Answering these questions helps ensure that the researcher is familiar with the data he or she will analyze and can help identify any problems with it. It is unlikely that you or any secondary data analyst will be able to develop complete answers to all these questions prior to starting an analysis, but it still is critical to make the attempt to assess what you know and don't know about data quality before deciding whether to conduct the analysis. If you uncover bases for real concern after checking documents, the other publications with the data, information on Web sites, and perhaps by making some phone calls, you may have to decide to reject the analytic plan and instead search for another data set. If your initial answers to these six questions give sufficient evidence that the data can reasonably be used to answer your research question, you should still keep seeking to add in missing gaps in you initial answers to the six questions; through this ongoing process, you will develop the fullest possible understanding of the quality of your data. This understanding can lead you to steer your analysis in the most productive directions and can help you write a convincing description of the data set's advantages and limitations.

This seems like a lot to ask, doesn't it? After all, you can be married for life after answering only one question; here, I'm encouraging you to attempt to answer six questions before committing yourself to a brief relationship with a data set. Fortunately, the task is not normally so daunting. If you acquire a data set for analysis from a trusted source, many of these questions will already have been answered for you. You may need to do no more than read through a description of data available on a Web site to answer the secondary data questions and consider yourself prepared to use the data for your own purposes. If you are going to be conducting major analyses of a data set, you should take more time to read the complete study documents, review other publications with the data, and learn about the researchers who collected the data.

Exhibit 13.8 contains the description of a data set available from the ICPSR. Read through it and see how many of the secondary data questions it answers.

You will quickly learn that this data set represents one survey conducted as part of the ongoing Detroit Area Studies, so you'll understand the data set better if you also read a general description of that survey project (Exhibit 13.9).

In an environment in which so many important social science data sets are instantly available for reanalysis, the method of secondary data analysis should permit increasingly rapid refinement of social science knowledge, as new hypotheses can be tested and methodological disputes clarified if not resolved quickly. Both the necessary technology and the supportive ideologies required for this rapid refinement have spread throughout the world. Social science researchers now have the opportunity to take advantage of this methodology as well as the responsibility to carefully and publicly delineate and acknowledge the limitations of the method.

CONTENT ANALYSIS

How are medical doctors regarded in American culture? Do newspapers use the term *schizophrenia* in a way that reflects what this serious mental illness actually involves? Does the portrayal of men and women in video games reinforce gender stereotypes? Are the body images of male and female college students related to their experiences with romantic love? If you are concerned with understanding culture, attitudes toward mental illness, or gender roles,

EXHIBIT 13.8 ICPSR Data Set Description

Description—Study No. 4120

Bibliographic Description

ICPSR Study No.:	4120
Title:	Detroit Area Study, 1997: Social Change in Religion and Child Rearing
Principal Investigator(s):	Duane Alwin, University of Michigan
Series:	*Detroit Area Studies Series*
Bibliographic Citation:	Alwin, Duane. DETROIT AREA STUDY, 1997: SOCIAL CHANGE IN RELIGION AND CHILD REARING [Computer file]. ICPSR04120-v1. Ann Arbor, MI: Detroit Area Studies [producer], 1997. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2005-06-02.
Scope of Study *Summary:*	For this survey, respondents from three counties in the Detroit, Michigan, area were queried about their work, health, marriage and family, finances, political views, religion, and child rearing. With respect to finances, respondent views were elicited on credit card purchases, recording expenditures, and savings and investments. Regarding political views, respondents were. . . .
Subject Term(s):	*abortion, Atheism, Bible, birth control, Catholicism, Catholics, child rearing, children, Christianity, church attendance, communism, Creationism, credit card use, divorce, drinking behavior, economic behavior, educational background, employment, ethnicity, families, . . .*
Geographic Coverage:	Detroit, Michigan, United States
Time Period:	1997
Date(s) of Collection:	1997
Universe:	Residents 21 years and older in the tri-county area (Wayne, Oakland, and Macomb) of Michigan.
Data Type:	survey data
Methodology *Sample:*	A random-digit dialing sample of residential telephone numbers in the Michigan counties of Wayne, Oakland, and Macomb. The sample was restricted to adults 21 years of age and older.
Mode of Data Collection:	telephone interview
Extent of Processing:	CDBK.ICPSR/ DDEF.ICPSR/ REFORM.DATA
Access & Availability *Extent of Collection:*	1 data file + machine-readable documentation (PDF) + SAS setup file + SPSS setup file + Stata setup file
Data Format:	Logical Record Length with SAS, SPSS, and Stata setup files, SPSS portable file, and Stata system file
Original ICPSR Release:	2005-06-02

