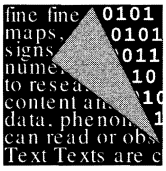


CHAPTER 11

Reliability

This chapter discusses two general purposes of reliability in scientific research. It distinguishes among three designs for generating data to measure reliability, which leads to three manifestations of reliability: stability, reproducibility, and accuracy. All turn out to be functions of the agreement achieved among observers, coders, judges, or measuring instruments. Krippendorff's agreement coefficient alpha is presented as a tool to assess such agreement, and its computation is demonstrated, starting with the simplest kind of data and moving to embrace the most common forms, nominal data, several metrics, multiple observers, and incomplete data. The chapter also discusses the statistical issues of sample sizes, alpha's distribution, and reliability standards.

WHY RELIABILITY? 11.1

Data, by definition, are the trusted ground for reasoning, discussion, or calculation. To stand on indisputable ground, content analysts must be confident that their data (a) have been generated with all conceivable precautions in place against known pollutants, distortions, and biases, intentional or accidental, and (b) mean the same thing for everyone who uses them. Reliability grounds this confidence empirically. There are two ways of operationalizing this confidence. In Kaplan and Goldsen's (1965) words: "The importance of reliability rests on the assurance it provides that data are obtained independent of the measuring event, instrument or person. Reliable data, by definition, are data that remain constant throughout variations in the measuring process" (pp. 83–84). Accordingly, a research procedure is reliable when it responds to the same phenomena in the same way regardless of the circumstances of its implementation. This is the measurement theory conception of reliability.

The other operationalization acknowledges that the phenomena of interest, which are encoded or inscribed in analyzable data, usually disappear right after they have been observed and recorded—human voices, historical events, radio transmissions, and even physical experiments. The analyst's ability to examine these phenomena in their absence, compare them with other phenomena, and, particularly, discuss them with members of a community of stakeholders relies heavily on a consensual reading and use of the data that represent, point to, or invoke experiences with the phenomena of interest. Empirical inquiries into bygone phenomena have no choice other than to presume that their data can be trusted to mean the same to all of their users. In content analysis, this means that the reading of textual data as well as of the research results is replicable elsewhere, that researchers demonstrably agree on what they are talking about. Here, then, reliability is the degree to which members of a designated community agree on the readings, interpretations, responses to, or uses of given texts or data. This is an interpretivist conception of reliability.

In either case, researchers need to demonstrate the trustworthiness of their data by measuring their reliability. If the results of reliability testing are compelling, researchers may proceed with the analysis of their data. If not, doubts as to what these data mean prevail, and their analysis is hard to justify.

To perform *reliability tests,* analysts require data in addition to the data whose reliability is in question. These are called *reliability data,* and analysts obtain them by duplicating their research efforts under various conditions—for example, by using several researchers with diverse personalities, by working in differing environments, or by relying on different but functionally equal measuring devices. Reliability is indicated by substantial agreement of results among these duplications.

In contrast to reliability, *validity* concerns truths. Researchers cannot ascertain validity through duplications. *Validity tests* pit the claims resulting from a research effort against evidence obtained independent of that effort. Thus, whereas reliability provides assurances that particular research results can be duplicated, that no (or only a negligible amount) of extraneous "noise" has entered the process and polluted the data or perturbed the research results, validity provides assurances that the claims emerging from the research are borne out in fact. Reliability is not concerned with the world outside of the research process. All it can do is assure researchers that their procedures can be trusted to have responded to real phenomena, without claiming knowledge of what these phenomena "really" are.

In content analysis, reliability and validity can be related by two propositions and a conjecture:

- *Unreliability limits the chance of validity.* In everyday life, disagreements among eyewitness accounts make it difficult for third parties to know what actually happened or whether the witnesses are reporting on the same event. For such accounts to be considered reliable, witnesses must concur well above chance. If the coding of textual matter is the product of chance,

it may well include a valid account of what was observed or read, but researchers would not be able to identify that account to a degree better than chance. Thus, the more unreliable a procedure, the less likely it is to result in data that lead to valid conclusions.

- *Reliability does not guarantee validity.* Two observers of the same event who hold the same conceptual system, prejudice, or interest may well agree on what they see but still be objectively wrong. Content analysts are not exempt from such concurrences. Because they have acquired a language and concepts that make them see the world from the unique perspective of their academic discipline, their observations and readings are based in a consensus that is not likely shared by many people outside of their scholarly community. Content analysts' shared worldview may deviate radically from the worldviews of those whose intentions, perceptions, and actions are at issue and could validate the intended inferences. A highly reliable research process may well be artificial and thus have little chance of being substantiated by evidence on the intentions, perceptions, actions, or events that were inferred. Even perfectly dependable mechanical instruments, such as computers, can be wrong—reliably. Thus a reliable process may or may not lead to valid outcomes.

This relationship is illustrated in Figure 11.1, which depicts reliability as repeating the same score and validity as being on-target. The top part of the figure suggests that with diminishing reliability, validity increasingly becomes a chance event. The bottom part suggests that reliability does not guarantee being on-target.

Thus reliability is a necessary, but not a sufficient, condition for validity. The following conjecture does not have the logical force of the preceding propositions, but it is born out of the experiences by numerous content analysts:

- *In the pursuit of high reliability, validity tends to get lost.* This statement describes the analyst's common dilemma of having to choose between interesting but nonreproducible interpretations that intelligent readers of texts may offer each other in conversations and oversimplified or superficial but reliable text analyses generated through the use of computers or carefully instructed human coders. Merritt's (1966) study of the rising national consciousness among the 13 original American colonies on the basis of newspaper accounts provides an example of a case in which complexity of interpretation was sacrificed for reliability. Because national sentiments are difficult to define and identify, Merritt elected to enumerate the mentions of American place-names instead. A shift in the use of the names of places in colonial England to the names of places in America may well be an attractive index, and counting words instead of themes causes fewer reliability problems, however, the use of place-names surely is only one manifestation of "national consciousness," and a richer account of this phenomenon could well have led to more interesting inferences. Merritt's index is attractive, as I have suggested, but its validity remains thin.

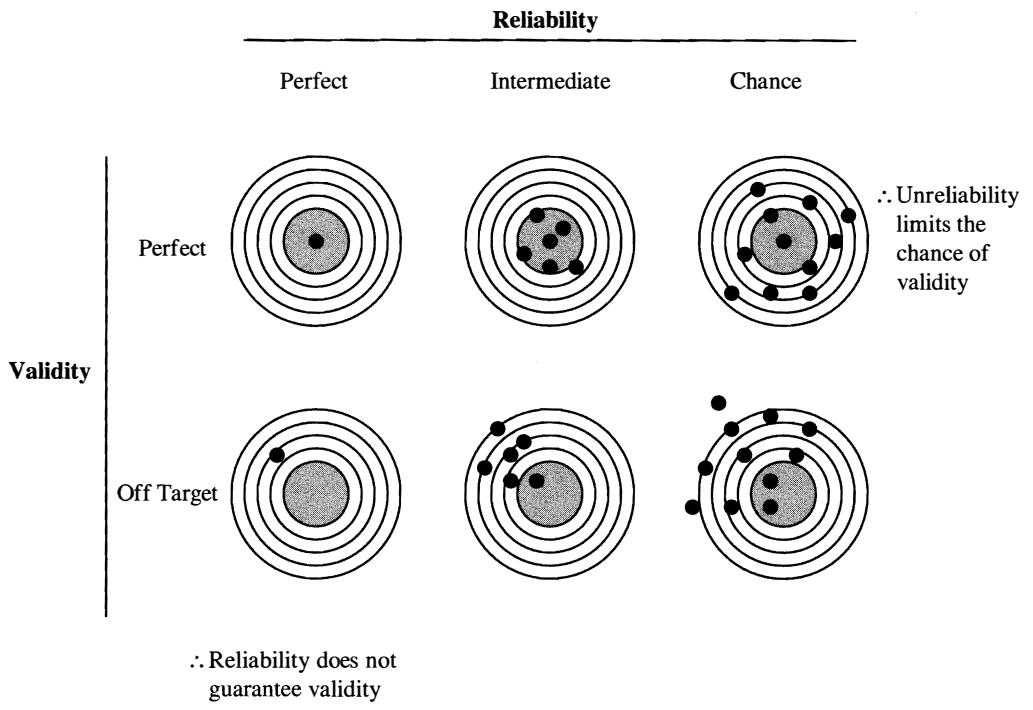


Figure 11.1 The Relationship Between Reliability and Validity

The use of computers in content analysis, praised for increasing reliability, has highlighted this dilemma even more clearly. Computers process character strings, not meanings. They sort volumes of words without making sense of them. Although it is possible to program computers to perform amazing functions, when analysts rely on them rather than on intelligent readers they run the risk of trivializing the meanings of texts (see Chapter 12, section 12.1). In content analysis, researchers should approach highly reliable procedures with as much caution as they approach fascinating interpretations that nobody can replicate.

11.2 RELIABILITY DESIGNS

11.2.1 Types of Reliability

There are three types of reliability: stability, reproducibility, and accuracy (see Table 11.1). These are distinguished not by how agreement is measured but by the way the reliability data are obtained. Without information about the

Table 11.1 Typese of Reliability

Reliability	Designs	Causes of Disagreements	Strength
Stability	test-retest	intraobserver inconsistencies	weakest
Reproducibility	test-test	intraobserver inconsistencies + interobserver disagreements	medium
Accuracy	test-standard	intraobserver inconsistencies, + interobserver disagreements, + deviations from a standard	strongest

circumstances under which the data for reliability assessments have been generated, agreement measures remain uninterpretable.

Stability is the degree to which a process is unchanging over time. It is measured as the extent to which a measuring or coding procedure yields the same results on repeated trials. The data for such assessments are created under *test-retest* conditions; that is, one observer rereads, recategorizes, or reanalyzes the same text, usually after some time has elapsed, or the same measuring device is repeatedly applied to one set of objects. Under test-retest conditions, unreliability is manifest in variations in the performance of an observer or measuring device. With reference to humans, such variations, also called *intraobserver disagreement* or *individual inconsistencies,* may be due to insecurity, carelessness, openness to distractions, difficulties in comprehending written instructions, or the tendency to relax performance standards when tired. Even the inherently human characteristic of learning through practice, creatively improving one's performance over time, shows up in disagreements over time. Stability, the weakest form of reliability, is insufficient as the sole criterion for accepting data as reliable. But because test-retest data are the easiest reliability data to obtain, and internal inconsistencies limit other reliabilities as well, measuring stability may be an analyst's first step in establishing the reliability of data.

