Routledge Communication Series
Jennings Bryant/Dolf Zillmann, Series Editors

Selected titles include:

Psychophysiological Measurement and Meaning
Cognitive and Emotional Processing of Media
Potter/Bolls

An Integrated Approach to Communication Theory and Research, Second Edition
Stacks/Salwen

Statistical Methods for Communication Science
Hayes

ANALYZING MEDIA MESSAGES

Using Quantitative Content Analysis in Research

Third Edition

Daniel Riffe, Stephen Lacy, and Frederick Fico

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

6

RELIABILITY

One of the questions posed in chapter 3 (this volume) was "How can the quality of the data be maximized?" To a considerable extent, the quality of data reflects the reliability of the measurement process. Reliable measurement in content analysis is crucial. If one cannot trust the measures, one cannot trust any analysis that uses those measures.

The core notion of reliability is simple: The measurement instruments applied to observations must be consistent over time, place, coder, and circumstance. As with all measurement, one must be certain that one's measuring stick does not develop distortions. If, for example, one had to measure day-to-day changes in someone's height, would a metal yardstick or one made of rubber be better? Clearly the rubber yardstick's own length would be more likely to vary with the temperature and humidity of the day the measure was taken and with the measurer's pull on the yardstick. Indeed, a biased measurer might stretch the rubber yardstick. Similarly, if one wanted to measure minority presence in television commercials, as described in chapter 2 (this volume), one would find different results by using an untrained coder's assessment or by using trained coders with explicit coding instructions.

In this chapter, we deal with reliability in content analysis. Specific issues in content analysis reliability involve the definition of concepts and their operationalization in a content analysis protocol, the training of coders in applying those concepts, and mathematical measures of reliability permitting an assessment of how effectively the content analysis protocol and the coders have achieved reliability.

Reliability: Basic Notions

Reliability in content analysis is defined as agreement among coders about categorizing content. Indeed, content analysis as a research tool is based on the assumption that explicitly defined and accepted concept definitions control assignment of content to particular categories by coders. If the category definitions do not

control assignment of content, then human biases may be doing so in unknown ways. If this is so, findings are likely to be uninterpretable and unreplicable by others. Yet replicability is a defining trait of science, as noted in chapter 2 (this volume). Reliability is thus crucial to content analysis as a scientific method. The problem of assessing reliability comes down ultimately to testing coder agreement to verify the assumption that content coding is determined by the concept definitions and operationalizations in the protocol.

Achieving reliability in content analysis begins with defining variables and categories (subdivisions of the variable) that are relevant to the study goals. Coders are then trained to apply those definitions to the content of interest. The process ends with the assessment of reliability through coder reliability tests. Such tests indicate numerically how well the concept definitions have controlled the assignment of content to appropriate analytic categories.

These steps obviously interrelate, and if any one fails, the overall reliability must suffer. Without clarity and simplicity of concept definition, coders will fail to apply them properly when looking at content. Without coder diligence and discernment in applying the concepts, the reliability assessment will prove inadequate. Without the assessment, an alternate interpretation of any study's findings could be "coder bias." Failure to achieve reliability in a content study means replication by the same or by other researchers will be of dubious value.

Concept Definitions and Category Construction

Reliability in content analysis starts with the variable and category definitions and the rules for applying them in a study. These definitions and the rules that operationalize them are specified in a content analysis protocol, a guidebook that answers the chapter 3 (this volume) question: "How will coders know the data when they see it?"

For example, two of the authors developed a protocol to study non-profit professional online news sites to compare them to news sites created by daily newspapers. The first step was to code which sites fit the concept of "non-profit professional online news sites." The variable was labeled "type of news site" and the categories were "one" or "two," with one representing these sites and two representing newspaper sites. The rules for giving the site a one were as follows: (a) It has 501c status, (b) it pays a salary to at least some of the staff, (c) its geographic market includes city, metro or regional areas, (d) it publishes general news and opinion information rather than niche information, and (e) it posts such information multiple times during the week. The protocol then explained how coders could find information to address each of these characteristics.

Conceptual and Operational Definitions

Conceptual and operational definitions specify how the concepts of interest can be recognized in the content of interest. Think of it this way: A *concept* is a broad, abstract idea about the way something is or about the way several things

interrelate. Each variable in a content analysis is the operationalized definition of that broader, more abstract concept. Each category of each content variable is an operational definition as well, but one subsumed by the broader operational definition of the variable it is part of.

A simple example makes this process clear. In the study of political visibility of state legislators (Fico, 1985), the concept of prominence was defined and measured. As an abstract concept, prominence means something that is first or most important and clearly distinct from all else in these qualities. In a news story about the legislative process, prominence can be measured (operationalized) in a number of ways. For example, a political actor's prominence in a story can be measured in terms of "how high up" the actor's name appears. Or, the actor's prominence can be assessed according to how much story space or time is taken up with assertions attributed to or about the actor. Prominence can even be assessed by whether the political actor's photo appears with the article, or his or her name is in a headline. Fico's (1985) study operationalized political prominence as the first legislator's name encountered in the news story (not in the headline or cutline).

Note several things about the concept of prominence and the way it was applied. Certainly, the concept is much more extensive than the way it was operationalized. Many concepts have more than one dimension in their meaning, and almost all have a number of ways in which one or more of those dimensions of meaning can be measured. For example, Fico's (1985) study also used the concept of political activity, citing previous research that operationalized that concept as references in the Congressional Directory—essentially a measure of a lawmaker's frequency of congressional addresses. However, the Fico study opted for a different operational definition of political activity: the number of bills a lawmaker proposed. Arguably, both—and more—really make up the various dimensions of political activity.

Certainly it can be argued that the concept of prominence is best tapped by several measures such as those noted—story position, space, accompanying photograph—but combined into an overall index. In fact, many concepts are operationalized in just this way. Of course, using several measures of a concept to create an index requires making sure that the various components indicate the presence of the same concept. For example, story space or number of paragraphs devoted to a politician may not be a good measure of prominence if he or she is not mentioned until the last paragraphs of a story.

Concept Complexity and Number of Variables

Thus, the more conceptually complex the variables and categories, the harder it will be to achieve acceptable reliability for reasons that we explain in the following section. Either more time and effort must be available for a complex analysis, or the analysis itself may have to be less extensive. That is, if the concepts are simple and easy to apply, reliability is more easily achieved. A large number of

complex concepts increases the chances that coders will make mistakes, diminishing the reliability of the study.

Reliability also is easier to achieve when a concept is more, rather than less, manifest. Recall from chapter 2 (this volume) that something manifest is observable "on its face" and, therefore, easier to recognize and count than latent content. The simpler it is to recognize when the concept exists in the content, the easier it is for the coders to agree and thus the better the chance of achieving reliability in the study. For example, recognizing when the name of a politician appears in a television situation comedy is easier than categorizing the context in which that official is mentioned. Or, if political visibility is operationalized simply as the number of times a legislator's name appears in a sample of news stories, coders will probably easily agree on the counted occurrences of that name. However, categorizing whether the legislator is discussing a specific bill or commenting on some more general political topic may require more complex judgment and thereby affect coder reliability.

Although reliability is easiest to achieve when content is more manifest (e.g., counting names), the most manifest content is not always the most interesting or significant. Therefore, content studies also may address content that also has some degree of latent meaning. Content rarely is limited to only manifest or only latent meaning. Two problems can ensue, one of which affects the study's reliability. First, as the proportion of meaning in content that is latent increases, agreement among coders becomes more difficult to achieve. Beyond within-study reliability, however, a second problem may occur that engages the interpretation of study results. Specifically, even though trained coders may achieve agreement on content with a high degree of latent meaning, it may be unclear whether naive observers of the content (e.g., newspaper readers, soap opera viewers, etc.) experience the meanings defined in the protocol and applied by the researchers. Few viewers of television commercials, for example, repeatedly rewind and review these commercials to describe the relationships among actors. Here, the issue of reliability relates in some sense to the degree to which the study and its operationalizations "matter" in the real world (and therefore to the study's validity, a topic we discuss in chapter 7, this volume).

These issues do not mean that studies involving latent meaning should not be done or that they fail to have broader meaning and significance. That depends on the goals of the research. For example, Simon, Fico, and Lacy (1989) studied defamation in stories of local conflict. The definition of defamation came from court decisions: words that tend to harm the reputation of identifiable individuals. Simon et al.'s study further operationalized "per se" defamation that harms reputation on its face, and "per quod" defamation that requires interpretation that harm to reputation has occurred. Obviously, what harms reputation depends on what the reader or viewer of the material brings to it. To call a leader "tough" may be an admirable characterization to some and a disparaging one to others. Furthermore, it is doubtful that many readers of these stories had the concept of defamation

in mind as they read (although they may have noted that sources were insulting one another). However, the goal of Simon et al.'s study was to determine when stories might risk angering one crucial population of readers: people defamed in the news who might begin a lawsuit.

These concepts of manifest and latent meaning can be thought to exist on a continuum. Some symbols are more manifest than others in that a higher percentage of receivers share a common meaning for those symbols. Few people would disagree on the common, manifest meaning of the word *car,* but the word *cool* has multiple uses as a verb and as a noun in a standard dictionary. Likewise, the latent meanings of symbols vary according to how many members of the group using the language share the latent meaning. The latent or connotative meaning also can change with time, as noted in chapter 2 (this volume). In the 1950s in America, a Cadillac was considered the ultimate automotive symbol of wealth by the majority of people in the United States. Today, the Cadillac still symbolizes wealth, but it is not the ultimate symbol in everyone's mind. Other cars, such as the Mercedes and BMW, have come to symbolize wealth as much or more so than the Cadillac.

The point is that variables requiring difficult coder decisions, whether because of concept complexity or lack of common meaning, should be limited. The more complex categories there are, the more the coding may have to be controlled by rules of procedure. Before each coding session, instructions should require that coders first review the protocol rules governing the categories. Coding sessions may be restricted to a set amount of content or a set amount of time to reduce the chance that coder fatigue will systematically degrade the coding of content toward the end of the session.

Content Analysis Protocol

However simple or complex the concepts, the definitions must be articulated clearly and unambiguously. This is done in the content analysis protocol. The protocol's importance cannot be overstated. It is the documentary record that defines the study in general and the coding rules applied to content in particular.

Purpose of the Protocol

First, the protocol sets down the rules governing the study, rules that bind the researchers in the way they define and measure the content of interest. These rules are invariant across the life of the study. Content coded on Day 1 of a study should be coded in the identical way on Day 100 of the study.

Second, the protocol is the archival record of the study's operations and definitions, or how the study was conducted. Therefore, the protocol makes it possible for other researchers to interpret the results and replicate the study. Such

replication strengthens the ability of science to build a body of findings and theory illuminating the processes and effects of communication.

The content analysis protocol can be thought of as a cookbook. Just as a cookbook specifies ingredients, amounts of ingredients needed, and the procedures for combining and mixing them, the protocol specifies the study's conceptual and operational definitions and the ways they are to be applied. To continue the analogy, if a cookbook is clear, one does not need to be a chef to make a good stew. The same is true for a content analysis protocol. If the concepts and procedures are sufficiently clear and procedures for applying them straightforward, anyone reading the protocol could code the content in the same way as the researchers. If the concepts and procedures are more complex, anyone trained in using the protocol could code the content in the same way as the researchers.