Note: Detailed file-level information (such as LRECL, case count, and variable count) may be found in the *file manifest.*

EXHIBIT 13.9 ICPSR Description of Detroit Area Studies

Detroit Area Studies Series

- View studies in the series
- Related Literature

The Detroit Area Studies series was initiated in 1951 at the University of Michigan and has been carried out nearly every year till the present. The Department of Sociology and the Survey Research Center of the Institute for Social Research are associated with the development of the series. It was initially supported by funds from the Ford Foundation, but since 1988 the University of Michigan has provided primary financial support for the series, with supplemental funding obtained frequently from outside sources. The purpose of these surveys is to provide practical social research training for graduate students and reliable data on the Greater Detroit community. Each survey probes a different aspect of personal and public life, economic and political behavior, political attitudes, professional and family life, and living experiences in the Detroit metropolitan area. The different specific problems investigated each year are selected by the executive committee of the project.

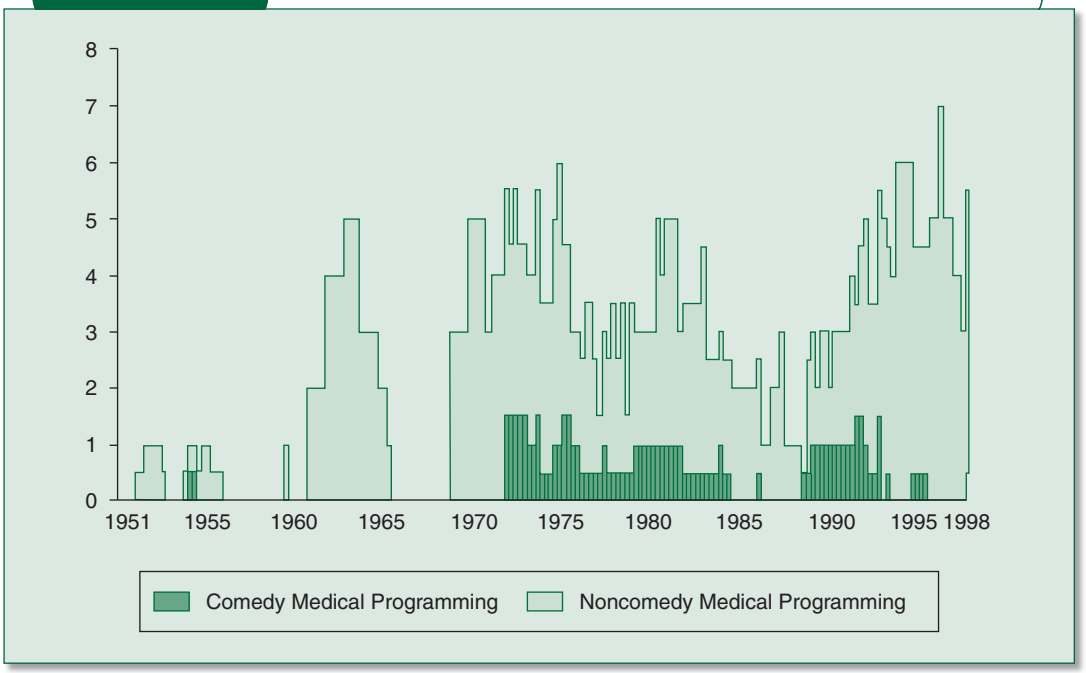
you'll probably find these to be important research questions. In this section, I introduce content analysis as an appropriate method for investigating such questions. Content analysis is "the systematic, objective, quantitative analysis of message characteristics" (Neuendorf 2002:1); you will soon see that we can learn a great deal about popular culture and many other issues through studying the characteristics of messages delivered through the mass media and other sources.