Reproducibility is the degree to which a process can be replicated by different analysts working under varying conditions, at different locations, or using different but functionally equivalent measuring instruments. Demonstrating reproducibility requires reliability data that are obtained under *test-test* conditions; for example, two or more individuals, working independent of each other, apply the same recording instructions to the same units of analysis. Disagreements between these observers' performances are due to both intraobserver inconsistencies and *interobserver differences* in the interpretation and application of given recording instructions. Compared with stability, reproducibility, which is also variously called *intercoder reliability, intersubjective agreement,* and *parallel-forms reliability,* is a far stronger measure of reliability.

Accuracy is the degree to which a process conforms to its specifications and yields what it is designed to yield. To establish accuracy, analysts must obtain data under *test-standard* conditions; that is, they must compare the performance of one or more data-making procedures with the performance of a procedure that is taken to be correct. Observed disagreements between the two kinds of

performances are due to intraobserver inconsistencies, interobserver differences, and *deviations from a given standard.* Because it responds to all three sources of variation, accuracy is the strongest reliability test available. It is surpassed in strength only by validity measures that appear to conform to the test-standard design, but the standard is truth, or at least what is known to be true, a requirement that lies outside reliability considerations (and that I take up in Chapter 13).

When data making is merely clerical or computational, the meaning of accuracy is clear. Typos, for instance, are errors by comparison to existing spelling standards. In linguistics, conversation analysis, and therapeutic contexts, analysts perform accuracy checks by comparing novices' uses of transcription conventions, for example, with those of acknowledged experts. In content analysis, accuracy measurements often include testing the work of trainee coders against standards that have been established by panels of experienced content analysts. In the more problematic parts of content analysis, such as the interpretation and transcription of complex textual matter, suitable accuracy standards are not easy to find. Because interpretations can be compared only with other interpretations, attempts to measure accuracy presuppose the privileging of some interpretations over others, and this puts any claims regarding precision or accuracy on epistemologically shaky grounds. Thus the use of accuracy is limited to coder training and other areas where objective standards are readily available.

Stability, on the other end of the spectrum, is too weak to serve as a reliability measure in content analysis. It cannot respond to individually stable idiosyncrasies, prejudices, ideological commitments, closed-mindedness, or consistent misinterpretations of given coding instructions and texts.

One method that some scholars have mentioned as a reliability test is the split-half technique. This technique would call on content analysts to divide a sample of recording units into two approximately equal parts and have the two parts coded by different observers, one unit at a time. The analysts would then compare the frequency distributions obtained for the two parts. If the difference between the two distributions is statistically insignificant, the data would be considered reliable; otherwise, they would be considered unreliable. However, this measure merely assesses the degree to which two subsamples resemble each other or whether the larger data set can be considered homogeneous. (The test can also be used to determine whether a sample is large enough to represent a population; see Chapter 6, section 6.3.3.) However, as there may be good reasons why two subsamples are different, or no reasons why they should be the same, homogeneity says nothing about whether data can be trusted. In content analysis—or, more generally, when the reliability of categorizing or describing units is at issue—the split-half technique is not a suitable reliability test, and its use must be discouraged as uninformative and misleading.

11.2.2 Conditions for Generating Reliability Data

As noted in Chapter 2 (section 2.1), content analysis must be *reproducible,* at least in principle. To check on this possibility, analysts must generate reliability

data at least under test-test conditions and account not only for individual instabilities but also for disagreements among observers, coders, or analysts. Any analysis using observed agreement as a measure of reproducibility must meet the following requirements:

- It must employ communicable coding instructions—that is, an exhaustively formulated, clear, and workable data language plus step-by-step instructions on how to use it. This widely accepted requirement may need to be extended to include rarely mentioned training programs that coders typically undergo before qualifying for the task—otherwise, one may not know what the data mean and how to reproduce them.
- It must employ communicable criteria for the selection of individual observers, coders, or analysts from a population of equally capable individuals who are potentially available for training, instruction, and coding elsewhere.
- It must ensure that the observers who generate the reliability data work independent of each other. Only if such independence is assured can covert consensus be ruled out and the observed agreement be explained in terms of the given instructions and the phenomena observed or the texts interpreted.

Inasmuch as reliability serves as a condition for research to proceed with the data in hand, the content analysis literature is full of evidence of researchers' well-intended but often misguided attempts to manipulate the process of data generation so as to increase the appearance of high levels of agreement. Most of these involve violations of one or more of the above conditions, as the following examples illustrate.

In the belief that consensus is better than individual judgment, some researchers have asked observers to discuss what they read or see and reach their decisions by compromise or majority vote. This practice may indeed moderate the effects of individual idiosyncrasies and take advantage of the possibility that two observers can notice more than one, but data generated in this way neither ensure reproducibility nor reveal its extent. In groups like these, observers are known to negotiate and to yield to each other in tit-for-tat exchanges, with prestigious group members dominating the outcome. Here, observing and coding come to reflect the social structure of the group, which is nearly impossible to communicate to other researchers and replicate. Moreover, subjective feelings of accomplishment notwithstanding, the data that are generated by such consensual coding afford no reliability test. They are akin to data generated by a single observer. Reproducibility requires at least two independent observers. To substantiate the contention that coding by groups is superior to coding by separate individuals, a researcher would have to compare the data generated by at least two such groups and two individuals, each working independently.

It is not uncommon for researchers to ask observers to work separately, but to consult each other whenever unanticipated problems arise. Such consultation is a response to a common problem: The writers of the coding instructions have not

been able to anticipate all possible ways of expressing relevant matter. Ideally, these instructions should include every applicable rule on which agreement is being measured. However, the very act of observers' discussing emerging problems creates interpretations of the existing coding instructions to cope with the newly discovered problems that are typical of the group and not communicable to others. In addition, as the instructions become reinterpreted, the process loses some of its stability over time: Data generated early in the process use instructions that differ from those that evolve. The higher measure of reliability in the end is partly illusory.

Because content analysts may not be able to anticipate all of the possible complications in their texts, it is a common practice to expand the written coding instructions by adopting new and written rules as the process unfolds. The idea is that the coding instructions evolve, eventually requiring no new rules. To avoid being misled by unwritten consensus among coders engaged in the dual task of interpreting text and expanding the common instructions to do so, content analysts should put the final instructions to a reliability test, using different coders, and reexamine the data generated before these final instructions have been reached.

Content analysts are sometimes tempted to assume, and act on the assumption, that data making is served best by experts, exceptionally acute observers, or individuals who have long histories of involvement with the subject of the research. They should be reminded, however, that the requirement of reproducibility means that any individual with specifiable qualifications could perform the same coding tasks as well and know exactly what is meant by the categories, scales, and descriptive devices used in the research. If there are no other experts against whom the performance of the available expert observers can be checked, the observers' interpretations may be insightful and fascinating, but the analyst cannot claim that they are reliable. This is the basis of arguments against a content analyst's doing his or her own coding (as a principal investigator may sometimes do, for instance) unless the analyst's performance is compared with that of at least one other coder. To satisfy this requirement, the analyst may be tempted to find that other coder among friends or close associates with whom the analyst has worked for a long time. Two such coders are likely to agree—not, however, because they carefully follow the written instructions, but because they know each other and the purpose of the research; they are likely to react similarly without being able to convey the source of their convenient commonalities to others. Analysts should choose observers from a specifiable population of potential observers from which other researchers can select as well.

Sometimes content analysts accept as data only those units of analysis on which observers achieve perfect agreement. This is a particularly problematic practice, because it gives researchers the illusion of perfect reliability without affording them the possibility of separating agreement due to chance from agreement based on the sameness of reading or observation. For binary or dichotomous data, agreement by chance is at least 50%. Omitting units on which coders happen to disagree cannot change the chance nature of those on which they do agree. There is no escape from true chance events. This is true also when agreement is well above chance but not perfect. Units that are coded by chance

populate both agreement and disagreement cells of a coincidence table, and in the agreement cells there is no way of separating units according to whether observers agreed by chance or by following the instructions. Most important, when analysts rely on data that are easily coded, the data become an artifact of the analytical procedure—they are no longer representative of the phenomena the researchers hope to analyze.

Content analysts who employ the following two-step procedure can achieve both data whose reliability is measurable and an improvement in their confidence in the data beyond the measured reliability. First, they have to employ three or more observers working independent of one another. This yields reliability data whose reliability can be measured. Second, they reconcile discrepancies in these data either by relying on a formal decision rule—majority judgments or average scores—or by reaching consensus in postcoding deliberations. The data *before* such reconciliation are reliability data proper and do yield reportable reliabilities. Although it is reasonable to assume that postcoding reconciliation improves the reliability of the data beyond the reliability of data generated by any one individual observer, this is an assumption without measurable evidence. The only publishable reliability is the one measured before the reconciliation of disagreements. The reliability of the data after this reconciliation effort is merely arguable.

Reliability Data 11.2.3

As noted above, the data that enable researchers to assess reliability, called *reliability data,* duplicate the very data-making process whose reliability is in question. Reliability data make no reference to what the data are about. Only the assignment of units to the terms of a data language matters.

In their most basic or *canonical* form, reliability data consist of records generated by two or more observers or measuring devices and concern the same set of phenomena. In content analysis, two data-making processes are distinguishable: unitizing and coding. In the practice of research, unitizing and coding may occur together, but the mathematical processes used in evaluating their reliability are different, and so are their reliability data.

Unitizing is identifying within a medium—within an initially undifferentiated continuum—contiguous sections containing information relevant to a research question. These sections become the units of analysis or recording units, and sections between the identified units are left unattended. Examples of unitizing include clipping relevant articles from a newspaper, earmarking pertinent sections in a video recording for subsequent analysis, isolating conversational moves within a conversation, and identifying historically significant events in time. Reliability calculations for unitizing are not as transparent as those for coding and, unfortunately, are still uncommon. Later in this chapter (in section 11.6), I will state, but not develop, reliability measures for unitizing and refer interested readers to work published elsewhere (Krippendorff, 1995a, in press-a). The

reliability data for unitizing may be depicted as a three-dimensional cube of observers-by-continuum-by-categories of identified units, as in Figure 11.2.

