost-effective devices for making text machine-readable, and the reduction of computer costs generally, have renewed interest in content analysis. Certainly, these tools will be increasingly applied to a wide range of social science questions during the next few years.

As a brief introduction to content analysis, much is omitted. Consequently, there are suggestions for further reading at the end of each chapter. These books and series address content analysis methods, substantive research, general issues in research methodology, or statistics at a much greater level of detail than is possible or even desirable here.

Suggestions for Further Reading

There are several books on content analysis that should be read by anyone seriously interested in the subject. Krippendorff (1980) is a good recent survey of the field and its problems. It is especially useful for those doing human-coded content analysis. His discussion of reliability is "must" reading. However, the book is not up-to-date regarding the use of computers.

Other books contain numerous methodological insights and practical information. One is the book on the General Inquirer system (Stone et al., 1966), the first widely used computer system for content analysis. Although the version discussed here is not the current one (see Kelly and Stone, 1975), this book presents a wide-ranging discussion of content analysis, its problems, and practical solutions. Stone and his associates also present several chapters illustrating the application of computer-aided content analysis to a variety of substantive problems. Another useful resource is a set of conference papers edited by Gerbier et al. (1969). This interdisciplinary collection addresses many issues still current in content analysis. In addition, Holsti's (1969) brief discussion is worthwhile reading. North et al. (1963) apply a variety of content-analytic techniques to the study of communications in international relations. Finally, there is an earlier, precomputer body of work on (or using) content analysis—notably, Berelson (1952), Lasswell, Lutes et al. (1965), Lasswell et al. (1952), and Pooo (1953, 1959).

2. CONTENT CLASSIFICATION AND INTERPRETATION

The central problems of content analysis arise mainly in the data-reduction process by which the many words of texts are classified into much fewer content categories. One set of problems concerns the consistency or reliability of text classification. In context analysis, reliability problems usually grow out of the ambiguity of word meanings or the ambiguity of 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 (under the assumptions of correct computer programs and well-functioning computer hardware). Once properly 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 is the case 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 common sense 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 whole paragraph in one and only one of these categories. 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 Administrations 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, or about 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 partly contrived by the present author, as it results from the lack of clear and detailed coding rules for each category and from the variety of subject matter. Large portions of text, such as paragraphs and whole texts, are usually more difficult to code as a unit than smaller portions, such as words and phrases, because large units typically contain more information and greater diversity of topics. Hence they are more likely to present coders with conflicting cues.

These examples illustrate the kind of problems investigators face when coding text. The next two sections look more systematically at coding problems from the perspectives of reliability and validity assessment, respectively.

**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 ascertained 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 private understandings, whereas reproducibility measures the consistency of shared understandings, or meanings.

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 when a standard coding for some text has already been established. Except for training purposes, standard codings are infrequently established 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, when 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. Krippendorff goes on to illustrate several ways of calculating reliabilities for human coders.

Readers who plan to do human-coded content analysis should pay close attention to Krippendorff's discussion. Subsequent sections of this chapter return to reliability issues in conjunction with category construction and word classification.
Validity

The term validity is potentially confusing because it has been used in a variety of ways in the methodology literature (comprehensive Brinberg and Kidder, 1982; Cook and Campbell, 1979; Campbell and Stanley, 1963). However, two distinctions may help clarify the concept. The first is between validity as correspondence between two sets of things, such as concepts, variables, methods, and data, and validity as generalizability of results, inferences, and theory (Brinberg and McGrath, 1982). For example, if the same construct is measured by two different methods and the resulting variables are highly correlated, then these variables are valid indicators of the construct. These results would also indicate that the concept can be generalized in the sense that its measurement is not dependent upon a particular operation. Imagine, for example, two categories indicating concern with Hostility—one based on classifying words, the other based on classifying sentences. Imagine further that for the same set of texts, these different measures are highly correlated. This finding would indicate that both variables measure the same construct, and thus are both valid indicators.

Furthermore, if results from a number of studies agree, this correspondence suggests that the results are valid because they can be generalized from a variety of research situations. Correspondence and generalizability are essential aspects of the several types of validity discussed in this chapter.

A second distinction, more specific to content analysis, is between the validity of the classification scheme or variables derived from it, and the validity of the interpretation relating content variables to their causes or consequences. To assert that a category or variable (Economic, for example) is valid is to assert that there is correspondence between the category and the abstract concept that it represents (concern with economic matters). To assert that a research result based on content analysis is valid is to assert that the finding does not depend upon, or is generalizable beyond the specific data, methods, or measurements of a particular study. For instance, if a computer-assisted content analysis of party platforms indicates a strong relationship between long-term economic fluctuations and concern with the well-being of economy and society, then the validity of the results would increase to the extent that other data (e.g., newspaper editorials), other coding procedures (e.g., human rather than computer-coded), or other classification schemes (dictionaries) produced similar substantive conclusions.