The goal of a content analysis is to develop inferences from text (Weber 1990:9). You can think of a content analysis as a "survey" of some documents or other records of prior communication—a survey with fixed-choice responses that produce quantitative data. This method was first applied to the study of newspaper and film content and then developed systematically for the analysis of Nazi propaganda broadcasts in World War II. Since then, content analysis has been used to study historical documents, records of speeches, and other "voices from the past" as well as media of all sorts (Neuendorf 2002:31–37). Content analysis techniques are also used to analyze in responses to open-ended survey questions.

Content analysis bears some similarities to qualitative data analysis, because it involves coding and categorizing text and identifying relationships among constructs identified in the text. However, since it usually is conceived as a quantitative procedure, content analysis overlaps with qualitative data analysis only at the margins—the points where qualitative analysis takes on quantitative features or where content analysis focuses on qualitative features of the text. This distinction becomes fuzzy, however, because content analysis techniques can be used with all forms of messages, including visual images, sounds, and interaction patterns, as well as written text (Neuendorf 2002:24–25).

Kimberly Neuendorf's (2002:3) analysis of medical prime-time network television programming introduces the potential of content analysis. As Exhibit 13.10 shows, medical programming has been dominated by noncomedy shows, but there have been two significant periods of comedy medical shows—during the 1970s and early 1980s and then again in the

EXHIBIT 13.10 Medical Prime-Time Network Television Programming, 1951 to 1998



early 1990s. It took a qualitative analysis of medical show content to reveal that the 1960s shows represented a very distinct "physician-as-God" era, which shifted to a more human view of the medical profession in the 1970s and 1980s. This era has been followed, in turn, by a mixed period that has had no dominant theme.

Content analysis is useful for investigating many questions about the social world. To illustrate its diverse range of applications, I will use in the next sections Neuendorf's (2002) analysis of TV programming, Kenneth Duckworth's (2003) (and my) analysis of newspaper articles, Karen Dill and Kathryn Thill's (2007) analysis of video game characters, and Suman Ambwani and Jaine Strauss's (2007) analysis of student responses to open-ended survey questions. These examples will demonstrate that the unites that are "surveyed" in a content analysis can range from newspapers, books, or TV shows to persons referred to in other communications, themes expressed in documents, or propositions made in different statements.

Content analysis proceeds through several stages.

Identify a Population of Documents or Other Textual Sources

This population should be selected so that it is appropriate to the research question of interest. Perhaps the population will be all newspapers published in the United States, college student newspapers, nomination speeches at political party conventions, or "state of the nation" speeches by national leaders. Books or films are also common sources for content analysis

projects. Often, a comprehensive archive can provide the primary data for the analysis (Neuendorf 2002:76–77). For a fee, the LexisNexis service makes a large archive of newspapers available for analysis. For her analysis of prime-time programming since 1951, Neuendorf (2002:3–4) used a published catalog of all TV shows. For my analysis with Duckworth and others (2003:1402) of the use by newspapers of the terms *schizophrenia* and *cancer*, I requested a sample of articles from the LexisNexis national newspaper archive. Dill and Thill (2007:855–856) turned to video game magazines for their analysis of the depiction of gender roles in video games. For their analysis of gender differences in body image and romantic love, Ambwani and Strauss (2007:15) surveyed students at a small Midwestern liberal arts college.

Determine the Units of Analysis

These could be items such as newspaper articles, whole newspapers, speeches, or political conventions, or they could be more microscopic units such as words, interactions, time periods, or other bits of a communication (Neuendorf 2002:71). The content analyst has to decide what units are most appropriate to her research question and how the communication content can be broken up into those units. If the units are individual issues of a newspaper, in a study of changes in news emphases, this step may be relatively easy. However, if the units are most appropriately the instances of interaction between characters in a novel or a movie, in a study of conflict patterns between different types of characters, it will require a careful process of testing to determine how to define operationally the specific units of interaction (Weber 1990:39–40).

Units of analysis varied across the four content analysis projects I have introduced. The units of analysis for Neuendorf (2002:2) were "the individual medically oriented TV program"; for Duckworth et al. (2003:1403), they were newspaper articles; for Dill and Thill (2007:856), they were images appearing in magazine articles; while for Ambwani and Strauss (2007:15), they were individual students.