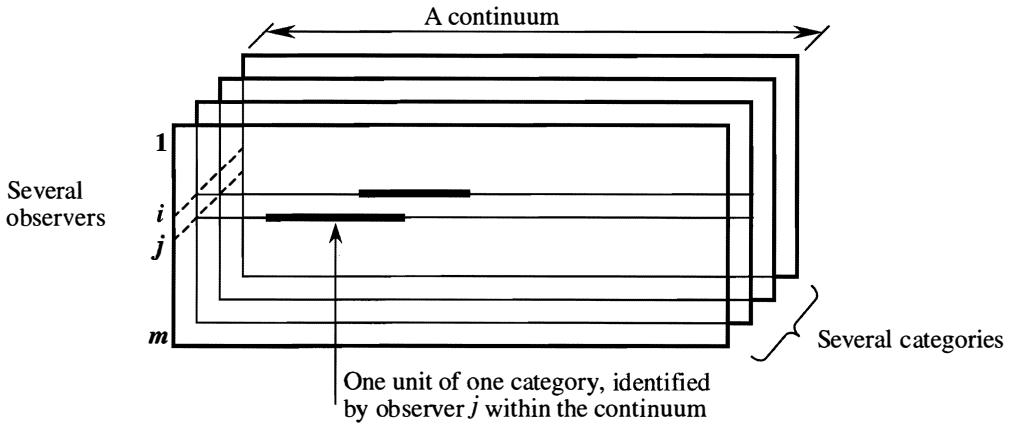


Figure 11.2 Reliability Data for Unitizing

Coding is the transcribing, recording, categorizing, or interpreting of given units of analysis into the terms of a data language so that they can be compared and analyzed. A distinction can be drawn between single-valued and multi-valued data. In single-valued data, each unit of analysis receives a unique description, one value from each variable. In multi-valued data, multiple descriptions or interpretations are allowed. As not too many statistics can handle the latter, I present reliability calculations of single-valued data only in this chapter. The data structure for coding single-valued data may be visualized as in Figure 11.3.

A comment on terminology: In discussing the generation of reliability data, I use the word *observers* in a general sense. They could be called coders, scorers, interpreters, unitizers, analysts, or judges. Outside of content analysis, they might be referred to as raters, interviewers, or acknowledged experts. Moreover, reliability data are not limited to those recorded by individual human beings. Business accounts, medical records, and court ledgers, for example, can be interpreted as the work of institutionalized observers. And mechanical measuring instruments—which convert phenomena into numbers—are included here as well.

In the process of unitizing a given continuum, observers characterize *units* by their length, duration, or size, and by their location in the continuum, using appropriate instructions. In coding, recording units are given or predefined, and the observers' efforts are directed toward their transcription, interpretation, or coding. In public opinion research, individuals often are the units of analysis; in educational research, units are often called (test) items, and elsewhere they may be known as cases. In content analysis, units may be single words or longer text segments, photographic images, minutes of video recordings, scenes in fictional television programs, Web pages, utterances, distinct experiences—anything that could have distinct meanings to an analyst.

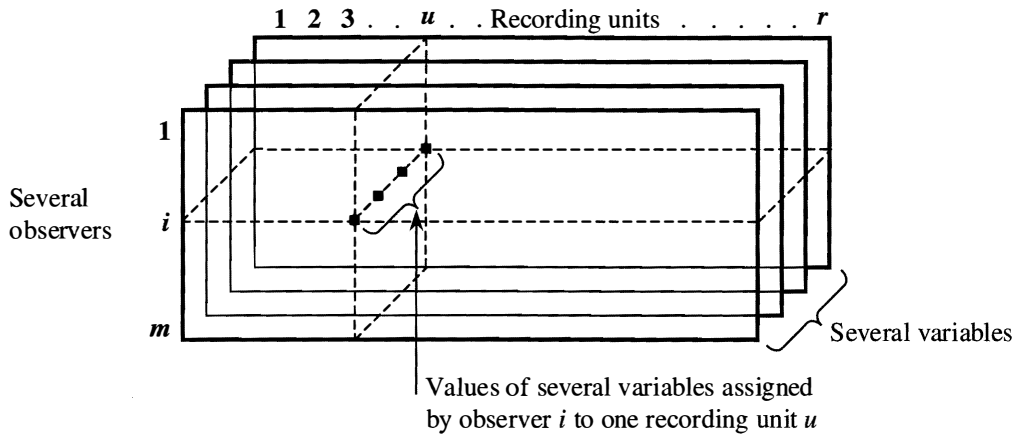


Figure 11.3 Reliability Data for Coding

In the context of coding, I use *values* as the generic term for names, categories, ranks, scores, scale points, measurements, answers to questions, even written text that describes itself and may be entered in the cells of a spreadsheet and located in the coordinates of rows and columns of reliability data. Values vary along dimensions called *variables.* As discussed in Chapter 8, a variable has mutually exclusive values.

Given these preliminaries, how can analysts assess agreement within reliability data as defined above?

α-AGREEMENT FOR CODING 11.3

Analysts always aim for the highest reliability achievable, of course. As perfect reliability may be difficult to achieve, especially when coding tasks are complex and so require elaborate cognitive processes, analysts need to know by how much the data deviate from the ideal of perfect reliability and whether this deviation is above or below accepted reliability standards. These are the two main questions that any agreement measure should answer.

Several coefficients for measuring agreement are available, specialized for particular kinds of data. Popping (1988) identifies 39 for nominal data, and his list is blatantly incomplete. In section 11.5, I will review a few popular ones regarding how they measure what their proponents claim they measure and explore their suitability for reliability assessments. Here, I am relying on Krippendorff's α, not because I invented it, but because it is the most general agreement measure with appropriate reliability interpretations in content analysis, as shall become clear (see Krippendorff, 1970a, 1970b, 1978, 1980b, 1987,

1992). α-agreement should not be confused with Cronbach's (1951) alpha, which is widely used in biometric and educational research for an entirely different purpose and is unsuitable for evaluating reliability in content analysis. Krippendorff's α allows uniform reliability standards to be applied to a great diversity of data:

- It is applicable to any number of values per variable. Its correction for chance makes α independent of this number.
- It is applicable to any number of observers, not just the traditional two.
- It is applicable to small and large sample sizes alike. It corrects itself for varying amounts of reliability data.
- It is applicable to several metrics (scales of measurements)—nominal, ordinal, interval, ratio, and more.
- It is applicable to data with missing values in which some observers do not attend to all recording units.

In its most general form, α is defined by

$$\alpha = 1 - \frac{D_o}{D_e},$$

where D_o is a measure of the observed disagreement and D_e is a measure of the disagreement that can be expected when chance prevails.

This definition reveals two reference points that are essential for any reliability measure: When agreement is observed to be perfect and disagreement is, therefore, absent, $D_o = 0$ and $\alpha = 1$, indicating perfect reliability. When agreement and disagreement are matters of chance and observed and expected disagreements are equal, $D_e = D_o$, $\alpha = 0$, indicating the absence of reliability. α could become negative, in fact, −1. However, as the aim is the highest reliability possible, negative values are too far removed from where differences in reliability matter. Negative values result from two kinds of errors: sampling errors and systematic disagreements. For reliability considerations, α 's limits are

$$1 \geq \alpha \geq 0 \begin{cases} \pm \text{ sampling error} \\ - \text{ systematic disagreement} \end{cases}$$

- *Sampling errors* happen for a variety of reasons; in this inequality, when sample sizes are too small. They manifest themselves when observations are few, each having large effects on α. These deviations from the true value of α occur when it turns out to be impossible for observed disagreements to equal the expected disagreements, causing α values to dance above and below zero.
- *Systematic disagreements* occur when observers agree to disagree or pursue opposing interpretations of the instructions given to them. All

observed disagreements distract from perfect reliability, but systematic disagreements can cause α values to drop below what would be expected by chance.

Foreshadowing the development of α, which I undertake in small steps, I offer a *first interpretation* of α, one of four I am proposing, in the form of this equation:

$$\alpha = 1 - \frac{D_o}{D_e} = 1 - \frac{\text{Average}_{\text{metric}}\, \delta^2_{ck} \text{ within all units}}{\text{Average}_{\text{metric}}\, \delta^2_{ck} \text{ within all data}},$$

where D_o is the average difference between values within units, regardless of who assigned them; D_e is the average difference between all values, regardless of who assigned them and to which units; and δ^2 is the squared difference between any two values c and k, a function of the applicable *metric*.

Inasmuch as differences within units count toward unreliability (are in error), and differences within all values include both the justified or true differences between units and the error differences within units, *α is the extent to which the proportion of the differences that are in error deviates from perfect agreement*, $\alpha = 1$ always being its largest value.

Before I get to the details of the agreement coefficient α, I should mention that I do assume that content analysts will use computer programs to analyze the reliabilities they need to know. However, to understand how α indicates agreement, it is important that content analysts be able to construct coincidence matrices by hand, calculate at least elementary α-measures, and thereby understand how observed disagreements affect reliability. In what follows, I introduce α, beginning with a simple example, and slowly add complications. Readers may find the next two sections to be a straightforward presentation of basic ideas, covering the kinds of data that are most common in content analysis. If these explanations suffice, they may want to skip to the discussion of standards in section 11.4.4 and, if still curious, the discussion of other coefficients in section 11.5.

Two Observers, Binary Data 11.3.1

The conceivably simplest reliability data are generated by two observers who assign one of two available values to each of a common set of units of analysis. For example, suppose a political scientist is interested in references to the United States in the Chinese press and, unable to read Chinese, hires two individuals (Jon and Han) who claim to be competent readers of Chinese to identify the presence or absence of such references. They mark each newspaper article "0" for absent or "1" for present. Their results may be tabulated as follows:

Article:	1	2	3	4	5	6	7	8	9	10
Jon's values:	1	1	0	0	0	0	0	0	0	0
Han's values:	0	1	1	0	0	1	0	1	0	0

This is called a reliability data matrix. It tells us that Jon and Han agree regarding 6 out of the 10 articles. Although 60% agreement sounds good, this figure does not tell us anything about the condition under which the assignment of values is the product of chance. In fact, 0% agreement is about as unlikely to achieve by chance as is 100%. Because 0% has no meaningful reliability interpretation, the %-agreement measure has no second reference point and does not constitute a reliability scale. When agreement deviates from 100%, as in this example, %-agreement is uninterpretable and can say nothing about the extent to which the data can be relied upon.