Perhaps the weakest form of validity is face validity, which is based on the correspondence between the investigator’s definition of a concept and her or his definition of the category that measures it. A category has face validity to the extent that it appears to measure the construct it is intended to measure. Even if a number of expert judges agree, face validity is still a weak claim as it rests on a single variable. Stronger forms of validity involve more than one variable. Unfortunately, content analysts tend to rely heavily on face validity, consequently, other social scientists often view their results with some skepticism. Much stronger validity is obtained by comparing content-analytic data with some external criterion. Four types of external validity are pertinent.

A measure has construct validity to the extent that it is correlated with some other measure of the same construct. Campbell and Fiske (1959) and others (e.g., Alwin, 1974; Atshauer, 1974; Fiske, 1982; Campbell and O’Connell, 1982) further differentiate convergent from discriminant validity. A measure has high construct validity when it correlates with other measures of the same construct (convergent) and is uncorrelated with measures of dissimilar constructs (discriminant). Construct validity entails the generalizability of the construct across measures or methods.

The research reported in Saris-Galilhofer et al. (1978) is a fine example of applying these ideas to content-analytic data. The object of this study was to validate a content analysis dictionary developed by Holsti (1969) using the main categories of the Semantic Differential (Osgood et al., 1957; Snider and Osgood, 1969; Osgood et al., 1975; Anderson, 1970). The Semantic Differential is a technique for assessing the primary categories people use in affective evaluation or classification. The details of the technique are not pertinent here. However, research in a variety of cultures indicates that people use three basic dimensions of classification. Each dimension is anchored by polar opposites:

- Evaluation: positive versus negative affect;
- Potency: strength versus weakness; and,
- Activity: active versus passive.

Each word in Holsti’s (1969) dictionary was assigned a number representing the extent to which it indicates each dimension of the
semantic differential. Saris-Gilhofer and her colleagues (1978) compared Holsti's assignment of scores with Osgood's and with scores assigned by a group of students. Thus, each word (or other unit of text) was classified by three different methods. Each method claims to classify text on the same constructs. Using statistical techniques designed to assess convergent and discriminant validity, Saris-Gilhofer found that Holsti's scoring for the Evaluation and Potency dimensions was much more valid than his scoring for the Activity dimension. It remains unclear why Holsti's scoring of the Activity dimension is less valid than the scoring for the other two. Additional research is required to determine the specific factors that affect the validity of content classification. Nonetheless, this study shows that sophisticated statistical techniques useful in assessing validity can be applied to content analysis data.

Hypothesis validity, the second type of validity, relies on the correspondence among variables and the correspondence between these relationships and theory. A measure has hypothesis validity if in relationship to other variables it "behaves" as it is expected to. For example, several studies based on political documents, such as party platforms in presidential campaigns, have shown that the preoccupation of society with economic issues increases during bad economic times and decreases when the economy is good (e.g., Namenwirth, 1969, 1973; Weber, 1982, 1984a). These results are consistent with theoretical arguments relating the cultural and social processes that generate political documents (such as party platforms) with changes in the economy. Thus, the observed inverse relationship between economic fluctuations and concern with economic matters indicates the hypothesis validity of measured variables and the constructs they represent.

A measure has predictive validity, the third type, 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 either future, past (postdiction), or concurrent events. Predictive validity is powerful because the inferences from data are successfully generalized beyond the study to situations not under the direct control of the investigator.

Content-analytic data are seldom shown to have predictive validity, but three examples illustrate the point.

- Ogilvie et al. (1966) analyzed real suicide notes from 33 males who had been matched for age, gender, occupation, religion,

and ethnicity with 33 nonsuicidal controls who were asked to produce simulated suicide notes. Using General Inquirer type computer-aided content analysis, Stone was able to correctly distinguish real from simulated suicide notes in 30 of the 33 pairs (90.9%) of notes.

- George (1959a) studied inferences made by The Foreign Broadcast Intelligence Service of the F.C.C. from German propaganda during World War II. He found that Allied intelligence analysts often could anticipate changes in German war tactics and strategy from changes in the content of radio broadcasts and other media.

- Namenwirth's (1973) analysis of party platforms in presidential campaigns, written in the late 1960s, suggested that America would experience severe economic difficulties that would peak about 1980. Events since seem to confirm this prediction.