Select a Sample of Units From the Population

The simplest strategy might be a simple random sample of documents. However, a stratified sample might be needed to ensure adequate representation of community newspapers in large and in small cities, or of weekday and Sunday papers, or of political speeches during election years and in off years (see Chapter 4) (Weber 1990:40–43). Nonrandom sampling methods have also been used in content analyses when the entire population of interest could not be determined (Neuendorf 2002:87–88).

The selected samples in our four content analysis projects were diverse. In fact, Neuendorf (2002:2) included the entire population of medically oriented TV programs between 1951 and 1998. For my content analysis with Ken Duckworth, I had my student, Chris Gillespie (2003) draw a stratified random sample of 1,802 articles published

> in the five U.S. newspapers with the highest daily circulation in 1996 to 1997 in each of the four regions identified in the LexisNexis database, as well as the two high-circulation national papers in the database, *The New York Times* and *USA Today*. (pp. 1402–1403)

Because individual articles cannot be sampled directly in the LexisNexis database, a random sample of days was drawn first. All articles using the terms *schizophrenia* or *cancer* (or several variants of these terms) were then selected from the chosen newspapers on these days. Dill and Thill (2007:855–856) used all images in the current issues (as of January 2006) of the six most popular video game magazines sold on Amazon.com. Ambwani and Strauss (2007:15) used an availability sampling strategy, with 220 students from Introductory Psychology and a variety of other sources.

Design Coding Procedures for the Variables to Be Measured

This requires deciding what variables to measure, using the unit of text to be coded such as words, sentences, themes, or paragraphs. Then, the categories into which the text units are to be coded must be defined. These categories may be broad such as "supports democracy" or narrow such as "supports universal suffrage." Reading or otherwise reviewing some of the documents or other units to be coded is an essential step in thinking about variables that should be coded and in developing coding procedures. Development of clear instructions and careful training of coders is essential.

As an example, Exhibit 13.11 is a segment of the coding form that I developed for a content analysis of union literature that I collected during a study of union political processes (Schutt 1986). My sample was of 362 documents: all union newspapers and a stratified sample of union leaflets given to members during the years of my investigation. My coding scheme included measures of the source and target for the communication, as well as measures of concepts that my theoretical framework indicated were important in organizational development: types of goals, tactics for achieving goals, organizational structure, and forms of participation. The analysis documented a decline in concern with client issues and an increase in focus on organizational structure, which were both trends that also emerged in interviews with union members.

Developing reliable and valid coding procedures deserves special attention in a content analysis, for it is not an easy task. The meaning of words and phrases is often ambiguous. Homographs create special problems (words such as *mine* that have different meanings in different contexts), as do many phrases that have special meanings (such as "point of not return") (Weber 1990:29–30). As a result, coding procedures cannot simply categorize and count words; text segments in which the words are embedded must also be inspected before codes are finalized. Because different coders may perceive different meanings in the same text segments, explicit coding rules are required to ensure coding consistency. Special dictionaries can be developed to keep track of how the categories of interest are defined in the study (Weber 1990:23–29).

After coding procedures are developed, their reliability should be assessed by comparing different coders' codes for the same variables. Computer programs for content analysis can enhance reliability by facilitating the consistent application of text-coding rules (Weber 1990:24–28). Validity can be assessed with a construct validation approach by determining the extent to which theoretically predicted relationships occur (see Chapter 4).

Neuendorf's (2002:2) analysis of medical programming measured two variables that did not need explicit coding rules: length of show in minutes and the year(s) the program was aired. She also coded shows as comedies or noncomedies, as well as medical or not, but she does not report the coding rules for these distinctions. We provided a detailed description of

EXHIBIT 13.11 Union Literature Coding Form*

I. *Preliminary Codes*

1. Document # _____
2. Date _____
   mo yr
3. Length of text _____ pp. (round up to next 1/4 page; count legal size as 1.25)
4. Literature Type
   1. General leaflet for members/employees
   2. Newspaper/Newsletter article
   3. Rep Council motions
   4. Other material for Reps, Stewards, Delegates (e.g., budget, agenda)
   5. Activity reports of officers, President's Report
   6. Technical information-filing grievances, processing forms
   7. Buying plans/Travel packages
   8. Survey Forms, Limited Circulation material (correspondence)
   9. Non-Union
   10. Other _____ (specify)