Protocol Development

Of course, making concepts sufficiently clear and the procedures straightforward may not be such a simple process. Concepts that remain in a researcher's head are not likely to be very useful. Therefore, the researcher writes it down. Although that sounds simple, the act of putting even a simple concept into words is more likely than anything else to illuminate sloppy or incomplete thinking. Defining concepts forces more discerning thinking about what the researcher really means by a concept. Others with different perspectives can then more easily apply these written definitions.

This dynamic of articulation and response, both within oneself and with others, drives the process that clarifies concepts. The process forces the researcher to formulate concepts in words and sentences that are to others less ambiguous and less subject to alternative interpretations that miss the concept the researcher had in mind.

Protocol Organization

Because it is the documentary record of the study, care should be taken to organize and present the protocol in a coherent manner. The document should be sufficiently comprehensive for other researchers to replicate the study without additional information from the researchers. Furthermore, the protocol must be available to any who wish to use it to help interpret, replicate, extend, or critique research governed by the protocol.

A three-part approach works well for protocol organization. The first part is an introduction specifying the goals of the study and generally introducing the major concepts. For example, in a study of local government coverage (Fico et al., 2013b; Lacy et al., 2012), the protocol introduction specified the content and news media to be examined (news and opinion stories in eight types of news outlets).

The second part specifies the procedures governing how the content was to be processed. For example, the protocol explained to coders which stories to be excluded and included.

The third part of the protocol specifies each variable used in the content analysis and, therefore, carries the weight of the protocol. For each variable, the overall operational definition is given along with the definitions of the values of each category. These are the actual instructions used by the coders to assign content to particular values of particular variables and categories. Obviously, the instructions for some variables will be relatively simple (news media) or complex (item topic).

How much detail should the category definitions contain? It should have only as much as necessary. As just noted, defining the concepts and articulating them in the protocol involves an interactive process. The protocol itself undergoes change, as coders in their practice sessions attempt to use the definitions, assessing their interim agreement at various stages in the training process. Category definitions become more coder friendly as examples and exceptions are integrated. Ironically, though, extremes in category definition—too much or too little detail—should be avoided. Definitions that lack detail permit coders too much leeway in interpreting when the categories should be used. Definitions that are excessively detailed may promote coder confusion or may result in coders forgetting particular rules in the process of coding.

The coding instructions shown in Table 6.1 provide an example of part of a protocol used with a national sample of news content from roughly 800 news outlets. The protocol was applied to more than 47,000 stories and had two sections. This is the first section that was applied to all stories. The second section was more complex and applied only to local government stories.

Coding Sheet

Each variable in the content analysis protocol must relate unambiguously to the actual coding sheet used to record the content attributes of each unit of content in the study. A coding sheet should be coder friendly. Coding sheets can be printed on paper or presented on a computer screen. Each form has advantages and disadvantages. Paper sheets allow flexibility when coding. With paper, a computer need not be available while analyzing content, and the periodic interruption of coding content for keyboarding is avoided. Not having interruptions is especially important when categories are complex and the uninterrupted application of the coding instructions can improve reliability. Paper sheets are useful particularly if a coder is examining content that is physically large, such as a newspaper.

Using paper, however, adds more time to the coding process. Paper coding sheets require the coder to write the value; someone else must then keyboard it into the computer. If large amounts of data are being analyzed, the time increase can be considerable. This double recording on paper and keyboard also increases the chance of transcribing error. Conversely, having the paper sheets provides a backup for the data should a hard drive crash.

TABLE 6.1 Coding Protocol for Local Government Coverage

Introduction

This protocol addresses the news and opinion coverage of local governments by daily newspapers, weekly newspapers, broadcast television stations, local cable networks, radio news stations, music radio stations, citizen blogs and citizen new sites. It is divided into two sections. The first addresses general characteristics of all local stories, and the second concerns the topic, nature and sources of local governments (city, county and regional) governments. The content will be used to evaluate the extent and nature of coverage and will be connected with environmental variables (size of market, competition, ownership, etc.) to evaluate variation across these environmental variables.

Procedure and Story Eligibility for Study

Our study deals with local public affairs reporting at the city/suburb, county, and regional government levels. These areas include the local governmental institutions closest to ordinary people and therefore more accessible to them.

A city government (also sometimes called a "township") is the smallest geopolitical unit in America. Many cities (townships) are included in counties, and many counties may be connected a regional governmental unit.

A story may NOT be eligible for coding for the following reasons:

1. The story deals with routine sports material.
2. The story deals with routine weather material.
3. The story deals with entertainment (e.g., plays).
4. The story deals with celebrities (their lives).
5. The story deals with state government only.
6. The story deals with national government only.

Read the story before coding. If you believe a story is NOT eligible for the study because it deals with excluded material noted above, go on to the next story. Consult with a supervisor on the shift if the story is ambiguous in its study eligibility.

Variable Operational Definitions

V1: **Item Number:** (assigned)

V2: **Item Date**: month/day/year (two digits: e.g., Aug. 8, 2008 is 080808)

V3: ID number **of the city** (assigned) – See list. Assign 999 if DMA sample

V4: **Item Geographic Focus**

Stories used in this analysis were collected based on their identification as "local" by the news organization. Stories that address state, national or international matters would not be included unless some "local angle" was present.

The geographic focus of the content is considered to be the locality that occurs first in the item. Such localities are indicated by formal names (e.g., a Dallas city council meeting) used first in the story.

In some cases, a formal name will be given for a subunit of a city (e.g., the "Ivanhoe Neighborhood" of East Lansing) and in these cases, the city is the focus.

Often the locality of a story is given by the dateline (e.g., Buffalo, N.Y.), but in many cases the story *must be read* to determine the locality because it may be different than

(Continued)

TABLE 6.1 (Continued)

that in a dateline. If no locality at all is given in the story, code according to the story's dateline.

1 = listed central city: see list
2 = listed suburb city: see list
3 = other local geographic area

V5: **ID number of DMA** (assigned number) – See list. Assign 99 if city council sample.
V6: **ID number of outlet:** (assigned) See list.
V7: **Type of Medium (Check ID number list):**

1 = Daily newspaper
2 = Weekly newspaper
3 = Broadcast Television
4 = Cable Television
5 = News Talk Radio
6 = Non-News Talk Radio
7 = Citizen Journalism News Site
8 = Citizen Journalism Blog Site

V8: **Organizational Origin of Content Item:**

1 = **Staff Member:** (Code story as 1 if there is any collaboration between news organization staff and some other story information source.)

a. Includes items from any medium that attribute content to a reporter's or content provider's NAME (unless the item is attributed to a source such as those under the code 2 below). A first name or a username suffices for citizen journalism sites.
b. Includes items by any medium that attribute content to the news organization name (e.g., by KTO; e.g., by the Blade; e.g., by The Jones Blog). Such attribution can also be in the story itself (e.g., KTO has learned; The Blade's Joe Jones reports).
c. Includes items that attribute content to the organization's "staff" or by any position within that organization (e.g., "editor," etc.)
d. FOR TV AND RADIO ONLY, assume an item is staff produced if
   1) a station copyright is on the story page (the copyright name may NOT match the station name.) However: If an AP/wire identification is the only one in the byline position or at the bottom of the story, code V7 as 2 even if there is a station copyright at the bottom of the page.
   2) a video box is on a TV item or an audio file is on a Radio item
e. FOR RADIO ONLY, assume an item is staff produced ALSO if the item includes a station logo inside the story box
f. FOR NEWSPAPER ONLY, assume an item is staff produced ALSO if the item includes:
   1) an email address with newspaper URL
   2) a "police blotter" or "in brief" section of multiple stories

(Continued)

TABLE 6.1 (Continued)

2 = **News and Opinion Services:**

a. This includes news wire services such as Associated Press, Reuters, Agence France Press and opinion syndicates such as King's World.
b. This includes news services such as The New York Times News Service, McClatchy News Service, Gannett News Service, Westwood One Wire.
c. This includes stories taken WHOLE from other news organizations as indicated by a different news organization name in the story's byline.

3 = **Creator's Online Site (for material identified as 7 or 8 in V8):**

a. Used ONLY for online citizen journalism sites whose content is produced by one person, as indicated by the item or by other site information.
b. If the site uses material from others (e.g., "staff," "contributors," etc.), use other V8 codes for those items.

4 = **Local Submissions:**

Use this code for WHOLE items that include a name or other identification that establishes it as TOTALLY verbatim material from people such as government or non-government local sources. The name can refer to a person or to an organization. Such material may include

a. Verbatim news releases
b. Official reports of government or nongovernment organizations
c. Letters or statements of particular people
d. Op-ed pieces or letters to the editor
e. Etc.

5 = **Can't Tell:** The item includes no information that would result in the assignment of codes 1, 2, 3, or 4 above.

The organization of the coding sheet will, of course, depend on the specific study. However, the variables on the coding sheet should be organized to follow the order of variables in the protocol, which in turn follows the flow of the content of interest. The coders should not have to dart back and forth within the content of interest to determine the variable categories. If, for example, an analysis involves recording who wrote a news story lead, that category should be coded relatively high up on the coding sheet because coders will encounter the byline first. Planning the sheet design along with the protocol requires the researcher to visualize what the process of data collection will be like and how problems can be avoided.

Coding sheets usually fall into two types: single case and multiple cases. The single case coding sheets have one or more pages for each case or recording unit.

TABLE 6.2 Coding Sheet

Content Analysis Protocol AAA for Assessing Local Government News Coverage

V1: Item Number _____

V2: Item Date _____

V3: ID number of the city _____

V4: Item Geographic Focus

1 = listed central city: see list
2 = listed suburb city: see list
3 = other local geographic area _____

V5: ID number of DMA _____

V6: ID number of outlet _____

V7: Type of Medium (Check ID number list)

1 = Daily newspaper 2 = Weekly newspaper
3 = Broadcast Television 4 = Cable Television
5 = News Talk Radio 6 = Non-News Talk Radio
7 = Citizen News Site 8 = Citizen Blog Site _____

V8: Organizational Origin of Content Item

1 = Staff Member 2 = News & Opinion Services
3 = Creator's Online Site 4 = Local Submissions
5 = Can't Tell _____

The analysis of suicide notes for themes might use a "sheet" for each note, with several content categories on the sheet.

Table 6.2 shows the single case coding sheet associated with the coding instructions given in Table 6.1. Each variable (V) and a response number or space is identified with a letter and numbers (Vl, V2, etc.) that corresponds with the definition in the coding protocol. Connecting variable locations on the protocol and on the coding sheet reduces time and confusion while coding.

Multicase coding sheets allow an analyst to put more than one case on a page. This type of coding sheet often appears as a grid, with the cases placed along the rows and the variables listed in the columns. It's the form used when setting up a computer database in Excel or SPSS. Figure 6.1 shows an abbreviated multicase coding sheet for a study of monthly consumer magazines. Each row contains the data for one issue of the magazine; this example contains data for six cases. Each column holds the numbers for the variable listed. Coders will record the number of photographs in column 4 for the issue listed on the row. For instance, the March issue in 1995 had 45 photographs in the magazine.