To gain a clearer view of these reliability data, one can create a matrix of the observed coincidences. Such a matrix accounts for all the values that the two observers used—20 in our example. Here, each unit contains two pairs of values, a Jon-Han pair and a Han-Jon pair. In the first unit, we find two pairs of values, 1-0 and 0-1. They do not match. In the second unit, the two pairs of values, 1-1 and 1-1, match perfectly and are in fact indistinguishable. Given that there are 10 units with two pairs of values in each, we need to tabulate 20 pairs, which equals the number of values in the above reliability data matrix. Let o_{ck} be the number of observed coincidences of the two values c and k. In our reliability data matrix we count ten 0-0 pairs, so, $o_{00} = 10$; four 0-1 pairs, $o_{01} = 4$; four 1-0 pairs, $o_{10} = 4$; and two 1-1 pairs, $o_{11} = 2$. A tabulation of these pairs yields the following matrix of observed coincidences:

Values:	0	1	
Matrix of observed coincidences 0	o_{00}	o_{01}	n_0
1	o_{10}	o_{11}	n_1
Number of values:	n_0	n_1	n

	0	1	
0	10	4	14
1	4	2	6
	14	6	20

Coincidence matrices should not be confused with the better-known *contingency matrices,* which are familiar in the tradition of assessing statistical associations or correlations, not agreements. Contingency matrices retain the identities of the two observers and treat them as independent variables. They tabulate the number of *units* being coded, not the number of *values* that participate in pair comparisons. There is a simple relationship between coincidence matrices and contingency matrices, however: Coincidences sum contingencies and their inverses, thereby omitting the references to the individual observers, considering them interchangeable. If x_{ck} is the number of times a particular observer uses c while the other uses k, then the number of coincidences $o_{ck} = x_{ck} + x_{kc}$. Accordingly, our 2-by-2-coincidence matrix can be seen as the sum of the contingencies in the Jon-Han matrix and in its inverse, the Han-Jon matrix:

Coincidence matrix				=	Contingency matrix + Inverse contingency matrix									
					Jon's values					Han's values				
Values:	0	1			0	1				0	1			
0	10	4	14		0	5	1	6		0	5	3	8	
1	4	2	6	=	Han's 1	3	1	4	+	Jon's 1	1	1	2	
Number of:	14	6	20 values			8	2	10 units			6	4	10 units	

Both kinds of matrices show mismatched pairs in their off-diagonal cells. However, coincidence matrices are symmetrical around the diagonal, whereas contingency matrices are not. The margins of coincidence matrices enumerate the *values* actually used by all observers. The margins of contingency matrices enumerate the *units* being recorded by each observer. The two margins of coincidence matrices are the same. The two margins of contingency matrices typically differ.

Disagreements in the matrix of observed coincidences determine the quantity of observed disagreement D_o, the numerator of α. We now need to consider what goes into the denominator of α, the expected disagreements D_e, which averages all differences between values that are pairable within the whole reliability data matrix, ignoring who assigned them to which units. The disagreements that determine the quantity of expected disagreement D_e are found in the matrix of the expected coincidences, which represents what could happen by chance. Ideally, and in our example, we would base such expectations on knowledge of the proportion of references to the United States that actually occur in the Chinese press. However, we cannot know this proportion without completing the very analysis whose reliability is in question. Lacking knowledge of this kind, we estimate the population of references in the Chinese press from what we do know, the proportion of references that all available observers jointly identify. With Jon finding U.S. references in 2 out of 10 articles and Han finding them in 4 out of 10, the two observers have jointly identified 6 out of 20, or 30%. This is our population estimate. It also happens to be, as it should, the proportion of ones in the original reliability data regardless of who contributed them.

Given these proportions, we now calculate how often one can expect Jon and Han to agree under conditions that they do not read anything at all and draw zeros and ones randomly out of a hat. Suppose we place 20 balls in that hat, 6 balls labeled "1" and 14 balls labeled "0," mix them thoroughly, and let two individuals draw balls out of the hat blindly. To draw two 1s in a row, the first individual can be expected to draw a 1 in 6 out of 20 cases. Having removed a 1 from the hat and thereby reduced the remaining number of 1s and the total by one, the second individual will draw a 1 in 5 out of 19 balls. To get to expected frequencies, just as in the matrix of observed coincidences, we multiply these two probabilities by the total number $n = 20$ and obtain the expected frequencies of 1-1 pairs as $(6/20)(5/19) \cdot 20 = 1.5789$. By contrast, if the first individual drew a 1 from the hat, the number of 0s stayed the same and the second individual would be expected to draw a 0 in 14 out of 19 cases. This gives us the following matrix of expected coincidences (the computational formulas are reproduced on the right of these matrices):

Values:	0	1	
Matrix of expected coincidences 0	e_{00}	e_{01}	n_0
1	e_{10}	e_{11}	n_1
Number of values:	n_0	n_1	n

	0	1	
0	9.6	4.4	14
1	4.4	1.6	6
	14	6	20

Where:

$e_{00} = n_0 \cdot (n_0 - 1)/(n - 1)$

$e_{01} = e_{10} = n_0 \cdot n_1/(n - 1)$

$e_{11} = n_1 \cdot (n_1 - 1)/(n - 1)$

We now have everything we need to calculate the agreement for binary reliability data. What counts as disagreement are the off-diagonal entries of the two matrices. Because coincidence matrices are symmetrical, $o_{01} = o_{10} = 4$ and $e_{01} = e_{10} = 4.4211$, we do not need to add the contents of both off-diagonal cells and can express α for binary data by

$$_{\text{binary}}\alpha = 1 - \frac{D_o}{D_e} = 1 - \frac{o_{01}}{e_{01}} = 1 - \frac{4}{4.4211} = 0.095.$$

Bypassing the construction of the matrix of expected coincidences, which is informative but can be cumbersome, a computationally more direct form for this α is

$$_{\text{binary}}\alpha = 1 - \frac{D_o}{D_e} = 1 - (n - 1)\frac{o_{01}}{n_0 \cdot n_1} = 1 - (20 - 1)\frac{4}{14.6} = 0.095.$$

By whichever form, in this numerical example, the reliability turns out to be barely 10% above what can be expected by chance. As the agreement that can be expected by chance is already (9.6 + 1.6)/20 = 56%, the 60% agreement that had been noted before looks far less impressive. Our data suggest that the two observers' performances are statistically equal to their having actually read (and reliably recorded) only about 10% of all newspaper articles, which is 1 article in our sample of 10, and assigned 0s or 1s to the remaining 90% by throwing dice. In light of the first of the three propositions in section 11.1, it is important to realize that we cannot know which the one correctly identified unit is, hence unreliability limits the chance of valid results.

Upon inspection of the circumstances of this devastating result, our political scientist may discover that the instructions to the two observers were incomprehensible or inappropriate to the coding task. Perhaps one or both observers failed to read the instructions carefully or did not know Chinese well enough to undertake the coding task. Whatever the reason, these data are far from being reliable, and content analysts would have to reject them without hesitation. Evidently, measuring 60% agreement means little by itself—a fact that should be a clear warning against the use of a reliability measure that does not account for chance, such as %-agreement.

For nominal data, I offer a *second interpretation.* It follows from the decomposition of the observed coincidences into two parts, α times the coincidences in a matrix of the ideal or perfect agreements, with all n values in the diagonal, plus $(1 - \alpha)$ times the coincidences in a matrix of what would be expected by chance:

$$\begin{bmatrix} o_{00} & o_{01} \\ o_{10} & o_{11} \end{bmatrix} = \alpha \begin{bmatrix} n_0 & \\ & n_1 \end{bmatrix} + (1-\alpha) \begin{bmatrix} e_{00} & e_{01} \\ e_{10} & e_{11} \end{bmatrix}$$

With reference to these three coincidence matrices: *α is the degree to which agreement exceeds expectations.* Specifically, *α is the proportion of the perfectly matching coincidences that, when added to the complementary proportion of chance coincidences, accounts for the coincidences that were observed.* This algebraic relationship can be demonstrated by means of the frequencies from our example:

α times *perfectly agreeing* coincidences: $.095 \begin{bmatrix} 14 & \\ & 6 \end{bmatrix} = \begin{bmatrix} 1.33 & \\ & .57 \end{bmatrix}$

$+$

$(1-\alpha)$ times *expected* coincidences: $(1-.095) \begin{bmatrix} 9.6 & 4.4 \\ 4.4 & 1.6 \end{bmatrix} = \begin{bmatrix} 8.67 & 4 \\ 4 & 1.43 \end{bmatrix}$

Total = *observed* coincidences: $\begin{bmatrix} 10 & 4 \\ 4 & 2 \end{bmatrix}$

Two Observers, Many Nominal Categories 11.3.2

I state this extension in three easily executable steps, first in general terms and then with a simple numerical example.

First step. Tabulate the values c and k that m observers, here the $m = 2$ observers A and B, respectively assign to each of r units, generically labeled u. This tabulation creates a *2-by-r reliability data matrix:*

Units *u*:	**1**	**2**	**...**	***u***	**...**	***r***
Observer A:	c_1	c_2	...	c_u	...	c_r
Observer B:	k_1	k_2	...	k_u	...	k_r

The following 2-by-12 reliability data matrix will serve as our example:

Units:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Mary:	a	a	b	b	b	b	b	c	c	c	c	c
Dave:	a	b	b	b	b	b	c	c	c	c	c	c

Second step. Construct the *matrix of observed coincidences.* With v different values in the reliability data, this matrix is a v-by-v matrix with v^2 cells containing all pairs of values assigned to units, or found in the columns of the reliability data matrix. Each value used in the reliability data matrix contributes one to the coincidence matrix, and each unit contributes two, one *c-k* pair of values and one *k-c* pair of values:

Values:	1	.	k	.	.	v	
1	o_{11}	.	o_{1k}	.	.	o_{1v}	n_1
.	.		.			.	.
.	.		.			.	
c	o_{c1}	.	o_{ck}	.	.	o_{cv}	$n_c = \Sigma_k\, o_{ck}$
.	.		.			.	.
v	o_{v1}	.	o_{vk}	.	.	o_{vv}	n_v
	n_1	.	n_k	.	.	n_v	$n = \Sigma_c \Sigma_k\, o_{ck}$

Sum the contents of the rows and columns in this matrix to their respective margins n_c and n_k and sum these margins to the total n. For this coincidence matrix to be accurate, its cells should be symmetrical around the diagonal, $o_{ck} = o_{kc}$. The total n should be $2r$, which is the number of values in the reliability data matrix. The vertical margin should equal the horizontal margin, and each n_c must equal the number of values c found in the reliability data matrix.