4A. If newspaper article
Position
1. Headline story
2. Other front page
3. Editorial
4. Other

4B. If Rep Council motion
Sponsor
1. Union leadership
2. Office
3. Leadership faction
4. Opposition faction
5. Other

5. Literature content-Special issues
   1. First strike (1966)
   2. Second strike (1967)
   3. Collective bargaining (1977)
   4. Collective bargaining (1979)
   5. Election/campaign literature
   6. Affiliation with AFSCME/SEIU/other national union
   7. Other

II. *Source and Target*

6. Primary source (code in terms of those who prepared this literature for distribution).
   1. Union-newspaper (Common Sense; IUPAE News)
   2. Union-newsletter (Info and IUPAE Bulletin)
   3. Union-unsigned
   4. Union officers
   5. Union committee
   6. Union faction (the Caucus; Rank-and-Filers; Contract Action, other election slate; PLP News; Black Facts)
   7. Union members in a specific work location/office
   8. Union members-other
   9. Dept. of Public Aid/Personnel

(Continued)

(Continued)

10. DVR/DORS
11. Credit Union
12. Am. Buyers' Assoc.
13. Other non-union

7. Secondary source (use for lit. at least in part reprinted from another source, for distribution to members)
   1. Newspaper-general circulation
   2. Literature of other unions, organizations
   3. Correspondence of union leaders
   4. Correspondence from DPA/DVR-DORS/Personnel
   5. Correspondence from national union
   6. Press release
   7. Credit Union, Am. Buyers'
   8. Other _____ (specify)
   9. None

8. Primary target (the audience for which the literature is distributed)
   1. Employees-general (if mass-produced and unless otherwise stated)
   2. Employees-DVR/DORS
   3. Union members (if refers only to members or if about union elections)
   4. Union stewards, reps, delegates committee
   5. Non-unionized employees (recruitment lit, etc.)
   6. Other _____ (specify)
   7. Unclear

III. *Issues*

A. Goal

B. Employee conditions/benefits (Circle up to 5)
   1. Criteria for hiring
   2. Promotion
   3. Work out of Classification, Upgrading
   4. Step increases
   5. Cost-of-living, pay raise, overtime pay, "money"
   6. Layoffs (non-disciplinary); position cuts
   7. Workloads, Redeterminations, "30 for 40", GA Review
   8. Office physical conditions, safety
   9. Performance evaluations
   10. Length of workday
   11. Sick Benefits/Leave—holidays, insurance, illness, vacation, voting time
   12. Educational leave
   13. Grievances—change in procedures
   14. Discrimination (race, sex, age, religion, national origin)
   15. Discipline-political (union-related)
   16. Discipline—performance, other
   17. Procedures with clients, at work
   18. Quality of work, "worthwhile jobs"—other than relations with clients

*Coding instruction available from author.

coding procedures in our analysis of newspaper articles that used the terms *schizophrenia* or *cancer* (Duckworth et al. 2003). This description also mentions our use of a computerized text-analysis program and procedures for establishing measurement reliability.

> Content coding was based on each sentence in which the key term was used. Review of the full text of an article resolved any potential ambiguities in proper assignment of codes. Key terms were coded into one of eight categories: metaphor, obituary, first person or human interest, medical news, prevention or education, incidental, medically inappropriate, and charitable foundation. Fifty-seven of the 913 articles that mentioned schizophrenia, but none of those that mentioned cancer, were too ambiguous to be coded into one of these defined categories and so were excluded from final comparisons. Coding was performed by a trained graduate assistant using QSR's NUD*IST program. In questionable cases, final decisions were made by consensus of the two psychiatrist coauthors. A random subsample of 100 articles was also coded by two psychiatry residents, who, although blinded to our findings, assigned the same primary codes to 95 percent of the articles. (p. 1403)

Dill and Thill (2007) used two coders and a careful training procedure for their analysis of the magazine images about video games.