ID #	Month	Year	# of Photos	Pages of Food Ads	# of Health Stories	Total space	# of Stories
01	01	95	42	15	09	102	29
02	02	95	37	21	10	115	31
03	03	95	45	32	15	130	35
04	04	95	31	25	08	090	27
05	06	95	50	19	12	112	30
06	01	96	43	19	11	120	25
07	02	96	45	23	17	145	29

FIGURE 6.1 Coding sheet for monthly consumer magazines.

Coder Training

The process of concept definition, protocol construction, and coder training is a reiterative process. Central to this process—how long it goes on and when it stops—are the coders. The coders, of course, change as they engage in successive encounters with the content of interest and the way that content is captured by the concepts defined for the study. A content analysis protocol will go through many drafts during pretesting as concepts are refined, measures specified, and procedures for coding worked through.

Coding Process

This process is both facilitated and made more complex depending on the number of coders. Along with everyone else, researchers carry mental baggage that influences their perception and interpretation of communication content. A single coder may not notice the dimensions of a concept being missed or how a protocol that is perfectly clear to him or her may be opaque to another. Several coders are therefore more likely to hammer out conceptual and operational definitions that are clearer and more explicit.

Conversely, the disadvantage of collaboration among several coders is that agreement on concepts may be more difficult, or their operationalization may reveal problems that would not occur with fewer or with only one coder. At some point, a concept or its measure may just not be worth further expenditure of time or effort, and recognizing that point may not be easy either.

Variable definitions are not very useful if they cannot be applied reliably. Although the protocol may be well organized and clearly and coherently written,

a content analysis must still involve systematic training of coders to use the protocol. An analogy to a survey is useful. Survey administrators must be trained in the rhythms of the questionnaire and gain comfort and confidence in reading the questions and recording respondent answers. Coders in a content analysis must grow comfortable and familiar with the definitions of the protocol and how they relate to the content of interest.

The first step in training coders is to familiarize them with the content being analyzed. The aim here is not to precode material, and indeed, content not in the study sample should be used for this familiarization process. The familiarization process is meant to increase the coders' comfort level with the content of interest, to give them an idea of what to expect in the content, and to determine how much energy and attention is needed to comprehend it.

To help minimize coder differences, the study should establish a procedure that coders follow in dealing with the content. For example, that procedure may specify how many pieces of content a coder may deal with in a session or a maximum length of time governing a coding session. The procedure may also specify that each coding session must start with a full reading of the protocol to refresh coder memory of category definitions.

Coders also should familiarize themselves with the content analysis protocol, discussing it with the supervisor and other coders during training and dealing with problems in applying it to the content being studied. During these discussions, it should become clear whether the coders are approaching the content from similar or different frames of reference. Obviously, differences will need to be addressed because these will almost certainly result in disagreements among coders and poor study reliability.

Sources of Coder Disagreements

Differences among coders can have a number of origins. Some are relatively easy to address, such as simple confusion over definitions. Others may be impossible to solve, such as a coder who simply does not follow the procedure specified in the protocol.

Category Problems

Differences over category definitions must be seriously addressed in training sessions. Does disagreement exist because a category is ambiguous or poorly articulated in the protocol? Or is the problem with a coder who just does not understand the concept or the rules for operationalizing it? Obviously, when several coders disagree on a category value, the strong possibility exists that the problem is in the category or variable. A problem may occur because of fundamental ambiguity or complexity in the variable or because the rules assigning content to the variable categories are poorly spelled out in the protocol.

The simplest approach to such a variable or category problem is to begin by revising its definition to remove the sources of ambiguity or confusion. If this revision fails to remove the source of disagreement, attention must be turned to the fundamental variable categories and definitions. It may be that an overly complex variable or its categories can be broken down into several parts that are relatively simpler to handle. For example, research on defamation (Fico & Cote, 1999) required initially that coders identify defamation in general and following that, coding copy as containing defamation per se and defamation per quod. Defamation per quod is interpreted by courts to mean that the defamation exists in the context of the overall meanings that people might bring to the reading. With this definition, coder reliability was poor. However, better reliability was achieved on recognition of defamation in general and defamation per se. The solution was obvious: Given defamation in general, defamation per quod was defined to exist when defamation per se was ruled out. In other words, if all defamation was either per se or per quod, getting a reliable measure of per se was all that was necessary to also define reliably the remaining part of defamatory content that was per quod.

However, researchers may also have to decide if a category must be dropped from the study because coders cannot use it reliably. In another study of how controversy about issues was covered in the news (Fico & Soffin, 1995), the coders attempted to make distinctions between "attack" and "defense" assertions by contenders on these issues. In fact, the content intermixed these kinds of assertions to such a degree that achieving acceptable reliability proved impossible.

Coder Problems

If only one coder is consistently disagreeing with others, the possibility exists that something has prevented that coder from properly applying the definitions. Between-coder reliability measures make it easy to identify problem coders by comparing the agreement of all possible pairs of coders. Attention must then be given to retraining that coder or removing him or her from the study.

There may be several reasons why a coder persistently disagrees with others on application of category definitions. The easiest coder problems to solve involve applications in procedure. Is the coder giving the proper time to the coding? Has the protocol been reviewed as specified in the coding procedures?

In some cases, if the content involves specialized knowledge, the coders may need to be educated. For example, some of the eight coders involved in the project about local government knew little about the structure, officials, and terms associated with local government. Therefore, the principal investigators created a booklet about local government and had the coders study it.

More difficult problems involve differences in cultural understanding or frame of reference that may be dividing coders. These differences will be encountered most frequently when coders deal with variables with relatively less manifest

content. As just noted, such content requires more coder interpretation about the meaning of content and its application to content categories.

One author recalls working as a student on a content study in a class of students from the United States, Bolivia, Nigeria, France, and South Africa. The study involved applying concepts such as terrorism to a sample of stories about international relations. As might be imagined, what is terrorism from one perspective may well be national liberation from another. Such frame of reference problems are not impossible to overcome, but they will increase the time needed for coder training. Such issues should also signal that the study itself may require more careful definition of its terms in the context of such cultural or social differences.

Peter and Lauf (2002) examined factors affecting intercoder reliability in a study of cross-national content analysis, which was defined as comparing content in different languages from more than one country. Peter and Lauf concluded that some coder characteristics affected intercoder reliability in bilingual content analysis. However, most of their recommendations centered on the failure to check reliability among the people who trained the coders. The conclusion was that cross-country content analysis would be reliable if three conditions are met: "First, the coder trainers agree in their coding with one another; second, the coders within a country group agree with one another; and, third, the coders agree with the coding of their trainers" (Peter & Lauf, 2002, p. 827).

Coder Reliability Assessment

Coder Reliability Tests

Ultimately, the process of concept definition and protocol construction must cease. At that point, the researcher must assess the degree to which the content definitions and procedures can be reliably applied. Reliability falls into three types (Krippendorff, 2004a, pp. 214–216): stability, reproducibility, and accuracy. Stability refers to *intracoder reliability* where a coder applies the protocol to the same content at two points in time. This "within-coder" assessment tests whether slippage has occurred in the single coder's understanding or application of the protocol definitions. Checking intracoder reliability is needed with coding that lasts for a long period of time, but there is no accepted definition of a "long period of time." However, if a project takes more than a month of coding, intracoder reliability testing would improve the argument for data validity.

Reproducibility involves two or more coders applying the protocol to the same content. Each variable in the protocol is tested for *intercoder reliability* by looking at agreement among coders in applying relevant category values to the content. For example, two coders code 10 Web site stories dealing with a conflict over abortion. Coding the variable for the fairness of stories, as indicated by source citation of both pro-life and pro-choice sides, they compute the percentage

of those stories on which they have agreed that the particular story is fair or unfair according to the coding definitions.

The third reliability is accuracy, which addresses whether or not the coding is consistent with some external standard for the content, much as one re-sets (or "calibrates") a household clock to a "standard" provided by one's mobile phone after a power outage. The problem in content analysis is how to come by a standard that is free of error. One way is to compare the content analysis data with a standard set by experts, but there is no way to verify that the expert's standards are free of bias. Therefore, most content analysis is limited to testing reproducibility.

Coder training sessions constitute a kind of reliability pretest. However, wishful thinking and post hoc rationalizations of why errors have occurred (e.g., "I made that mistake only because I got interrupted by the phone while coding that story") mean a more formal and rigorous procedure must be applied. In fact, formal coder reliability tests should be conducted during the period of coder training itself as a indicator of when to proceed with the study, as noted in chapter 3 (this volume). Such training tests should not, of course, be conducted with the content being used for the actual study because a coder must code study content independently, both of others and of herself or himself. If content is coded several times, prior decisions contaminate subsequent ones. Furthermore, repeated coding of the same content inflates the ultimate reliability estimate, thus giving a false confidence in the study's overall reliability.

At some point, the training formally stops, and the actual assessment of achieved reliability must take place. Two issues must be addressed: The first concerns selection of content used in the reliability assessment. The second concerns the actual statistical reliability tests that will be used.

Selection of Content for Testing

If the number of content units being studied is small, protocol reliability can be established by having two or more coders code all the content. Otherwise, researchers need to randomly select content samples for reliability testing. Most advice has been arbitrary and ambiguous about how much content to use when establishing protocol reliability. One text (Wimmer & Dominick, 2003) suggests that between 10% and 25% of the body of content should be tested. Others (Kaid & Wadsworth, 1989) suggested that between 5% and 7% of the total is adequate. One popular online resource (http://matthewlombard.com/reliability/index_print.html) suggests that the reliability sample "should not be less than 50 units or 10% of the full sample, and it rarely needs to be greater than 300 units." However, the foundations for these recommendations are not always clear.

The number of units needed will be addressed later, but probability sampling should be used when a census is impractical. Random sampling, relying on unbiased mathematical principles for selection of observations, accomplishes

two things. First, it controls for the inevitable human biases in selection. Second, the procedure produces, with a known probability of error, a sample that reflects the appropriate characteristics in the overall population of content being studied. Without a random sample, inference that the reliability outcome represents all the content being studied cannot be supported.

Given a random sample of sufficient size, the coder reliability test should then reflect the full range of potential coding decisions that must be made in the entire body of material. The problem with nonrandom selection of content for reliability testing is the same as the problem with a nonrandom sample of people: Tested material may be atypical of the entire body of content that will be coded. A nonrepresentative sample yields reliability assessments whose relation to the entire body of content is unknown.

Using probability sampling to select content for reliability testing also enables researchers to take advantage of sampling theory to answer the question of how much material must be tested. Random sampling can specify sampling error at known levels of confidence. For example, if two researchers using randomly sampled content achieve a 90% level of agreement, the actual agreement they would achieve coding all material could vary above and below that figure according to the computed sampling error. That computed sampling error would vary with the size of the sample—the bigger the sample the smaller the error and the more precise the estimate of agreement. Therefore, if the desired level of agreement is 80%, and the achieved level on a coder reliability test is 90% plus or minus 5 percentage points, the researchers can proceed with confidence that the desired agreement level has been reached or exceeded. However, if the test produced a percentage of 84%, the plus or minus 5% sampling error would include a value of 79% that is below the required standard of 80%.