In our example, there are $v = 3$ values: a, b, and c. The observed coincidence matrix has 3-by-3 = 9 cells. As each unit is represented by two pairs of values, it will have to contain a total of $n = 24$ values. For example, the first unit contributes 2 entries in the a-a cell, the second contributes 1 to the a-b cell and 1 to the b-a cell, the next four units together contribute 8 to the b-b cell, and so on:

	a	*b*	*c*	
a	2	1	0	3
b	1	8	1	10
c	0	1	10	11
	3	10	11	24

Third step. Compute the *agreement coefficient* α as follows:

$$_{\text{nominal}}\alpha = 1 - \frac{D_o}{D_e} = 1 - (n-1)\,\frac{n - \sum_c o_{cc}}{n^2 - \sum_c n_c^2}$$

$$= 1 - (24-1)\,\frac{24 - (2+8+10)}{24^2 - (3^2 + 10^2 + 11^2)} = .734.$$

We can also construct the *matrix of expected coincidences.* It can serve as a convenient aid to interpreting reliability data, but, as demonstrated above, it is not necessary for computing α:

Values c:	1	.	k	.	.	v	
1	e_{11}	.	e_{1k}	.	.	e_{1v}	n_1
.	.		.			.	.
.	.		.			.	.
c	e_{c1}	.	e_{ck}	.	.	e_{cv}	$n_c = \Sigma_k\, e_{ck}$
.	.		.			.	.
v	e_{v1}	.	e_{vk}	.	.	e_{vv}	n_v
	n_1	.	n_k	.	.	n_v	$n = \Sigma_c \Sigma_k\, o_{ck}$

where:

$$e_{ck} = \begin{cases} n_c(n_k - 1)/(n - 1) \text{ iff } c = k \\ n_c \cdot n_k/(n - 1) \text{ iff } c \neq k \end{cases}$$

This matrix has the same margins as the matrix of the observed coincidences. The reason for treating the expectations in cells of matching values, $c = k$, differently from those of mismatching cells, $c \neq k$, follows from the discussion in section 11.3.1. In our example, expectations are computed as follows:

$$e_{aa} = n_a(n_a - 1)/(n - 1) = 3(3 - 1)/(24 - 1) = 0.2609,$$
$$e_{ab} = n_a \cdot n_b/(n - 1) = 3 \cdot 10/(24 - 1) = 1.3043,$$
$$e_{bc} = n_b \cdot n_c/(n - 1) = 10 \cdot 11/(24 - 1) = 4.7826,$$
and so on.

The expected coincidences are tabulated as follows:

	a	b	c	
a	.26	1.30	1.44	3
b	1.30	3.91	4.78	10
c	1.44	4.78	4.78	11
	3	10	11	24

By comparing the expected with the observed coincidences, one can locate sources of disagreement in the coincidences that fail to deviate from expectations. Comparisons of this kind give rise to several algebraic expressions within which we can recognize a *third interpretation* of α—restricted to nominal data, however:

$$_{\text{nominal}}\alpha = 1 - \frac{D_o}{D_e} = \frac{A_o - A_e}{A_{max} - A_e} = \frac{\sum_c o_{cc} - \sum_c e_{cc}}{n - \sum_c e_{cc}} = \frac{\sum_c (o_{cc} - e_{cc})}{\sum_c (n_c - e_{cc})}.$$

In the second version of α, A_o is the observed agreement, A_e is the expected agreement, and $A_{\max}$ is the largest possible agreement. Agreements, A, can be proportions, percentages, or frequencies. In the third version, agreements A appear as the sum of the diagonal entries in the observed and expected coincidence matrices, respectively, where $A_{\max} = n$ is the total number of values. In the forth version we see the same differences but now expressed as the sum of the differences between the observed and expected coincidences in the diagonal cells. Thus *α is the proportion of the observed to expected above-chance agreement.*

Entering the numbers from our example into each version yields the following—excepting rounding errors:

$$_{\text{nominal}}\alpha = 1 - \frac{.1667}{.6268} = \frac{83\% - 37\%}{100\% - 37\%} = \frac{(2 + 8 + 10) - (.26 + 3.91 + 4.78)}{24 - (.26 + 3.91 + 4.78)}$$

$$= \frac{(2 - .26) + (8 - 3.91) + (10 - 4.78)}{(3 - .26) + (10 - 3.91) + (11 - 4.78)} = 0.734$$

As one may recognize, the proportions are the same whether of the difference between %-agreements (second version), the difference between the frequencies in the diagonals of the two coincidence matrices (third version), or the sum of the differences between agreements in each of the diagonal cells in the two coincidence matrices (fourth version).

11.3.3 Many Observers, Many Nominal Categories, Missing Values

Applying the *first step* in section 11.3.2 now to any number m of observers, we start with an m-by-r reliability data matrix of these observers' values. For example:

Recording units u:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Observer A:	🕮	✉	☎	☎	✉	🕮	💻	🕮	✉			
Observer B:	🕮	✉	☎	☎	✉	✉	💻	🕮	✉	🗁		
Observer C:		☎	☎	☎	✉	☎	💻	✉	✉	🗁	🕮	☎
Observer D:	🕮	✉	☎	☎	✉	💻	💻	🕮	✉	🗁	🕮	
Number of values m_u:	3	4	4	4	4	4	4	4	4	3	2	1

Here, m = 4 observers categorized r = 12 messages, which are the units of analysis, by their sources, which are represented by icons. Note that 7 out of the 4-by-12 = 48 cells are empty. These are missing values. Observer C failed to consider unit 1. Observer A stopped coding after unit 9, B after unit 10, and D after unit 11. To account for these irregularities, we add one row to this reliability data matrix, which lists the number m_u of values assigned to unit u. Note the lone ☎ in unit 12, m_{12} = 1. It cannot be compared with anything in that unit.

Let us now focus on the *second step,* the construction of the *matrix of observed coincidences.* This step generalizes the second step in section 11.3.2. For two observers and 2-by-r reliability data matrices without missing values, the pairing of values and their tabulation was obvious. We now extend the idea of counting mismatches to more than two observers, to multiple pair comparisons among any number m_u of values in units or columns. To start, note that from m_u values, we can form $m_u(m_u - 1)$ pairs of values. In unit 1 there are

$m_1(m_1 - 1) = 3(3 - 1) = 6$ matching 🕮–🕮 pairs. In unit 2, containing 3 ✉s and 1 ☎, we can form a total of $m_2(m_2 - 1) = 4(4 - 1) = 12$ pairs, $3(3 - 1) = 6$ matching ✉–✉ pairs, $3 \cdot 1 = 3$ ✉–☎ pairs, and $1 \cdot 3 = 3$ ☎–✉pairs. One of the extremes can be seen in unit 6. Here all 4 values are different and so are the 12 pairs of values that can be formed from them. The other extreme can be seen in unit 12. The lone ☎ does not participate in any pair, which is indicated by $m_{12}(m_{12} - 1) = 1(1 - 1) = 0$ pairs of values. To be pairable, $m_u > 1$.

By our earlier definition, a matrix of observed coincidences accounts for the number of values that are pairable within units of analysis—not for the numbers of *units* being coded and not for the number of *pairs* that can be formed from these values. In order for each value to contribute exactly one entry to a coincidence matrix, each of the $m_u(m_u - 1)$ possible pairs of values contained in unit u must contribute $1/(m_u - 1)$ to the coincidence matrix. Except for now proportioning the contribution of each pair of values to the matrix of observed coincidences, everything else is exactly as in section 11.3.2.

Values:

	1	.	***k***	.	.	***v***	
1	o_{11}	.	o_{1k}	.	.	o_{1v}	n_1
.	.		.			.	.
.	.		.			.	
c	o_{c1}	.	o_{ck}	.	.	o_{cv}	$n_c = \Sigma_k\, o_{ck}$
.	.		.			.	.
v	o_{v1}	.	o_{vk}	.	.	o_{vv}	n_v
	$n1$	.	n_k	.	.	n_v	$n = \Sigma_c \Sigma_k\, n_{ck}$

where:

$$o_{ck} = \sum_u \frac{\text{Number of } c\text{-}k \text{ pairs in } u}{m_u - 1}$$

Accordingly, the 6 matching 🕮–🕮 pairs in unit 1 contribute $6/(3 - 1) = 3$ to the 🕮–🕮 cell of that matrix. The three kinds of pairs in unit 2 add $6/(4 - 1) = 2$ to the ✉–✉ cell, $3/(4 - 1) = 1$ to the ✉–☎ cell, and 1 to the ☎–✉ cell of that matrix. The 12 pairs of values in unit 6 add $1/(4 - 1) = 1/3$ each to 1 of 12 cells of that matrix. The contributions of these three units are seen in the first three matrices below. The fourth matrix sums the contributions of all 12 units, whereby it is important to realize that the lone ☎ in unit 12 makes no contribution to this matrix as it does not have anything with which to compare that ☎.

	Unit 1					Unit 2					Unit 6					Sum over all 12 units					
🕮	3										1/3	1/3	1/3			7	4/3	1/3	1/3		$9 = n_{🕮}$
✉						2	1				1/3		1/3	1/3		4/3	10	4/3	1/3		$13 = n_{✉}$
☎						1					1/3	1/3		1/3		1/3	4/3	8	1/3		$10 = n_{☎}$
💻											1/3	1/3	1/3			1/3	1/3	1/3	4		$5 = n_{💻}$
🗁																				3	$3 = n_{🗁}$
																9	13	10	5	3	$40 = n$

The construction of the rightmost of these four coincidence matrices completes the second step. For the *third step,* the computation of the agreement coefficient α, we follow the definition in section 11.3.2:

$$_{\text{nominal}}\alpha = 1 - \frac{D_o}{D_e} = 1 - (n-1)\,\frac{n - \sum_c o_{cc}}{n^2 - \sum_c n_c^2}$$

$$= 1 - (40-1)\,\frac{40 - (7+10+8+4+3)}{40^2 - (9^2 + 13^2 + 10^2 + 5^2 + 3^2)} = .734$$

11.3.4 Data With Different Metrics

Calculations of such data follow the first and second steps from section 11.3.2 when there are two observers and the first and second steps from section 11.3.3 when there are more than two observers, but they differ in the *third step,* the *computation of the agreement coefficient* α, which we will now generalize to any metric.