> One male and one female rater, both undergraduate psychology majors, practiced on images from magazines similar to those used in the current investigation. Raters discussed these practice ratings with each other and with the first author until they showed evidence of properly applying the coding scheme for all variables. Progress was also checked part way through the coding process, as suggested by Cowan (2002). Specifically, the coding scheme was re-taught by the first author, and the two raters privately discussed discrepancies and then independently assessed their judgments about their ratings of the discrepant items. They did not resolve discrepancies, but simply reconsidered their own ratings in light of the coding scheme refresher session. Cowan (2002) reports that this practice of reevaluating ratings criteria is of particular value when coding large amounts of violent and sexual material because, as with viewers, coders suffer from desensitization effects. (p. 856)

Ambwani and Strauss (2007) also designed a careful training process to achieve acceptable levels of reliability before their rates coded the written answers to their open-ended survey questions.

> We developed a coding scheme derived from common themes in the qualitative responses; descriptions of these themes appear in the Appendix. We independently developed lists of possible coding themes by reviewing the responses and then came to a consensus regarding the most frequently emerging themes. Next, a team of four independent raters was trained in the coding scheme through oral, written, and group instruction by the first author. The raters first coded a set of hypothetical responses; once the raters reached an acceptable level of agreement on the coding for the hypothetical responses, they began to code the actual data. These strategies, recommended by Orwin (1994), were employed to reduce error in the data coding. All four independent coders were blind to the hypotheses. Each qualitative response could be coded for multiple themes. Thus, for example, a participant who described the influence of body image in mate selection and quality of sexual relations received positive codes for both "choice" and "quality of sex." (p. 16)

Develop Appropriate Statistical Analyses

The content analyst creates variables for analysis by counting occurrences of particular words, themes, or phrases and then tests relations between the resulting variables. These analyses could use some of the statistics that will be introduced in Chapter 14, including frequency distributions, measures of central tendency and variation, cross-tabulations, and correlation analysis (Weber 1990:58–63). Computer-aided qualitative analysis programs, like those you learned about in Chapter 10 and like the one I selected for the newspaper article analysis (above), can help, in many cases, develop coding procedures and then to carry out the content coding.

The simple chart that Neuendorf (2002:3) used to analyze the frequency of medical programming appears in Exhibit 13.10. Duckworth et al.'s (2003) primary analysis was simply a comparison of percentages showing that 28% of the articles mentioning schizophrenia used it as a metaphor, compared with only 1% of the articles mentioning cancer. We also presented examples of the text that had been coded into different categories. For example, "the nation's schizophrenic perspective on drugs" was the type of statement coded as a metaphorical use of the term *schizophrenia* (p. 1403). Dill and Thill (2007:858) presented percentages and other statistics that showed that, among other differences, female characters were much more likely to be portrayed in sexualized ways in video game images than were male characters. Ambwani and Strauss (2007:16) used other statistics that showed that body esteem and romantic love experiences are related, particularly for women. They also examined the original written comments and found further evidence for this relationship. For example, one woman wrote, "[My current boyfriend] taught me to love my body. Now I see myself through his eyes, and I feel beautiful" (p. 17).

The criteria for judging quantitative content analyses of text are the same standards of validity applied to data collected with other quantitative methods. We must review the sampling approach, the reliability and validity of the measures, and the controls used to strengthen any causal conclusions.

The various steps in a content analysis are represented in the flowchart in Exhibit 13.12. Note that the steps are comparable to the procedures in quantitative survey research. Use this flowchart as a checklist when you design or critique a content analysis project.

ETHICAL ISSUES IN SECONDARY DATA ANALYSIS AND CONTENT ANALYSIS

Analysis of data collected by others, as well as content analysis of text, does not create the same potential for harm as does the collection of primary data, but neither ethical nor related political considerations can be ignored. Because in most cases the secondary researchers did not collect the data, a key ethical obligation is to cite the original, principal investigators, as well as the data source, such as the ICPSR.

Subject confidentiality is a key concern when original records are analyzed. Whenever possible, all information that could identify individuals should be removed from the records to be analyzed so that no link is possible to the identities of living subjects or the living descendants of subjects (Huston & Naylor 1996:1698). When you used data that have already