Selection Procedures

Assuming content for a reliability test will be selected randomly, how many units of content must be selected? Lacy and Riffe (1996) noted that this will depend on several factors: the total number of units to be coded, the desired degree of confidence in the eventual reliability assessment, and the degree of precision desired in the reliability assessment.

Although each of these three factors is under the control of the researcher, a fourth factor must be assumed on the basis of prior studies, a pretest, or a guess. That is the researcher's estimate of the actual agreement that would have been obtained had all the content of interest been used in the reliability test. For reasons that we explain later, it is our recommendation that the estimate of actual agreement be set 5 percentage points higher than the minimum required reliability for the test. This 5-percentage point buffer will ensure a more rigorous test, that is, the achieved agreement will have to be higher for the reliability test to be judged adequate.

The first object in applying this procedure is to compute the number of content cases required for the reliability test. When researchers survey a population they use

the formula for the standard error of proportion to estimate a minimal sample size necessary to infer to that population at a given confidence level. A similar procedure is applied here to a population of content. One difference, however, is that a content analysis population is likely to be far smaller than the population of people that is surveyed. This makes it possible to correct for a finite population size when the sample makes up 20% or more of the population. This has the effect of reducing the standard error and giving a more precise estimate of reliability.

The formula for the standard error can be manipulated to solve for the sample size needed to achieve a given level of confidence. This formula is

$$n = \frac{(N-1)(SE)^2 + PQN}{(N-1)(SE)^2 + PQ}$$

in which

N = the population size (number of content units in the study)
P = the population level of agreement
$Q = (1 - P)$
n = the sample size for the reliability check

Solving for n gives the number of content units needed in the reliability check. Note that standard error gives the confidence level desired in the test. This is usually set at the 95% or 99% confidence level (using a one-tailed test because interest is in the portion of the interval that may extend below the acceptable reliability figure).

For the rest of the formula, N is the population size of the content of interest, P is the estimate of agreement in the population, and Q is 1 minus that estimate.

As an example, a researcher could assume an acceptable minimal level of agreement of 85% and P of 90% in a study using 1,000 content units (e.g., newspaper stories). One further assumes a desired confidence level of .05 (i.e., the 95% confidence level). A one-tailed z score—the number of standard errors needed to include 95% of all possible sample means on agreement—is 1.64 (a two-tailed test z score would be 1.96). Because the confidence level is 5% and our desired level of probability is 95%, SE is computed as follows:

$$.05 = 1.64(SE)$$

or

$$SE = .05/1.64 = .03.$$

Using these numbers to determine the test sample size to achieve a minimum 85% reliability agreement and assuming P to equal 90% (5% above our minimum), the results are

$$n = \frac{(999)(.0009) + .09(1000)}{(999)(.0009) + .09}$$

$$n = 92$$

In other words, 92 test units out of the 1,000 are used (e.g., newspaper stories) for the coder reliability test. If a 90% agreement in coding a variable on those 92 test units is achieved, chances are 95 out of 100 that at least an 85% or better agreement would exist if the entire content population were coded by all coders and reliability measured.

Once the number of test units needed is known, selection of the particular ones for testing can be based on any number of random techniques. For example, if study content has been numerically ordered from 1 to 1,000, a random number program can identify the particular units for the test, or a printed table of random numbers can be used.

The procedure just described is also applicable to studies in which coding categories are at interval or ratio scales. The calculation of standard error is the only difference.

If these formulas seem difficult to use, two tables may be useful. Tables 6.3 and 6.4 apply to studies that deal with nominal-level percentage of agreement; Table 6.3 is configured for a 95% confidence level, and Table 6.4 is configured for the more rigorous 99% confidence level. Furthermore, within each table, the number of test cases needed has been configured for 85%, 90%, and 95% estimates of population coding agreement.

Researchers should set the assumed level of population agreement (P) at a high enough level (we recommend 90% or higher) to assure that the reliability sample includes the range of category values for each variable. Otherwise, the sample will not represent the population of content.

For some studies, particularly smaller ones, the selection process for the reliability test just recommended will result in a large proportion of all the cases being used in the test. Researchers can always use all study content to test reliability, which eliminates sampling error. If they do not use a census of study content, the cases used for coder reliability testing must be randomly selected from the population of content of interest to have confidence in the reliability results. The level of sampling error for reliability samples should always be reported.

TABLE 6.3 Content Units Need for Reliability Test Based on Various Population Sizes, Three Assumed Levels of Population Intercoder Agreement, and a 95% Level of Probability

	Assumed Level of Agreement in Population		
	85%	*90%*	*95%*
Population Size			
10,000	141	100	54
5,000	139	99	54
1,000	125	92	52
500	111	84	49
250	91	72	45
100	59	51	36

TABLE 6.4 Content Units Need for Reliability Test Based on Various Population Sizes, Three Assumed Levels of Population Intercoder Agreement, and a 99% Level of Probability

	Assumed Level of Agreement in Population		
	85%	*90%*	*95%*
Population Size			
10,000	271	193	104
5,000	263	190	103
1,000	218	165	95
500	179	142	87
250	132	111	75
100	74	67	52

When to Conduct Reliability Tests

The process of establishing reliability involves two types of tests. The first is really a pretest that occurs during training. This process involves developing a reliable protocol rather than testing how well coders apply that protocol. As mentioned earlier, reliability pretesting serves as part of an iterative process of coding, examining reliability, adjusting the protocol, and coding again. Just how long this process continues reflects several factors, but it should continue until the reliability has reached an acceptable level, as discussed later. Formal reliability pretests will determine this point.

Once the pretests demonstrate that the protocol can be applied reliably, the study coding should begin. It is during the coding of the study content units that protocol reliability is established. As the coding gets underway, the investigators must select the content to be used for the reliability test, as described earlier. Generally, it is a good idea to wait until about 10% of the coding has been completed to begin the reliability test. This will allow coders to develop a routine for coding and become familiar with the protocol.

The content used for establishing and reporting reliability should be coded by all the coders, and the reliability content should be interspersed with the study content so the coders do not know which content units are being coded by everyone. This "blind" approach to testing reliability will yield a better representation of the reliability than having an identifiable set of reliability content coded separately from the normal coding process. If coders are aware of the test content, they might try harder or become nervous. In either case, the reliability results could be influenced.

If the study's coding phase will exceed a month in length, the investigators should consider establishing intracoder reliability, as discussed previously, and administering multiple intercoder tests. As with the initial reliability tests, content for intracoder and additional intercoder tests should be randomly selected from the study content. However, if the initial reliability test demonstrated sufficient reliability, the additional samples do not need to be as large as in the initial test. Samples in the 30 to 50 ranges should be sufficient to demonstrate the maintenance of reliability.

Most content analysis projects do not involve enough content to require more than two reliability tests, but in some cases, the investigators should consider more tests. For example, the coding of local government news coverage mentioned previously (Fico et al., 2013a; Lacy et al., 2012) lasted more than 4 months and involved three reliability tests. When additional tests are involved, a second one should take place after half of the content has been coded and before 60% has been coded. If a third tests occurs, it should be in the 80% and 90% completion range.

A major concern with longer projects is what to do if reliability falls below acceptable levels. If the reliability of the protocol in the initial test is high, any deterioration will likely reflect problems with coders. If this happens, coders whose reliability has slipped have to be identified and either retrained or dropped from the study. Of course, any content they coded since the last reliability test will need to be recoded by other coders.

Reliability Coefficients

The degree of reliability that applies to each variable in a protocol is reported with a reliability coefficient, which is a summary statistic for how often coders agreed on the classification of the content units. The research literature contains a number of such coefficients, but four are most often used in communication studies: percentage of agreement (also called Holsti's coefficient), Scott's pi, Cohen's kappa, and Krippendorff's alpha. The first overstates the true level of reliability because it does not take chance agreement into consideration. However, the latter three coefficients do consider chance agreement.

A number of sites and/or software programs are available to calculate reliability coefficients. However, it helps us understand what the computer output means by examining the process by which the coefficients are calculated. The first step regardless of the coefficient used is to select the cases to use in the calculation. Cases for these tests must either be a census of all content units being analyzed or be randomly sampled. As discussed previously, this sample must contain content units that fit into all the variables and all the categories within the variables to be a true test of reliability. This is likely to happen with an adequate sample size. If the sample is selected and reliability tested and some categories are empty, then a larger sample should be selected.

Percentage of Agreement

The coefficient that has been around the longest is the percentage of agreement among two or more coders. In this reliability test, coders determine the proportion of correct judgments as a percentage of total judgments made. All coding decisions can be reduced to dichotomous decisions for figuring simple agreement. In such cases, each possible pair of coders is compared for agreement or disagreement. For example, if three coders categorize an article, the total number of dichotomous coding decisions will equal three: Coders A and B, Coders B and C, and Coders A and C. Four coders will yield six decisions for comparison (A and B, A and C, A and D, B and C, B and D, and C and D, etc.).

Percentage of agreement coefficients overestimate reliability because the chances of accidentally agreeing increase as the number of coders decreases. However, the fact that agreement can take place by chance does not mean it does. It is not automatically true that 50% of the agreements between two coders were due to chance. All agreements could easily be the result of a well-developed protocol.

Although percentage of agreement inflates reliability, it is useful during protocol development and coder training as way of identifying where and why disagreements are occurring. Percentage of agreement also helps understanding the nature of the data by comparing it with other reliability coefficients. As discussed later, sometimes a study has high level of agreement but low coefficients that take chance into agreement. Examining these together can help future studies improve the protocol. Because of this, content analysis studies should report both a simple agreement figure and one or more of the coefficients mentioned later. The simple agreement figures should be placed in an endnote as information for researchers conducting replications. However, decisions about the reliability of a variable in the protocol should be based on a coefficient that takes chance agreement into consideration.

Coefficients That Evaluate Chance Agreement

Consider the possibility that some coder agreements might occur among untrained coders who are not guided by a protocol. These would be "chance agreements." One of the earliest reliability coefficients that "corrects" for chance agreement is Scott's pi (Scott, 1955). It involves only two coders and is used with nominal data. Correcting for chance leads to the calculation of "expected agreement" using basic probability theory. Scott's pi computes expected agreement by using the proportion of times particular values of a category are used in a given test.

Here is an example. Assume that a variable has four categories (values) for topic of news content (government, crime, entertainment, and sports) and that two coders have coded 10 items of content for a total of 20 coding decisions. Government has been used 40% of the time (i.e., eight of the combined decisions

by the two coders selected government as the correct coding category), sports has been used 30% of the time (in six decisions), and crime (in three decisions) and entertainment (three decisions) have each been used 15% of the time. Here is where the multiplication rules of probability apply. We multiply because chance involves two coders and not one. The probability of a single "event" (a story's being about government is) equals .4, but the probability of two such events (two coders coding the same variable as government) requires .4 to be multiplied by .4. This, of course, makes intuitive sense: A single event is more likely to occur than two such events occurring.

In this example, the expected agreement is .4 times .4 (government stories) or .16; plus .3 times .3 (sports stories) or .09, plus .15 times .15 (crime stories) or .022, plus .15 times .15 (entertainment stories) or .022. The expected agreement by chance alone would then be the sum of the four products, .29 (29%), or .16 + .09 + .022 + .022

The computing formula for Scott's pi is

$$\text{Pi} = \frac{\%OA - \%EA}{1 - \%EA}$$

in which
OA = observed agreement
EA = expected agreement

In this formula, OA is the agreement achieved in the reliability test, and EA is the agreement expected by chance, as just illustrated. Note that the expected agreement is subtracted from both the numerator and denominator. In other words, chance is eliminated from both the achieved agreement and the total possible agreement.