Words or other coding units classified together need to possess similar connotations in order for the classification to have semantic validity, the fourth and final type of validity. According to Krippendorff (1980: 1990), 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 to be an obvious requirement for valid content analysis, many difficulties arise because of the ambiguities of words and of category definitions. For example, the early systems for computer-aided content analysis could not distinguish between 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" lacked semantic validity. Various aspects of semantic validity are discussed later in this and in subsequent chapters.

Creating and Testing a Coding Scheme

Many studies require the investigator to design and implement a coding scheme. Whether the coding is to be done by humans or computer, the process of creating and applying a coding scheme includes several basic steps. Assuming that the investigator has identi-
found the substantive question to be investigated, relevant theories, previous research, and the text to be classified, he or she next proceeds 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 common options:

- **Word**—One choice is to code each word. As noted, early computer software for text analysis did not have the ability to distinguish among the various senses of words with more than one meaning, and hence produced ambiguous results.

- **Word Sense**—More recent computer programs have the ability 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"). These issues are discussed in detail later in this chapter.

- **Sentence**—An entire sentence is often the recording unit when the investigator is interested in words or phrases that occur 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. "Talks with the Soviet Union continue," would be coded Neutral 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 are often 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 theme format. Parts of speech shared between themes must be repeated. In addition, ambiguous phrases and pronouns must be identified manually. These steps are taken prior to 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. See Holsti (1963, 1966, 1969) for further details.

- **Paragraph**—When computer assistance is not feasible and when for resources for human codings are limited, investigators sometimes code whole paragraphs in order to reduce the effort required. Evidence discussed later in this chapter indicates 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 whole text is relatively short, like newspaper headlines, editorials, or stories, it is difficult to achieve high reliability when coding whole texts.

(2) Define the categories. In creating category definitions, investigators must make two basic decisions. (Other related issues are taken up later.) The first is whether or not 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 the basic statistical assumptions of the analysis are violated and the results dubious.

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, Agriculture, and so on.

(3) Test 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 reveal ambiguities in the rules, but it often leads to insights suggesting revisions of the classification scheme.

(4) Assess accuracy or reliability. Accuracy in this sense means the text is correctly coded by the computer, not in the sense of a 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 coding rules. If the reliability is low, or if errors in computer procedures are discovered, the coding rules must be revised.

(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 of 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 reliably coded, then the whole corpus of text will also be reliably coded. Human coders are subject to fatigue and are likely to make more mistakes as the coding continues. Their understanding of the coding rules may change in subtle ways as the text is coded, and this leads to unreliability.

If the coding was done by computer, the output should be carefully checked to ensure that the coding rules were correctly applied. 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 cause errors in classification.

Dictionaries and Computerized Text Classification

Content analysts have used several strategies to create categories and variables. Some investigators have counted to have a few key words or phrases. Tufte (1978: 75), for example, counted certain words in the 1976 Democratic and Republican party platforms, including indicators of distributional issues, such as “inequity,” “progressive,” “equal,” and “redistribution,” and indication of concern with inflation, such as “inflation,” “inflationary,” “price stability,” and “rising prices.”

Others have constructed a set of content categories based on a single concept. For example, the early version of Stone’s General Inquirer computer system was used to analyze achievement imagery (McClelland’s N-Achievement; Stone et al., 1966: 191ff). This approach offers several advantages. It permits the intensive and detailed analysis of a single theoretical construct. It also provides an explicit rationale not only for what is retained, but also for what is excluded from the analysis.

Furthermore, single-concept coding schemes often have high validity and reliability.

Another approach to content analysis involves the creation and application of general dictionaries. Categorization analysis dictionaries consist of category names, the definitions of rules for assigning words to categories, and the actual assignment of specific words. This strategy provides the researcher with a wide range of categories (60 to 150+) into which most words in most texts can be classified. Once created, general dictionaries are advantageous because they:

- provide a wide range of categories to choose from (see, for example, Stone et al., 1966: 42–44);
- minimize the time needed for dictionary construction, validation, and revision;
- standardize classification; and
- encourage the accumulation of comparable results when used in many studies.

It is worth noting here that dictionary construction is commonly misperceived to be merely a preface or preparatory step for quantification. Although researchers commonly use dictionaries to define variables for quantification, they also employ categories to locate and retrieve text based on the occurrence of semantically equivalent symbols. Chapter 3 presents examples of retrievals based on categories.

Certain problems arise in the creation of any content category or set of categories. These problems stem from the ambiguity of both the category definitions and of the words that are to be assigned to categories. To facilitate discussion of these difficulties, two general dictionaries are used as examples, the Harvard IV Psychosocial Dictionary, developed by Dexter Dunphy and his associates (Dunphy et al., 1978; Kelly and Strom, 1975), and the Lasswell Value Dictionary