In nominal data, values either match or they do not. This had simplified the calculations of α in the previous situations. We now recognize other and more quantitative relationships between values, and they depend on the metric underlying a variable. As noted in Chapter 8 (section 8.4), a metric is defined by the operations that are applicable to the values of a variable. Although researchers generally can choose the metrics with which they want to analyze their data, they also need to acknowledge the nature of their data, whether the operations that define a metric make sense. For example, one cannot add two names or multiply ranks. The values of a ratio metric cannot be negative. Thus, in choosing an appropriate metric for α, researchers must take into account what a variable represents, which mathematical operations it can afford. Equally important is that they keep in mind the demands made by the data analysis methods they anticipate using. An analysis of variance requires interval data, contingencies are computed from nominal data, and so forth.

The way α accounts for diverse metrics is by using metric-specific difference functions δ^2_{ck} to weigh the observed and expected coincidences of *c-k* pairs of values. To make the role of these differences transparent, we restate the first interpretation of α in section 11.3 as

$$\alpha = 1 - \frac{D_o}{D_e} = 1 - \frac{\text{Average}_{\text{metric}}\,\delta^2_{ck} \text{ within all units}}{\text{Average}_{\text{metric}}\,\delta^2_{ck} \text{ within all data}}$$

$$= 1 - \frac{\sum_c \sum_k o_{ck\,\text{metric}}\delta^2_{ck}}{\sum_c \sum_k e_{ck\,\text{metric}}\delta^2_{ck}} = 1 - \frac{\sum \boxed{o_{ck}} \times \boxed{\delta^2_{ck}}}{\sum \boxed{e_{ck}} \times \boxed{\delta^2_{ck}}}.$$

The last of these four versions of α depicts coincidences and differences as square matrices whose entries are multiplied and summed as indicated in the third version. We now attend to the difference functions for the most common metrics—nominal, ordinal, interval, and ratio metrics—which we will state in two ways, in mathematical terms, which are general, and in terms of difference matrices, here exemplified with six typical values.

Nominal metric. For the nominal data used so far, reference to a metric was not needed, as they entail no quantitative differences. Values are merely distinct and freely permutable. When nominal values are represented numerically—area codes, banking PINs, the numbers on the jerseys of football players—adding or subtracting them from one another makes no sense. Two values are either the same or different—they match or they do not match. For generality's sake, we define and tabulate this property as a difference function, albeit of a primitive kind:

$$_{\text{nominal}}\delta^2_{ck} = \begin{cases} 0 \text{ iff } c = k \\ 1 \text{ iff } c \neq k \end{cases}$$

	📖	✉	☎	💻	🖫	🗁
📖	0	1	1	1	1	1
✉	1	0	1	1	1	1
☎	1	1	0	1	1	1
💻	1	1	1	0	1	1
🖫	1	1	1	1	0	1
🗁	1	1	1	1	1	0

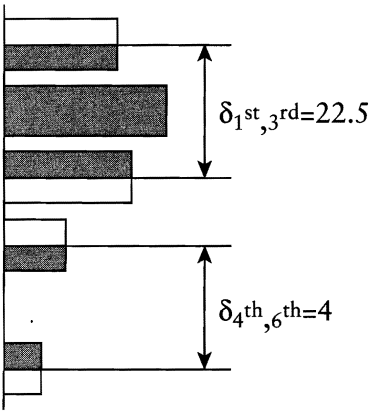
Naturally, all differences δ^2_{cc} between matching values are zero, which can be seen in the diagonal entries of their tabular forms. Off-diagonal differences are metric specific. It is a property of nominal data that all differences between mismatching values are identical, here 1.

Ordinal metric. In data, values are ranks. They have the meaning of 1st, 2nd, 3rd, 4th, and so forth. Ordinal differences are a function of how many ranks there are between any two ranks. The numerals used to label these ranks merely indicate their ordering.

We demonstrate the idea of ordinal differences with the example from section 11.3.3 whose marginal frequencies n_c. are now interpreted as of rank c: $n_{1st} = 9$, $n_{2nd} = 13$, $n_{3rd} = 10$, $n_{4th} = 5$, $n_{5th} = 0$, $n_{6th} = 3$. We add an unused rank, the 5th rank, for illustrative purposes.

$$_{\text{ordinal}}\delta^2_{ck} = \left(\frac{n_c}{2} + \sum_{g>c}^{g<k} n_g + \frac{n_k}{2}\right)^2 \text{ where } c < k$$

	1st	2nd	3rd	4th	5th	6th	
1st	0	11^2	22.5^2	30^2	32.5^2	34^2	$n_{1st} = 9$
2nd	121	0	11.5^2	19^2	21.5^2	23^2	$n_{2nd} = 13$
3rd	506	132	0	7.5^2	10^2	11.5^2	$n_{3rd} = 10$
4th	900	361	56	0	2.5^2	4^2	$n_{4th} = 5$
5th	992	462	100	6.3	0	1.5^2	$n_{5th} = 0$
6th	1,156	529	132	16	2.3	0	$n_{6th} = 3$



How ordinal differences are defined may be seen on the right of this difference matrix. The marginal frequencies of the ranks used by all observers are depicted as bar graphs, and shaded areas illustrate what goes into the differences between two ranks. That ordinal differences are not affected by the numerical values assigned to them may be seen in the example of the 5th rank. It is not used and does not make a difference in how far apart the 4th and 6th ranks are. Only a rank's ordering matters.

Note that the mathematical expressions state what the difference matrices exemplify. All difference matrices are symmetrical, $\delta_{ck} = \delta_{kc}$, and all of their diagonal entries are zero, $\delta_{cc} = \delta_{kk} = 0$.

Interval metric. One cannot add and subtract ranks, but the more familiar interval scales do afford these mathematical operations. In interval data, it is the simple algebraic differences that specify how far apart any two values are:

$$_{\text{interval}}\delta^2_{ck} = (c-k)^2$$

	−1	0	1	2	3	4
−1	0	1^2	2^2	3^2	4^2	5^2
0	1	0	1^2	2^2	3^2	4^2
1	4	1	0	1^2	2^2	3^2
2	9	4	1	0	1^2	2^2
3	16	9	4	1	0	1^2
4	25	16	9	4	1	0

Imagine drawing lines of equal differences in this matrix. These lines would parallel the diagonal, and their intervals would rapidly narrow with increasing distance from that diagonal.

Note that when all ranks are of equal frequency, $_{\text{interval}}\delta_{ck}$s and $_{\text{ordinal}}\delta_{ck}$s are proportional, and their agreement coefficients are equal: $_{\text{ordinal}}\alpha = {}_{\text{interval}}\alpha$.

Ratio metric. In ratio scales, algebraic differences between two values matter only in relation to how remote they are from zero, which is their reference point. Guessing the age of an older person within a year of accuracy may be remarkable, whereas guessing the age of a baby within a year is not. Losing a dollar may not be noticeable to a millionaire, but would mean losing everything for somebody who has only one. Age and income are ratio scales, as are frequencies. Algebraic differences between small values weigh more than the same differences between large values. The following difference function reflects these intuitions:

$$_{\text{ratio}}\delta^2_{ck} = \left(\frac{c-k}{c+k}\right)^2$$

	0	1	2	3	4	5
0	0	$(\frac{1}{1})^2$	$(\frac{2}{2})^2$	$(\frac{3}{3})^2$	$(\frac{4}{4})^2$	$(\frac{5}{5})^2$
1	1	0	$(\frac{1}{3})^2$	$(\frac{2}{4})^2$	$(\frac{3}{5})^2$	$(\frac{4}{6})^2$
2	1	.11	0	$(\frac{1}{5})^2$	$(\frac{2}{6})^2$	$(\frac{3}{7})^2$
3	1	.25	.04	0	$(\frac{1}{7})^2$	$(\frac{2}{8})^2$
4	1	.36	.11	.02	0	$(\frac{1}{9})^2$
5	1	.44	.18	.06	.01	0

Whereas in interval metrics the lines of equal differences are parallel to the diagonal, in ratio metrics they all join in the zero-point and extend, fanlike, into infinity, with ratio differences being the tangent of the angular deviation from the 45-degree diagonal line.

By acknowledging the above metrics, our third computational step generalizes the way the nominal α was calculated in sections 11.3.2 and 11.3.3:

$$D_o = \frac{1}{n} \sum_c \sum_k o_{ck\,\text{metric}} \delta^2_{ck},$$

$$D_e = \frac{1}{n} \sum_c \sum_k e_{ck\,\text{metric}} \delta^2_{ck} = \frac{1}{n(n-1)} \sum_c n_c \sum_k n_{k\,\text{metric}} \delta^2_{ck}.$$

In the form recommended for computational convenience, we bypass references to the matrix of expected coincidences by computing these expectations indirectly, from the margins of the matrix of observed coincidences. And because coincidence and difference matrices are symmetrical, we can reduce the number of algebraic operations by half when summing only one of the two triangles with off-diagonal coincidences:

$$_{\text{metric}}\alpha = 1 - \frac{D_o}{D_e} = 1 - (n-1) \frac{\sum_c \sum_{k>c} o_{ck\,\text{metric}} \delta^2_{ck}}{\sum_c n_c \sum_{k>c} n_{k\,\text{metric}} \delta^2_{ck}}.$$

To demonstrate this now generalized third step in the calculation of α, we continue to make use of the coincidences we already know from the example in section 11.3.3, for which purpose we recode their values by 📖 = 1, 📖 = 2, ☎ = 3, 💻 = 4, and 🗁 = 5. According to the above, we need the observed coincidences, their marginal frequencies, and differences for each coincidence:

Observed coincidences o_{ck}

	1	2	3	4	5	
1	7	4/3	1/3	1/3		9
2	4/3	10	4/3	1/3		13
3	1/3	4/3	8	1/3		10
4	1/3	1/3	1/3	4		5
5					3	3
	9	13	10	5	3	40