To continue with the example, suppose the observed agreement between two coders coding the four-value category for 10 news stories is 90% (they have disagreed only once). In this test, Scott's pi would be

$$\text{Pi} = \frac{90 - .29}{1 - .29} = \frac{.61}{.71} = .86$$

That .86 can be interpreted as the agreement that has been achieved as a result of the category definitions and their diligent application by coders after a measure of possible chance agreement has been removed. Finally, Scott's pi has an upper limit of 1.0 in the case of perfect agreement and a lower limit of −1.0 in the case of perfect disagreement. Figures around 0 indicate that chance is more likely governing coding decisions than the content analysis protocol definitions and their application. The example cited applies to two coders. If more than two coders are used, Krippendorff's alpha would be appropriate.

A number of other forms for assessing the impact of chance are available. Cohen (1960) developed kappa, which has the same formula as Scott's pi:

$$\text{Kappa} = \frac{P_O - P_e}{1 - P_e}$$

in which:
P_O = observed agreement
P_e = expected agreement

Kappa and pi differ, however, in the way expected agreement is calculated. Recall that Scott (1955) squared the observed proportions used for each value of a category assuming all coders are using those values equally. In other words, if 8 of 20 decisions were to select government (value 1) of a category, .4 is squared regardless of whether one of the coders used that value six times and the other only two. However, kappa uses expected agreement based on the proportion of a particular value of a category used by one coder multiplied by the proportion for that value used by the other coder. These proportions are then added for all the values of the category to get the expected agreement.

In the example, one coder has used the value of 1 in 6 of 10 decisions (.6), and the second coder has used the value of 1 in 2 of 10 decisions (.2). Therefore, whereas pi yielded the expected value of .16 (.4 × .4), kappa yields an expected value of .12 (.6 × .2). Kappa will sometimes produce somewhat higher reliability figures than pi, especially when one value of a category is used much more often than others. For further explanation of kappa, see Cohen (1960).

Kappa is used for nominal-level measures, and all disagreements are assumed to be equivalent. However, if disagreements vary in their seriousness (e.g., a psychiatrist reading a patient's diary's content concludes the person has a personality disorder when the person is really psychotic), then a weighted kappa (Cohen, 1968) has been developed.

Krippendorff (1980) developed a coefficient, alpha, that is similar to Scott's pi. Krippendorff's alpha is presented by the equation

$$\text{Alpha} = \frac{D_o}{D_e}$$

in which:
D_o = observed disagreement
D_e = expected disagreement

The process of calculating D_o and D_e depends on the level of measurement (nominal, ordinal, interval, and ratio) used for the content variables. The difference between alpha and pi is that Krippendorff's (1980) statistic can be used with non-nominal data and with multiple coders. The alpha also corrects for small samples

(Krippendorff, 2004a) and some computer programs for alpha can accommodate missing data. When nominal variables with two coders and a large sample are used, alpha and pi are equal. For more details about alpha, see Krippendorff (2004a).

Krippendorff (2004b) stated that an agreement coefficient can be an adequate measure of reliability under three conditions. First, content to be checked for reliability requires two or more coders working independently applying the same instructions to the same content. Second, a coefficient treats coders as interchangeable and presumes nothing about the content other than that it is separated into units that can be classified. Third, a reliability coefficient must control for agreement due to chance. Krippendorff pointed out that most formulas for reliability coefficients have similar numerators that subtract observed agreement from 1. However, they vary as to how the denominator (expected agreement) is calculated. Scott's pi and Krippendorff's alpha are the same, except alpha adjusts the denominator for small sample bias, and Scott's pi exceeds alpha by $(1 - \pi)/n$, where n is the sample size. As n increases, the difference between pi and alpha approaches zero.

In an article, Krippendorff (2004b) criticized Cohen's kappa because expected disagreement is calculated by multiplying the proportion of a category value used by one coder by the proportion used by the other coder (as described earlier). Krippendorff said the expected disagreement is based, therefore, on the coders' preferences, which violates the second and third of his three conditions. Krippendorff concluded that Scott's pi is acceptable with nominal data and large samples, although what qualifies as large was not defined. In situations in which data other than nominal measures are used, multiple coders are involved, and the samples are small, alpha is recommended.

One criticism of alpha offered by Lombard, Snyder-Duch, and Bracken (2004) is that it can be difficult to calculate; they called for the development of software that will calculate it for all levels of data. Hayes and Krippendorff (2007) responded by offering an SPSS Macro to calculate alpha. Hayes has made the Macro ("KALPHA") available for free on his Web site (http://www.afhayes.com/spss-sas-and-mplus-macros-and-code.html).

Pearson's Product-Moment Correlation

Pearson's correlation coefficient (r) is sometimes used as a check for accuracy of measurement with interval- and ratio-level data. This statistic, which we explain more fully in chapter 8 (this volume), measures the degree to which two variables, or two coders in this case, vary together. Correlation coefficients can be used when coders are measuring space or minutes. With this usage, the coders become the variables, and the recording units are the cases. If, for example, two coders measured the length in seconds of news stories on network evening news devoted to international events, a correlation coefficient would measure how similarly the coders were in their use of high and low scale values to describe the length of those stories relative to one another.

Krippendorff (1980) warned against using correlations for reliability because association is not necessarily the same as agreement. However, this is not a problem if agreement and accuracy of measurement are determined separately. The correlation coefficient is used not to measure category assignment but to measure the consistency of measuring instruments such as clocks and rulers.

Controversy About Reliability Coefficients

Recent years have seen a growing debate about which of the many reliability coefficients is most appropriate as an omnibus coefficient for estimating reliability, including a call to use newly developed coefficients (Gwet, 2008; Krippendorff, 2012; Zhao, 2012; Zhao, Liu, & Deng, 2012). The details of this debate are more complex than can be discussed in the space available here. However, the debate revolves around the process of calculating expected agreement. The usual reliability coefficients have been criticized because they can produce low coefficients despite high percentages of agreements among coders (Gwet, 2008; Krippendorff, 2011; Lombard, Synder-Duch, & Bracken, 2004; Potter & Levine-Donnerstein, 1999; Zhao, 2012; Zhao et al., 2012) and because they conservatively assume—and correct for—a *maximum* level of chance agreement (Zhao, 2012; Zhao et al., 2012).

More than 30 years ago, Kraemer (1979) pointed out the high agreement/low reliability conundrum with Cohen's kappa. Potter and Levine-Donnerstein (1999) have discussed the same phenomenon occurring with Scott's pi. They note that with a two-valued measure (e.g., is a terms-of-use agreement link clearly visible on a Web site's front page—yes or no?), the frequent occurrence of one value and the infrequent occurrence of the other creates an imbalance (e.g., 97% of sites have the agreement link while 3% do not), which is "overcorrected" in the chance agreement component of Scott's pi. Other scholars have recently joined the discussion, noting that this phenomenon occurs with kappa (Gwet, 2008) and with kappa, pi, and alpha (Zhao, 2012; Zhao et al., 2012).

Solutions offered by scholars have varied. Potter and Levine-Donnerstein (1999) suggested the use of the normal approximation of the binomial distribution to calculate chance agreement (e.g., use 50% with two coders, 33% with three coders, etc.). Two authors (Cicchetti & Feinstein, 1990; Feinstein & Cicchetti, 1990) examined the use of kappa as a way to evaluate diagnosis agreement in a clinical setting. They suggested that the problem could be solved by using kappa but also adding two other measures—positive agreements and negative agreements, which would allow for the diagnosis of a low kappa.

Gwet (2008) addressed the high agreement/low reliability phenomenon with pi and kappa by developing a new coefficient: AC1. He divided agreement and disagreements into groups based on four conditions for two coders and two categories: (a) both coders assign values based on the correct application of the protocol and (b) both assign based on randomness; (c) one assigns randomly and

(d) the other assigns based on the correct application of the protocol. He argued that kappa and pi assume only two types of agreement—one based on the correct application of the protocol and one based on randomness—which ignores the other two conditions. He conducted a Monte Carlo study with data from psychology and psychiatry and concluded that AC_1 produces lower variances than pi and kappa and provides a better estimate of agreement.

An effort to deal with the high agreement/low reliability phenomenon in communication studies was produced by Zhao (2012), who criticized kappa, pi, and alpha because they depend on the distribution of agreements and number of categories to calculate chance agreement. Zhao argues that it is coding difficulty that determines chance agreement and not distribution of agreements and categories. He, therefore, developed $alpha_i$, which calculates chance agreement based on disagreements. His paper also includes a Monte Carlo study based on human coders and concludes that actual level of agreement correlated highest with his new coefficient ($alpha_i$), followed by percentage of agreement, and Gwet's AC_1. Kappa, pi, and alpha correlated with actual level of agreement at only .559. The "behavior Monte Carlo" study is limited, however, because it used coding of visual representations rather than symbolic meaning usually found in communication content analysis.

Krippendorff (2011, 2012) has responded to criticisms of alpha by saying variables that do not vary are not very useful and that the uneven distribution could represent an inadequate reliability sample. These criticisms may be valid in some situations, but there are situations when the population distribution may actually be extremely uneven in its distribution among categories. Historically, for example, the percentage of television characters who are people of color is small (Mastro & Greenberg, 2000). Similar situations have been found in television representation of the elderly (Signorielli, 2001). Certainly, the study of representation in media is worthwhile, and the resulting distribution among categories will often be uneven. Depending on the exact distribution, kappa, pi, and alpha could be considerably lower than percentage of agreement. It would not matter how large the reliability sample is because the population itself has the same "one-sided" distribution. In some situations, even if the coding has only a small amount of error, the reliability will appear lower than it actually is.

Selecting a Reliability Coefficient

The discussion about the limits and advantages of various reliability coefficients will continue. Different positions reflect the assumptions on which those positions are based. Much of the discussion, however, misses the point of reliability tests, which is to help scholars establish that their data have achieved a sufficient level of reliability to proceed with their analysis. Too often, the discussion of reliability occurs without reference to validity, which is the real goal of data creation. In addition, the examples of imbalanced distributions used in the debate are often too extreme to help scholars who must deal with skewed distributions that are not as extreme as those examples. Finally, most of the discussion has concentrated on the mathematical evaluation of reliability coefficients, while more empirical research is needed to illustrate the practical implications of the mathematics (Gwet, 2008; Zhao, 2012).

In addition, much of this discussion ignores the role of sampling error. If reliability is established with a probability sample, as it should be, whatever coefficient used is just an estimate of what the coefficient would be if the entire population were tested. As such, all reports of reliability should include sampling error.