Interval differences $_{\text{interval}}\delta^2_{ck}$

	1	2	3	4	5
1	0	1^2	2^2	3^2	4^2
2	1^2	0	1^2	2^2	3^2
3	2^2	1^2	0	1^2	2^2
4	3^2	2^2	1^2	0	1^2
5	4^2	3^2	2^2	1^2	0

$$_{\text{interval}}\alpha = 1 - \frac{D_o}{D_e} = 1 - (n-1) \frac{\sum_c \sum_{k>c} o_{ck} (c-k)^2}{\sum_c n_c \sum_{k>c} n_k (c-k)^2} = 1 - (40-1)$$

$$\frac{4/3 \cdot 1^2 + 1/3 \cdot 2^2 + 1/3 \cdot 3^2 + 4/3 \cdot 1^2 + 1/3 \cdot 2^2 + 1/3 \cdot 1^2}{9(13 \cdot 1^2 + 10 \cdot 2^2 + 5 \cdot 3^2 + 3.4^2) + 13(10.1^2 + 5.2^2 + 3.3^2) + 10(5.1^2 + 3.2^2) + 5.3.1^2} = .849$$

For interval α coefficients, we offer this computational simplification:

$$
\begin{aligned}
{}_{\text{interval}}\alpha &= 1-(n-1)\,\frac{\sum_c \sum_{k>c} o_{ck}\,(c-k)^2}{\sum_c n_c \sum_{k>c} n_k\,(c-k)^2} \\
&= 1-(n-1)\,\frac{\sum_c n_c c^2 - \sum_c \sum_k o_{ck}\,ck}{n\sum_c n_c c^2 - \left(\sum_c n_c c\right)^2}.
\end{aligned}
$$

Referring to previous findings regarding our example, one may ask why the ${}_{\text{interval}}\alpha$, just calculated as .849, is larger than the ${}_{\text{nominal}}\alpha$, earlier calculated to be .743. Although this difference is small and our example has far too few units to allow us to draw interesting conclusions, if neither were the case, one might notice that observed mismatches are more frequent near the diagonal than away from it, which would be typical of interval data, and one may then be tempted to conclude that coders were using these values as if they had some kind of ordering, as if the intervals between them mattered.

11.4 STATISTICAL PROPERTIES OF α

The agreement coefficient α is a statistical measure. Below, I briefly discuss four conditions and precautions for accepting or rejecting measured reliabilities: insufficient variation, sampling considerations, statistical significance, and standards for the reliability of data.

11.4.1 Insufficient Variation

A frequently puzzling condition arises when reliability data show insufficient variation. Consider this rather extreme example:

Reliability Data

Units:	1	2	3	4	5	6	7	8
Observer A:	0	0	0	0	0	0	0	0
Observer B:	0	1	1	0	0	0	0	1

Observed coincidences

	0	1	
0	14	1	15
1	1	0	1
	15	1	16

Expected coincidences

	0	1	
0	14	1	15
1	1	0	1
	15	1	16

Here, the matrices of observed and expected coincidences are identical, $D_o = D_e = 0$ and $\alpha = 0$ by definition. Technically, α would be indeterminate, as $1 - 0/0$ can be either 0 or $-\infty$. I use zero not only because extreme negative values have no meanings here but because such data could not be relied upon, although here for lack of variation. Time and time again, statistical novices have found this condition

difficult to accept. They have argued that there evidently is considerable agreement on the value "0," in fact in 7 out of 8 units or 88%. How could α suggest reliability to be absent? This argument overlooks the requirement that reliability data must exhibit variation. Perhaps the material coded was mostly of the same kind, perhaps the observers found their task boring and settled on scoring habitually—we do not know. However, in the 8th unit, in which one observer noticed something out of line, in the only unit that seems to be at variance with all the others, the two observers fail to agree, and $\alpha = 0$, as it should. When one wants to report on the overwhelming frequency of one value, exceptions are particularly important.

Now, suppose observer B had assigned a "0" to unit 8 as well. Then we would have even less evidence that the observers did their job. They could have been too tired to notice unusual variations, or they could have been lazy—assurances to the contrary aside—and agreed in advance simply to label everything "0" without reading. In effect, they functioned just as two measuring instruments with frozen numerals would, just as a broken clock does, showing the same time all the time. Variability is a prerequisite of any measuring instrument's responsiveness to phenomena external to it.

Suppose further that the two observers agree on a "1" for unit 8, then observer A's change of heart concerning this value would cause α to jump from 0 to +1. This might seem surprising. Although the two 1s are still rare values, by chance alone, they could show up either in one unit and yield $\alpha = 1.00$, as suggested, or in two different units and yield $\alpha = -.071$. There is at least some variation to which α does respond.

Statistical Significance 11.4.2

A common mistake that researchers make is to accept or reject data as reliable when the null hypothesis that agreement occurs by chance can be rejected with statistical confidence. However, the whole reason for measuring the reliability of data is to ensure that they do not deviate too much from perfect agreement, not that they deviate from chance. In the definition of α, chance agreement merely serves as one of two anchors for the agreement scale to be interpretable, the more important reference point being perfect agreement. As the distribution of α is unknown, approximating a χ^2 distribution only in appearance, we have resorted to generating distributions of α by bootstrapping—that is, by drawing several thousand subsamples from the available reliability data, computing α for each, and thereby generating a probability distribution of hypothetical α values that could occur within the constraints of the observed coincidences. Figure 11.4 depicts a typical distribution of bootstrapped α values. It gives us two statistical qualifications of the observed α:

- α's *confidence interval* for the chosen level of significance p (two-tailed):

$$\alpha_{\text{largest}} \geq \alpha_{\text{observed}} \geq \alpha_{\text{smallest}}$$

■ The probability q of failing to reach the smallest acceptable reliability α_{min}:

$$q \mid \alpha < \alpha_{min}$$

The confidence interval spells out the range within which the observed α can be expected to vary with $(1 - p)\%$ certainty. The probability q of failing to reach the required reliability is the probability of making the wrong decision of accepting data as reliable when they could well be below the accepted standard.

For the numerical example in section 11.3.3, we measured $_{nominal}\alpha = .743$. After drawing 20,000 samples from these data, the 99% confidence interval is between $\alpha_{smallest} = .615$ and $\alpha_{largest} = .860$, and the probability of failure to exceed $\alpha_{min} = .667$ is $q = .198$.

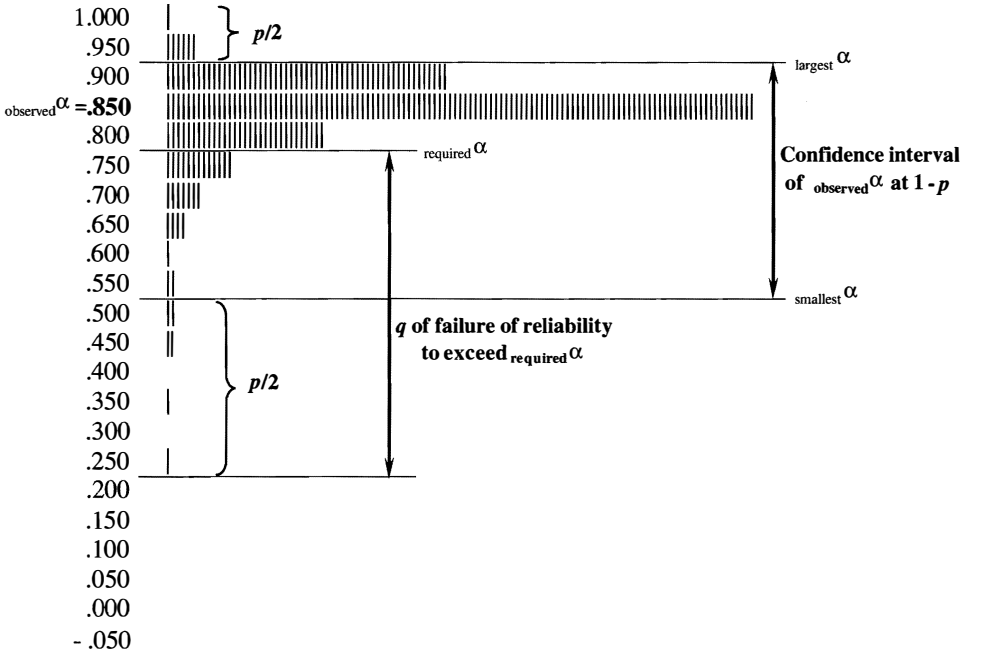


Figure 11.4 Bootstrapped Distribution of α and Areas of Interest

11.4.3 Sampling Considerations

A frequent question concerns how large the sample of units should be for α to be meaningful. There is no simple answer to this question. I consider two distinct issues of sampling reliability data below and then make a practical recommendation.

To establish the *reliability of data,* reliability data need to be representative of the population of data whose reliability is in question. The size of adequate samples is inversely related to the proportion of units in different categories. Rare but important categories

must occur in sufficient numbers so as to make the same difference in the reliability calculations as subsequent analyses. As a rule of thumb, each category of units should occur often enough to yield at least five agreements by chance (I will be more specific below, however). Where the research results are particularly important or data are not very numerous, content analysts have resorted to coding all data at least twice. This can be time-consuming and costly, but it bypasses sampling problems altogether, and the measured α is the α of the population of data.

To establish the *reliability of instructions*—data languages, systems of categories, or the measuring instruments that are to be applicable to all possible data—the diversity of the reliability data is more important than their representativeness of the data in hand. When the reliability of data is at issue, the unreliability of categories that do not occur in the data does not need to enter the agreement measures, as it will not affect the trustworthiness of the given data. Moreover, in assessing the reliability of data, there may be good reasons for disagreements on rare categories to be compensated by agreements on frequently used ones. In contrast, when the reliability of coding instructions is in question—reliable coding instructions should generate reliable data regardless of which categories are used and whatever their frequencies may be—reliability data need to contain all the categories of units or values that the instrument lists as possible. Ideally, reliability data here contain each category of units that the instrument distinguishes with equal and sufficiently large frequency. This would justify the oversampling of rare kinds of units and the undersampling of frequently occurring ones.

To obtain *required sample sizes* for either situation, we rely on Bloch and Kraemer's (1989, p. 276) Formula 3.7, which estimates α's variance for binary data (2-by-2 matrices). The minimum number of units,

$$N_c = z^2 \left(\frac{(1 + \alpha_{min})\ (3 - \alpha_{min})}{4(1 - \alpha_{min})\ p_c(1 - p_c)} = \alpha_{min} \right),$$

turns out to be a function of the smallest estimated proportion p_c of values c in the population, the smallest acceptable reliability α_{min} below which data would have to be rejected as unreliable, and the desired level of statistical significance, represented by the corresponding z value for one-tailed tests.