Recommendations for Establishing Reliability

In light of the debate about how best to establish reliability we suggest the following:

1. Report the reliability for each variable in the study. Replication requires this. An average or "overall" measure of reliability can hide weak variables, and not every study will want to use every variable from previous studies.
2. An important issue is what is an acceptable level for a given coefficient. Krippendorff (2004a) suggested that an alpha of .8 indicates adequate reliability. However, Krippendorff has also written that variables with alphas as low as .667 could be acceptable for drawing tentative conclusions. Unpublished research by the authors of this text suggests pi coefficients below .8 can lead to some invalid conclusions from nominal-level data and that pi coefficients below .75 are particularly problematic in reaching valid conclusions.
3. Compute and report multiple measures of reliability. The true level of reliability appears to fall somewhere between percentage of agreement and alpha. Until research sheds light on which, if any, reliability coefficient is an acceptable omnibus coefficient, we suggest scholars report the percentage of agreement, alpha (http://www.afhayes.com/spss-sas-and-mplus-macros-and-code.html) and $alpha_i$ (http://reliability.hkbu.edu.hk/). The confidence intervals for alpha and $alpha_i$ also should be reported. If the low side of the confidence intervals for both these coefficients exceeds the .80-level, the reliability is acceptable. If the confidence intervals for both coefficients do not exceed the .80 level, the authors must provided a more detailed argument for why the data are reliable and valid.
4. Adhering to these rules will require that an adequate sample size be randomly selected (see the section earlier in this chapter on sample size). All of the categories for each variable should be in the sample. If they are not, then the sample size should be increased.

Equally important to establishing reliability for variables in a given protocol is the establishment of variable reliability across time. Social science advances through improved measurement, and improved measurement requires consistent reliability. If scholars aim to standardize protocols for commonly used variables, the reliability of these protocols will increase over time.

Summary

Conducting and reporting reliability assessments in content analysis are a necessity, not a choice. However, a study (Riffe & Freitag, 1997) of 25 years (from 1971–1995) of content analysis research in *Journalism & Mass Communication Quarterly* indicated that only 56% of such studies reported that assessment. During this same period, the percentage of content analysis articles in *Journalism & Mass Communication Quarterly* climbed from 6% to nearly 35%, and after 1978, no fewer than 20% of the studies in that journal were content analyses. Yet, even in the most recent years of that period studied, from 1991 to 1995, nearly 29% of content studies failed to report an intercoder reliability assessment.

This situation has hardly improved since then. An unpublished review of 80 quantitative content analysis studies published in *Journalism & Mass Communication Quarterly* since 1998 showed that some 26% failed to conduct or report the results of reliability testing. Only 16% of the studies conducted the reliability testing on randomly selected content, included a test for chance agreement, and reported the reliability figures for all relevant variables—the three reliability requirements emphasized in this book.

Only one in three studies randomly selected content for the test. About 46% of the studies used a measure of chance agreement, and only 54% of the studies provided the test results for all variables or at least provided the range of results for the relevant variables.

Moreover, full information on the content analysis should be disclosed or at least made available for other researchers to examine or use. A full report on content analysis reliability would include protocol definitions and procedures. Because space in journals is limited, the protocol should be made available by study authors on request. Furthermore, information on the training of judges, the number of content items tested, and how they were selected should be included in footnotes or appendixes to the protocol. At a minimum, the specific coder reliability tests applied and the achieved numeric reliability along with confidence intervals should be included for each variable in the published research.

In applying reliability tests, researchers should randomly select a sufficient number of units for the tests (Lacy & Riffe, 1996), apply and make decisions on whether the variables reach acceptable reliability levels based on coefficients that take chance into consideration, and report simple agreement in an endnote to assist in the replication of the study.

Failure to systematize and report the procedures used as well as to assess and report reliability virtually invalidates whatever usefulness a content study may have for building a coherent body of research. Students must be taught the importance of assessing and reporting content analysis reliability. Journals that publish content studies should insist on such assessments.

7

VALIDITY

When we introduced the definition of quantitative content analysis in chapter 1 (this volume), it was noted that if the categories and rules are conceptually and theoretically sound and are reliably applied, the chance increases that the study results will be valid. The focus of chapter 6 (this volume) on reliability leads easily to one of the possible correlates of reliable measurement: valid measurement.

What does the term *valid* mean? "I'd say that's a valid point," one person might respond to an argument offered by another. In this everyday context, validity can relate in at least two ways to the process by which one knows things with some degree of certainty.

First, valid can mean the speaker's argument refers to some *fact or evidence,* for example, that the national debt in 2012 topped $15 trillion. A reference to some fact suggests, of course, that the fact is part of objective reality. Second, valid can mean the speaker's *logic* is persuasive, because observation of facts leads to similar plausible inferences or deductions from them.

The social science notion of validity relates more rigorously to both these everyday ways in which we make inferences and interpretations of our reality. Social science does this in two ways. First, social science breaks up reality into conceptually distinct parts that we believe actually exist, and that have observable indicators of their existence. And second, social science operates with logic and properly collected observations to connect those concepts in ways that helps us to predict, explain and potentially control that reality.

Content analysis, then, must also incorporate these two processes in the way this method illuminates reality. First, we must address how a concept we have defined about some part of communication reality actually exists in that reality. And even if this is true, we must address how our category measurement of that communication concept is a good one. If we are mistaken about that

communication part, or if we are measuring it wrongly, then our predictions about the communication process fail.

But even with good concepts and measures of communication reality, we then have the second problem of validly linking those concepts through data collection and analysis methods that have the highest chance of producing successful predictions. So our second validity problem focuses on these "linking processes" that tie together our concepts, our measurements of them, our observations of their interconnections, and our predictions about their future states.

Although all that may sound formidable enough, philosophers of science urge humility even when our successful predictions suggest we have a good handle on understanding and measuring reality. Bertrand Russell illustrated this with a homely story of the chicken that day after day was well fed and watered and cared for in every way by a farmer. That chicken had a very solid record of very good predictions about interconnected events, suggesting a very good understanding of reality. Until, of course, the day the farmer showed up with his axe. That chicken did not even know, much less understand, the larger context in which it was a part.

In the sections of this chapter that follow, we first deal with the validity of the concepts in our theories and then with the validity of the observational processes we use to link those concepts. Finally, taking a lesson from the experience of Bertrand Russell's chicken, we address a wider context that we call "social validity" in which we ask how a scientifically valid content analysis relates to the wider communication world experienced by people.

The Problem of Measurement Reliability and Validity

Content analysis studies the reality of communication in our world. It does so through the creation of reliable and valid categories making up the variables we describe and relate to one another in hypotheses or models of the communication process. As we've emphasized in earlier chapters, we must operationally define content categories for the terms in these hypotheses and questions.

For example, here is a hypothesis from a study that related circulation of a newspaper to measures of its quality (Lacy & Fico, 1991).

The higher the circulation of a newspaper, the better the quality of that newspaper.

Now while newspaper circulation size may seem nonproblematic, the word *quality* should certainly give us pause. Just what on earth can quality be in a newspaper? The observational measures Lacy and Fico (1991) used to assess newspapers included the ratio of news-hole to advertising, the number of news services carried, and the amount of local coverage. However, who says measures such as those are good measures of quality? Is quality, like beauty, in the eye of the beholder?

The answer to that second question is, of course, yes: quality is often in the eye (or mind) of the beholder. That question also nicely illustrates the validity

problem in content analysis measurement. Communication is not simply about the occurrence or frequency of communication elements. It is also about the meanings of all the words, expressions, gestures, etc. that we use to communicate in life. So when we ask about the content analysis validity of our measurement of something like newspaper "quality," we frequently have an operational definition that has reduced ambiguity in the measurement of communication reality rather than genuinely apprehended that reality. And we ought not to assume that such ambiguities can always be resolved.

This becomes a critical problem because of our efforts to achieve reliability in our content category measures. Measurement reliability is a necessary but not sufficient condition for measurement validity. However, a measure can be reliable in its application but wrong in what researchers assume it is really measuring. A valid measure must be both reliable in its application and valid for what it measures.

A special problem in content analysis may occur because reliable measurement can come at the expense of valid measurement. Specifically, to get high levels of coder agreement on the existence or state of a content variable, the operational definition may have only tenuous connection with the concept of interest. Much of the concern with computer content analysis (dealt with in chapter 9 [this volume]) is that the validity of concepts is compromised by the focus on single words absent any context that gives them meaning. Part of the solution to this problem is multiple measures of the concept of interest, as used in the example study of newspaper quality referenced earlier. But ultimately, content analysts must ask the most consequential question of their measures: do they validly assess something meaningful beyond their utility in the particular study? Sometimes the validity of particular measures used in a content study may have been assessed in the broader research stream of which the particular study is a part. Too often, however, research will discuss the reliability of measures at length and ignore or assume the validity of those measures.

Tests of Measurement Validity

Analysts such as Holsti (1969) and Krippendorff (2004a) have discussed validity assessment at length. In particular, Holsti's familiar typology identifies four tests of measurement validity—face, concurrent, predictive, and construct—that apply to the operational terms we use in our hypotheses and questions.

Face Validity

The most common validity test used in content analysis and certainly the minimum one required is face validity. Basically, the researcher makes a persuasive argument that a particular measure of a concept makes sense on its face. Examining civil rights era letters to the editors of Southern newspapers for changing references to literacy tests, poll taxes, integration, and states' rights might, on the face of it, index the changing focus and nature of public debate. In essence, the

researcher assumes that the adequacy of a measure is obvious to all and requires little additional explanation. Relying on face validity sometimes can be appropriate when agreement on a measure is high among relevant researchers.

However, assuming face validity of measures is sometimes chancy, especially in broader contexts. Certainly we'd expect problems with face validity in studies that crossed cultures or language groups because meanings can vary so drastically. But problems can occur even where one wouldn't expect to find them. One of the authors of this text participated in a study assessing fairness and balance in reporting a state political race (Fico & Cote, 1997). The measure of such fairness and balance—equivalent treatment of the candidates in terms of story space and prominence—were consistent with invocations for "balance" and "nonpartisanship" in codes of ethics. These measures were subsequently discussed with seven reporters who wrote the largest number of campaign stories. None agreed with the research definition. That is not to say that either the research definition or the professional definition was wrong per se. However, what seems obvious on its face sometimes is not.

Concurrent Validity

Even face validity, however, can be strengthened for purposes of inference. One of the best techniques is to correlate the measure used in one study with a similar one used in another study. In effect, the two methods can provide mutual or *concurrent validation*.

In the study of newspaper quality and circulation mentioned previously (Lacy & Fico, 1991), a number of different measures were operationalized into an overall indicator. Those included amount of news hole devoted to local news, number of wire services carried, and proportion of news to advertising copy. These measures were validated by a previous survey of some 700 editors who answered questions about quality in journalism and who rated a number of indicators subsequently used in the content study. Presumably, the researchers reasoned, editors were in a very good position to recognize quality when they see it. Therefore, in addition to the face validity of each measure of the index, the research incorporated a cross-check with a relevant (and large) sample of people who could be considered experts on journalism quality.

Predictive Validity

A test of *predictive validity* correlates a measure with some predicted outcome. If the outcome occurs as expected, our confidence in the validity of the measure is increased. More specifically, if a hypothesized prediction is borne out, then our confidence in the validity of the measures making up the operational definitions of concepts in the hypothesis is strengthened. The classic example cited by Holsti (1969, p. 144) concerns a study of suicide notes left in real suicides and a companion sample from nonsuicides. Real notes were used to put together a linguistic model predicting suicide. Based on this model, coders successfully classified notes from real suicides, thereby validating the predictive power of the content model. In the newspaper quality study just cited (Lacy & Fico, 1991), theory predicting circulation based on news quality was consistent with empirical results.

Construct Validity

Construct validity involves the relation of an abstract concept to the observable measures that presumably indicate the concept's existence and change. The underlying notion is that a construct exists but is not directly observable except through one or more measures. Therefore, some change in the underlying abstract concept will cause observable change in the measures. Statistical tests of construct validity assess whether the measures relate only to that concept and to no other concept (Hunter & Gerbing, 1982). If construct validity of measures exists, then any change in the measures and the relation of the measures to one another is entirely a function of their relation to the underlying concept. If construct validity does not exist, then measures may change because of their relation to some other, unknown concepts. In other words, construct validity enables the researcher to be confident that when the measures vary, only the concept of interest is actually varying.

Put another way, the issue of construct validity involves whether measures "behave" as theory predicts and only as theory predicts. Wimmer and Dominick (2003) wrote that "construct validity is present if the measurement instrument under consideration does not relate to other variables when there is no theoretic reason to expect such a relationship. Therefore, if an investigator finds a relationship between a measure and other variables that is predicted by theory and fails to find other relationships that are not predicted by theory, there is evidence for construct validity" (p. 60).

Construct validity must exist if a research program in a field such as mass communication is to build a cumulative body of scientific knowledge across a multitude of studies. Common constructs used across studies help bring coherence and a common focus to a body of research. Valid constructs also make for more efficient research, enabling researchers to take the next step in extending or applying theory without needing to duplicate earlier work. The newspaper quality study, for instance, uses the "financial commitment" (Lacy, 1992) construct, which in turn is related to broader economic theory. Few studies in our field, however, validate measures this way.

Validity in Observational Process

Given that we have enough confidence in the validity of our concept measures to use them to address hypotheses and research questions, the question then becomes how we link them in a way that validly describes social reality. Every

social science method has a set of procedures meant to ensure that observations are made that minimize human biases in the way such reality is perceived. In survey research, such procedures include random sampling to make valid inferences from characteristics in a sample to characteristics in a population. In an experiment, such procedures include randomly assigning subjects to create equivalent control and experimental treatment groups, thereby permitting a logical inference that only the treatment variable administered to the experimental group could have caused some effect. In our previous chapter on reliability and in the section before this one, we discussed how content analysis uses protocol definitions and tests for chance agreement to minimize the influence of human biases and chance on coding defined and measured variables. Valid application of these procedures strengthens one's confidence in what the survey, experiment, or content analysis has found. But given science's largest goal, the prediction, explanation, and potential control of phenomena, how does content analysis achieve validity?

Internal and External Validity

Experimental method provides some ways of thinking about validity in research process that can be related to content analysis. In their assessment of experimental method in educational research, Campbell and Stanley (1963) made the distinction between an experimental research design's internal and external validity. By internal validity, Campbell and Stanley meant the ability of an experiment to illuminate valid causal relationships. An experiment does this by the use of controls to rule out other possible sources of influence, the rival explanations we mentioned in chapter 3 (this volume). By external validity, Campbell and Stanley meant the broader relevance of an experiment's findings to the vastly more complex and dynamic pattern of causal relations in the world. An experiment may increase external validity by incorporating "naturalistic settings" into the design. This permits assessment of whether causal relations observed in the laboratory are in fact of much importance relative to other influences operating in the world.

These notions of internal and external validity in experimental design also are useful for thinking about content analysis validity. A first and obvious observation is that content analysis, used alone, cannot possess internal, causal validity in the sense that Campbell and Stanley (1963) used because it cannot control all known and unknown "third variables." Recall from chapter 3 (this volume) that inferring causal relations requires knowledge of the time order in which cause and effect operate, knowledge of their joint variation, control over the influence of other variables, and a rationale explaining the presumed cause-effect relationship.

However, content analysis can incorporate other research procedures that strengthen the ability to make such causal inferences. For instance, if some content is thought to produce a particular effect on audiences, content analysis could be paired with survey research designs to explore that relationship, as in the

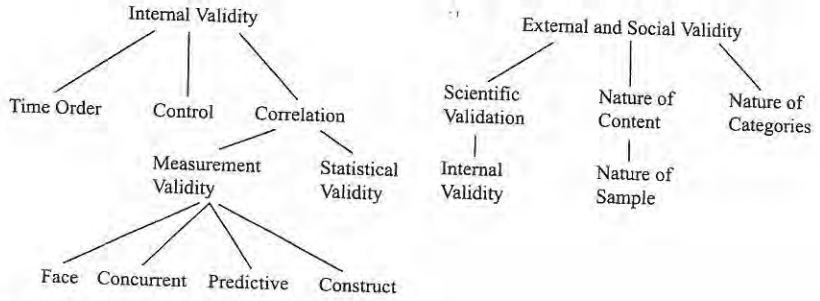

FIGURE 7.1 Types of content analysis validity.

agenda-setting and cultivation studies we described in chapters 1 and 3 (this volume). We discuss some of these designs in more detail in the following sections.

Content analysis can be a very strong research technique in terms of the external validity or generalizability of research that Campbell and Stanley (1963) discussed. This, of course, will depend on such factors as whether a census or appropriate sample of the content has been collected. However, the notion of external validity can also be related to what we call a study's social validity. This social validity will depend on the social significance of the content that content analysis can explore and the degree to which the content analysis categories created by researchers have relevance and meaning beyond an academic audience. We explore some of these issues in the following pages.

Figure 7.1 summarizes several types of validity. Note first that internal validity deals with the design governing data collection and how designs may strengthen causal inference. Data collection also requires assessment of measurement validity consisting of face, concurrent, predictive, and construct validity. Statistical validity is a subset of internal validity that deals with measurement decisions and the assumptions about data required by particular statistical analyses. Finally, the external and social validity of a content analysis presupposes the internal validity of measurement and design that makes content analysis a part of scientific method. However, the notion of external and social validity used here goes beyond those qualities to assess the social importance and meaning of the content being explored. The overall validity of a study therefore depends on a number of interrelated factors we discuss in the following section.

Internal Validity and Design

Content analysis by itself can best illuminate patterns, regularities, or variable relations in some content of interest. Content analysis cannot alone establish the antecedent causes producing those patterns in the content, nor can it explain as causal the subsequent effects that content produces in some social system. Of

course, the analyst may make logical inferences to antecedent causes or subsequent effects of such content, as we discussed in chapter 1 (this volume), with the model showing the centrality of content to communication processes and effects. Also, certain research designs pairing content analysis with other methods strengthen the ability to infer such causal relationships, thereby enhancing internal validity. In one way or another, then, content analysis designs must address issues of control, time order, and correlation of variables included in a communication model.

Control in Content Analysis

Designs that attempt to explain patterns of content must look to information outside the content of interest. This requires a theoretical or hypothesized model including the kinds of factors that may influence content. In other words, this model is assumed to control for other sources of influence by bringing them into the analysis. The model itself is derived from theory, previous research, or hunch. Consider a simple example. A researcher, noting the collapse of communist regimes, predicts rapid growth of new newspaper ventures as alternative political voices seek audiences, even as existing newspapers "open up" to previously taboo topics.

Two problems need to be emphasized. First, however plausible a model may be, there is always something—usually a lot of things—that is left out and that may be crucial. However, the second problem is what makes the first one so interesting: Such a model can always be empirically tested to assess how well it works to explain patterns in content. If the model does not work as planned, interesting and engaging work to figure out a better model can be undertaken. Unimportant variables can be dropped from the model and other theoretically interesting ones added. In this example, failure to find new newspapers might simply reflect limited access to printing facilities rather than a lack of dissent. Also, failure to find criticism of former party programs in existing papers may simply indicate residual citizen distrust of journalists who were viewed for years as political party tools.

Time Order in Content Analysis

Furthermore, in designing such a model, these presumed influences must incorporate the time element into the design, as we noted in chapter 3 (this volume). Such incorporation may be empirical—data on the presumed cause occurs and is collected and measured before the content it presumably influences. For example, studies have looked at the effects on a newspaper's community coverage after the loss of a city's competing daily or after the takeover of a daily by a group.

Incorporation of the time element may also be assumed from the logic of the design. For example, circulation size and population in a city at one point in time may be measured to help predict and explain news hole at a subsequent point in time. Clearly, the logic here would rule out news hole influencing population or, certainly not in the short run, circulation. Similarly, Atwater, Fico, and Pizante (1987) incorporated time into a design which illuminated how newspapers, local broadcast outlets, and the Associated Press set one another's state legislative news agenda across a 2-week period.

Obviously, exploring the effects of content is the converse of the situation just discussed. Here, time also must be part of the design, but other methods to assess effect are mandatory as well. As in the case of exploring antecedents of content, the logic of the design may be sufficient for controlling time order. For example, Lacy and Fico (1991) explored the effect of newspaper quality at one point in time on subsequent circulation. Quality was assessed from measures of newspaper content in a national sample of newspapers. Circulation was obtained from national audit figures. Clearly circulation at a subsequent time could not possibly influence the measures of quality at an earlier time.

Perhaps the most frequent multi-method example of content analysis research that assesses effect is the agenda-setting research we described in chapter 1 (this volume). This line of research explores whether differences in news media coverage frequency of various topics at one point in time creates a similar subsequent ordering of importance among news consumers. Of course, the possibilities that the news priorities of consumers really influence the media or that both influence one another must be taken into account in the design.

A technique called *cross-lag correlation* has often been used to do this. In this technique, both content analysis and survey research are performed at two different points in time. At each point, the media agenda is assessed through content analysis, and the public agenda is measured through a survey. The cross-lag correlation is then computed between the Time 1 media agenda and Time 2 public agenda and between the Time 1 public agenda and the Time 2 media agenda. If the correlation between the Time 1 media agenda and Time 2 public agenda is stronger than the correlation between the Time 1 public agenda and Time 2 media agenda, that must logically be the result of media's having an effect on the public rather than the reverse. Obviously, cross-lagged correlation has methodological limits including other potential variables that may be having an effect but are left out of the analysis. Still, the technique allows for establishing correlation between two variables while controlling for time order in a nonexperimental design.

Correlation in Content Analysis

As Figure 7.1 showed, the validity of content analysis research can be assessed in terms of requirements for causation we introduced in chapter 3 (this volume): specification of time order, control, and demonstration of joint variation or correlation. This last requirement brings our discussion to the special issue of statistical validity.

Given that we have brought variables into a content analysis research design, proper statistical use of them is necessary. First, our assumptions about independent and dependent variable influences must be explicit in any kind of multivariate analysis. Further, the analysis must consider direct and indirect causal flows, and whether any variables moderate or control the nature of the model's relationships.

We discuss statistics used for analyzing content data in chapter 8 (this volume). These techniques range from simple correlation measures for relating two variables, to multivariate techniques enabling the analysis to more fully control and assess the effects of multiple variables. Different statistics have different assumptions that must be considered. The specific techniques that can be employed will also depend on the level at which variables have been measured. Furthermore, if content data have been randomly sampled, tests of statistical significance must be employed for valid inferences to content populations. These issues relate to the statistical validity of the analysis of content.

External Validity and Meaning in Content Analysis

A study may have strong internal validity in the senses discussed earlier. But study findings may be so circumscribed by theoretical or methodological considerations that they have little or no relevance beyond the research community to which it is meaningful. Certainly any research should first be a collective communication among a group of researchers. Isaac Newton summed this up in his often-quoted saying, "If I have seen farther it is by standing on the shoulders of giants" (Oxford University, 1979, p. 362). The researcher interacts professionally within a community of scientists. But the researcher is also part of the larger society, interacting with it in a variety of roles such as parent, neighbor, or citizen.

In these broader researcher roles, the notion of validity can also have a social dimension that relates to how such knowledge is understood, valued, or used. In the hypothetical conversation discussed at the start of this chapter, two persons communicate knowledge that is meaningful to both. This meaningfulness results from a common language; a common frame of reference for interpreting the concepts being communicated; and a common evaluation of the relevance, importance, or significance of those concepts. In this social dimension of validity as meaning, the broader importance or significance of what has been found can be assessed.

Earlier in chapter 3 (this volume) on research design, we discussed the Pasteur Quadrant that is ideal for research. In that quadrant, research fits into and moves forward the body of scientific knowledge that may have interest and meaning to only a small body of similar researchers. But ideally, research on a topic engaging to a small body of similar researchers may also engage much larger audiences in much more important ways.

External Validity and the Scientific Community

But first, research must be placed before the scientific community for their assessment of a study's meaningfulness as valid scientific knowledge. The minimum required is validation of the research from the peer-review process in which competent judges assess a study's fitness to be published as part of scientific knowledge. Specifically, scientific peers must agree that a particular study should be published in a peer-reviewed venue.

As in scientific method generally, the peer-review process is meant to minimize human biases in the assessment of studies. In this process, judges unknown to the author review the work of the author who is unknown to them. The judges, usually two or three, apply scientific criteria for validating the research's relevance, design and method, analysis, and inference. The requirements for the scientific validation of research are relatively straightforward. Presumably, the current research demonstrably grows out of previous work, and the researcher explicitly calls attention to its relevance for developing or modifying theory, replicating findings, extending the research line and filling research gaps, or resolving contradictions in previous studies. Only after this process is the research deemed fit to be presented or published as part of scientific knowledge. Earlier in this volume, an author said he always asked applicants looking for academic jobs about their dependent variable. His second question was, "What journals will you publish in?"

Researchers submit their final work to a peer-review process for several reasons. First, researchers are given comments and criticisms for improving the study. No research is without flaws and limitations, and other experts in the field can help to illuminate them so they can be corrected or taken into account. Second, researchers submit work to the peer-review process so that the study might inform and assist other researchers in the collective building of a body of knowledge that advances the goals of science to predict, explain, and potentially control phenomenon.

The judgment of the scientific community provides the necessary link between the internal and external validity of research. Clearly, research that is flawed because of some aspect of design or measurement cannot be trusted to generate new knowledge. The scientific validation of research is necessary before that research can (or should!) have any broader meaning or importance. In essence, internal validity (the study is deemed fit as scientific knowledge) is a necessary (but not necessarily sufficient) condition for external validity (the study has wider implications for part or the whole of society).

However, the status of any one study as part of scientific knowledge is still tentative until other research provides additional validation through study replication and extension. This validation can take place through direct replication and extension in similar studies as in the example of agenda-setting research. Replication of findings by other studies is important because any one study, even given

the most critical scrutiny, may through chance sampling error produce an atypical result. However, if study after study finds similar patterns in replications, the entire weight of the research as a whole strengthens one's confidence in the knowledge that has been found. Recall in chapter 5 (this volume) that it was noted that even data sets drawn consistently from non-probability samples can be useful if their findings make a cumulative contribution. Scientific community validation of a study can also happen through the use, modification, or further development of that study's definitions or measures or through more extensive work into an area to which some study has drawn attention. The attention to media agenda setting across multiple decades now is an example of collective validation from multiple studies by multiple researchers.

External Validity as Social Validity in Content Analysis

The validation of research method and inference is usually determined by the scientific community acting through the peer review and replication process just discussed. This validation is necessary but not sufficient to establish the broader meaning and importance of research to audiences beyond the scientific community.

The external validity of a content analysis beyond the scientific community is strengthened in two ways that place a content analysis in the Pasteur Quadrant in which social validity is also maximized. These concern the social importance of the content and how it has been collected and the social relevance of the content categories and the way they have been measured and analyzed. In the following sections, we address these issues in the social validity of content studies.

Nature of the Content

The social validity of a content analysis can be increased if the content being explored is important. The more pervasive and important the content of interest to audiences, the greater will be the social validity of the analysis exploring that content. One dimension concerns the sheer size of the audience exposed to the content. Much of the research and social attention to the Internet, for example, emerges from the fact that its content is readily available and that large numbers of people use its content for many hours on a daily basis.

Another dimension of the importance of the content being analyzed deals with the exposure of some critical audience to its influence. Children's television advertising is explored because it may have important implications for the social development of a presumably vulnerable and impressionable population.

Finally, content may be important because of some crucial role or function it plays in society. For example, advertising is thought to be crucial to the economic functioning of market societies. Obviously, the effectiveness of advertising in motivating consumers to buy products will affect not only producers and consumers but also the entire fabric of social relations linked in the market. Furthermore, advertising messages can have cultural by-products because of social roles or stereotypes they communicate. Similarly, news coverage of political controversy is examined because it may influence public policy affecting millions. The political ethic of most Western societies is that an informed citizenry, acting through democratic institutions, determines policy choices. Clearly, then, the way these choices are presented has the potential to influence the agendas and opinions of these citizens.

Whatever the importance of the content, the social validity of the analysis will also be affected by how that content has been gathered and analyzed for study. Specifically, whether content has been selected through a census or a probability sample will influence what generalizations can be validly made.

A major goal in most research is knowledge about populations of people or documents. Knowledge of an unrepresentative sample of content is frequently of limited value for knowing or understanding the population. Probability sampling, however, enables researchers to generalize to the population from which the sample was drawn. Taking a random sample or even a census of the relevant population of content obviously enables the researcher to speak with authority about the characteristics being measured in that population.

Findings from content selected purposively or because of convenience cannot be generalized to wider populations. However, a strong case for the social validity of purposively selected content may be made in specific contexts. For example, the news content of the "prestige" press is clearly atypical of news coverage in general. However, those newspapers influence important policymakers and other news outlets as well and therefore have importance because they are so atypical.

Nature of the Categories

Whatever the size or importance of the audience for some communication, content analysis creates categories for the study of the content of that communication. These categories serve three purposes. First, they are created because the researcher believes they describe important characteristics of the communication. Second, they are created because the researcher believes these communication characteristics are themselves systematically produced by other factors that can be illuminated. And third, they are created because the researcher believes these categories have some kind of meaning or effect for the audiences experiencing them in a communication.

The conceptual and operational definitions of a content category that are relevant for these last two reasons are therefore also relevant for a study's social validity. Such concepts and their operational definitions may be interpretable by only a small body of researchers working in some field, or they may be accessible and relevant to far broader audiences. Krippendorff's (1980) "semantical validity" (p. 157) relates to this notion of relevance in content analysis. Krippendorff

asserts that semantical validity "assesses the degree to which a method is sensitive to the symbolic meanings that are relevant within a given context" (p. 157). In particular, Krippendorff considered a study to be high in semantical validity when the "data language corresponds to that of the source, the receiver or any other context" (p. 157). To what extent, therefore, do content analysis categories have corresponding meanings to audiences beyond the researchers? This question is particularly important when a researcher explicitly attempts to locate a content analysis in the "Pasteur Quadrant" that deals with both the theoretical and practical relevance of the research.

This question is also critical when a researcher chooses to focus on either manifest or latent content in a content analysis. Manifest content, as we discussed earlier in this text, is relatively more easily recognized and counted than latent content. Person A's name is in a story, maybe accompanied by a picture; a television show runs *X* number of commercials; a novel's average sentence length is *X* number of words. Analyses that attempt to capture latent content deal with more holistic or "gestalt" judgments, evaluations, and interpretations of content and its context. Studies attempting to analyze latent content assume that the most important characteristics of communication may not be captured through sampling, category definition, reliability assessment, or statistical analysis of the collected content data. Instead, the proper judgment, evaluation, or interpretation of communication content rests with the researcher.

This assumption about the ability of the researcher to do these things is troublesome on several grounds. In particular, seldom or ever is it argued explicitly in the analysis of latent content that the researcher's experience, intuition, judgment, or whatever, is actually competent to make those judgments. We must simply believe that the meaning of content is illuminated by the discernment of the researcher who brings the appropriate context to the communication. In other words, the researcher analyzing latent content knows what that content in a communication "actually is" and what that content "actually does" to audiences getting that communication.

A study of latent content must assume, therefore, that the researcher possesses one or both of two different, even contradictory, qualities that displace explicit assessments of the reliability and validity of content studies.

The first is that the researcher is an authoritative interpreter who can intuitively identify and assess the meaning embedded in some communication sent to audiences. The researcher is, therefore, the source of the study's reliability and validity of measurement. But this requires a large leap of faith in researchers. Specifically, we must believe that while human biases in selective exposure, perception, and recall exist in the naive perceiver of some communication, the researcher is somehow so immune that he or she can perceive the "real" content. For example, interpretations of the media's power to control the social construction of reality emerge from the assumed ability of the researcher—but not of media audience members—to stand enough apart from such an effect to observe and recognize it. But if the researcher can observe it, couldn't others? And if others can't observe it, then why can the researcher?

A second, but contrary quality assumed for the researcher in the analysis of latent content, is that the researcher is a kind of representative of the audience getting a communication. In other words, we must believe that the researcher is a "random sample" (with an *N* of 1) who "knows" the content's effects on audiences because he or she experiences and identifies them. Of course, few would trust the precision or generality of even a well selected random sample with an *N* of 1. Are we to believe that a probably very atypical member of the audience—the researcher—can experience the content in the same way that other audience members would?

It should be noted that these problems also may exist in the quantitative analysis of manifest content. For example, Austin, Pinkleton, Hust, and Coral-Reaume Miller (2007) found large differences in the frequency with which trained coders and a group of untrained audience members would assign content to categories. Although content analysis standards are satisfied by appropriate reliability tests, the social validity of a study may be limited if the content categories have little or no meaning to some broader audience. And no solution to the problem of inferring the effects of content on audiences exists unless content analysis is paired with techniques such as audience surveys or experiments that can better illuminate such effects. The claims made in the content analysis of manifest content should, therefore, always be tentative and qualified. But that's the nature of the entire scientific enterprise. In other words, we argue here that quantitative content analysis is necessary, even if sometimes not sufficient, for the development of a science of human communication.

Summary

The assumption of this text is that scientific method enables research to speak as truthfully as possible to as many as possible. Accomplishing this is the essence of the validity in content analysis as well as in other research.

Indeed, the parallels and perils of establishing validity for other research techniques are as serious. Survey researchers ask questions that assume implicitly a context and frame of reference that interviewees may not share. Experimenters achieve strong causal inference at the cost of so isolating phenomena that little may actually be learned about the broader world of interest.

However, validity as truth is what enquiry is all about. When it comes to illuminating truth about content, quantitative content analysis is the best way we have. We should accept nothing less in our communication scholarship.