For convenience, Table 11.2 lists the sample sizes for the three smallest acceptable reliabilities α_{min}, four levels of statistical significance, and 10 probabilities p_c. The latter are expressed here as the probabilities of the number of equally likely values. For example, suppose the least frequent of all categories is expected to occur with $p_c = 0.125$. If α is to exceed the smallest acceptable reliability of, say, $\alpha_{min} = .667$, to be sure that 95% of all αs (at the .05 level of statistical significance) satisfy this condition, the table suggests a minimum reliability sample size of 71 units. If one has no clue of the expected probabilities, one may start sampling assuming that all categories are equally likely but then add units to the sample to compensate for the unequal proportions found in the data.

Table 11.2 Required Sample Sizes

Smallest acceptable α	**.667**				**.800**				**.900**			
Significance level	**.100**	**.050**	**.010**	**.005**	**.100**	**.050**	**.010**	**.005**	**.100**	**.050**	**.010**	**.005**
10 values or $p_c = .100$	53	86	172	211	90	147	294	360	181	298	595	730
9 values or $p_c = .111$	48	79	157	192	82	134	267	328	165	271	542	665
8 values or $p_c = .125$	43	71	141	173	74	121	241	295	149	245	489	600
7 values or $p_c = .143$	39	63	126	154	66	108	215	263	133	218	437	535
6 values or $p_c = .167$	34	56	111	136	58	95	189	232	117	192	384	471
5 values or $p_c = .200$	29	48	96	117	50	82	164	200	101	167	333	408
4 values or $p_c = .250$	25	41	81	99	43	70	139	170	86	142	284	348
3 values or $p_c = .333$	21	34	68	83	36	59	117	143	73	120	239	292
2 values or $p_c = .500$	19	30	60	74	32	52	103	127	65	106	212	259

Bloch and Kraemer's Formula 3.7 assumes that α is normally distributed, which is not quite so, as suggested by Figure 11.4. Moreover, the formula for N_c and Table 11.2 do not take into account the number of observers involved in the data-making process. To understand how this number affects the confidence in the required sample sizes, it is important to keep the purpose of reliability evaluations in mind. Reproducibility amounts to predicting from a measured agreement among actual observers the agreement that potential observers working elsewhere would achieve as well. The ability to generalize the reliabilities from a sample to a population of data is only half of the problem—the representativeness of the observers is the other half. If observers with similar qualifications are unavailable elsewhere, the measured agreement may not be interpretable as reproducibility. An increase in the number m of observers grants added assurances that the process is replicable elsewhere. Our estimated sample sizes do not address this experience. However, they err merely by being conservative.

Standards for Data Reliability 11.4.4

The ultimate aim of testing reliability is to ensure that unreliabilities are negligible so as to justify continuing the coding or starting an analysis of the data toward answering research questions. Below, I answer three commonly asked questions regarding reasonable standards.

What is an acceptable level of reliability? Facing the real difficulties of obtaining perfect agreement, can one require that α be at least .95, .90, or .80? Unfortunately, although every content analyst faces this question, there is no set answer. To shed light on how different levels of reliability can be interpreted, Marten Brouwer, a colleague of mine from the Netherlands, designed an experiment. He gave coders who spoke only English a set of complicated Dutch words and asked them to describe U.S. television characters using those words. The Dutch words had no resemblance to any words in English, and the English speakers could hardly pronounce them, but the words must have invoked some consistent associations with perceived personality characteristics because the agreement was $\alpha = .44$. Knowing the observers' unfamiliarity with these words, nobody in their right mind would draw conclusions from the records these subjects created to what they had observed or read. The agreement was well above chance, but on account of entirely unknown associations in the observers' minds, associations that the researcher and the users of findings based on them can hardly imagine. This finding gives us another reference point on the scale of α's values, one that one should not approach. After further explorations of the relationship between achieved agreement and understanding of the categories involved, we adopted the following policies:

- Rely only on variables with reliabilities above $\alpha = .800$.
- Consider variables with reliabilities between $\alpha = .667$ and $\alpha = .800$ only for drawing tentative conclusions.

These standards have been adopted in numerous content analyses in the social sciences and they might continue to serve as guidelines. Similar guidelines have been proposed for other coefficients—for example, Fleiss (1981) has proposed guidelines for Cohen's κ (kappa). However, relying on α's distribution gives us criteria that are more justifiable, as a distribution responds to the sample size as well. In these terms, the recommendations could be rephrased:

- Do not accept data with reliabilities whose confidence interval reaches below the smallest acceptable reliability α_{min}, for example, of .800, but no less than .667.
- Ensure that the probability q of the failure to exceed the smallest acceptable reliability α_{min} is reasonably small, for example, .050, or the tolerable risk of drawing wrong conclusions.

I recommend such levels with considerable hesitation. The choice of reliability standards should always be related to the validity requirements imposed on the research results, specifically to the costs of drawing wrong conclusions. If the outcome of a content analysis will affect someone's life—such as in court proceedings—the analyst should not rely on data whose probability of leading to a wrong decision is less than what is commonly accepted (for example, the probability of being killed in a car accident). The results of most content analyses do not have drastic consequences, however, and so the researchers can adopt far lower standards. Even a cutoff point of $\alpha = .800$—meaning only 80% of the data are coded or transcribed to a degree better than chance—is a pretty low standard by comparison to standards used in engineering, architecture, and medical research.

Whether a content analysis is exploratory or intended to be decisive, no researcher should ignore reliability, set reliability standards so low that findings cannot be taken seriously, use deceitful ways of generating reliability data, or apply deceptive agreement measures to prop up the appearance of reliability. In content analysis, the famous phrase "beyond reasonable doubt" has an operationalizable meaning.

Given the α values of separate variables, how reliable are the data as a whole? α is defined for separate variables, and most content analyses involve many. Ideally, the variables of a data language are logically independent, free of conceptual redundancies, and observed disagreements affect the research results equally. Under these conditions, every variable counts and every variable must also be reliable.

It is a serious mistake to average the reliabilities of the variables of a complex instrument and take this average as a measure of overall data reliability. Computing averages assumes that higher values compensate for lower ones. Typical content analyses include clerical variables, publication, date, length, and mechanically obtained measures that tend to be perfectly reliable, whereas the variables that matter are most typically more difficult to code and end up being less reliable.

Researchers who average such reliabilities will have an unwarranted sense of trust that may lead them astray in their conclusions. A condition in which averaging may make sense arises when the values of several variables are subsequently summed or averaged to form a composite index. Averaging their reliabilities is justifiable only if this index is such that scoring on one account is as good as scoring on another and omissions in one variable compensate for commissions in another, so that agreements in one variable balance disagreements in another. These rather stringent conditions are not easy to meet.

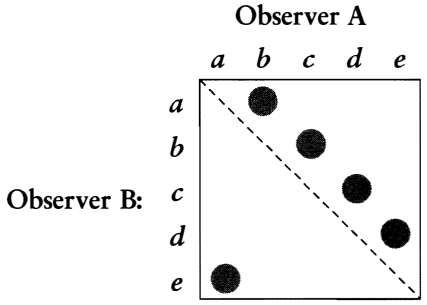
Generally, when variables are equally important to the research effort, any unreliable variable can become a bottleneck for confidence in the data as a whole. Thus, *for multivariate data, the lowest α among the variables is the joint reliability of the data as a whole.* This might appear a harsh criterion, but it is entirely consistent with the common practice of dropping unreliable variables from further analysis. Trading the information that one hoped unreliable variables would provide for the reliability of the data as a whole is the only valid strategy for improving joint reliability once data have been gathered.

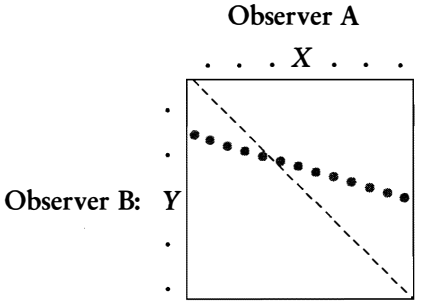
How does unreliability affect the quality of findings? Part of this question has already been answered in section 11.1: Unreliability limits the chance that results will be valid. Here we are concerned with where reliability should be measured. Because data generation—the reading, coding, or transcribing of texts—typically is the most uncertain part of a study, content analysts routinely assess the reliability of their data at the *front end* of the research effort. Indeed, when the data can be shown to be reliable, the remainder of the work is generally unproblematic. Yet some content analyses are robust in that the unreliabilities that enter the data-making process are barely noticeable in the results. In others, small differences may tip the scale in important decisions, turning affirmative answers into negatives. To appreciate this sensitivity, analysts would have to know how disagreement in the data is transmitted through the analytical process to its outcome. Ideally, this would entail analyzing not just one set of data but as many as the researchers can obtain by permuting the values among which disagreements were encountered. This would generate a distribution of possible results. If this distribution is too wide to allow the analysts to draw conclusions from these data, the reliability standard for front-end coding would need to be set higher. If the distribution does not limit the drawing of conclusions, front-end reliability standards may be relaxed. Analysts can achieve a simple but not quite sufficient approximation of this distribution by performing separate analyses on each observer's data and ascertaining how the results would differ. At the very minimum, content analysts need to trace observed disagreements through the analysis to the results. Generally, data reduction techniques—creating frequency accounts, for example, or combining several variables into an index—tend to reduce the effects that front-end disagreements have on the results. If disagreements are amplified, customary standards for front-end reliabilities may not suffice. Although the reliability of data is surely critical to the success of a content analysis, this is not the only reliability measure that counts.

11.5 OTHER COEFFICIENTS AND CORRESPONDENCES

In a survey of content analyses published in the journal *Journalism & Mass Communication Research* from 1971 through 1995, Riffe and Freitag (1996; cited in Riffe, Lacy, & Fico, 1998) found that only 56% reported assessments of reliability. In their review of the consumer research literature from 1981 through 1990, Kolbe and Burnett (1991) noted similar deficiencies: 31% of the published content analyses showed no concerns for reliabilities; 19% mentioned reliability without revealing any method of calculation; 35% reported %-agreement, including Holsti's; 8% used one of several less well-known measures (each mentioned only once or twice in the sample); and 7% used Krippendorff's α. Lombard, Snyder-Duch, and Bracken (2002) found reliability discussed in 69% of 200 content analyses indexed in *Communication Abstracts* from 1994 through 1998. These are discouraging findings. In this section, I review the most popular agreement indices found in the literature and discuss their shortcomings and relationships to the α coefficient.

Some content analysts make the common but serious mistake of considering the performances of individual observers as variables; applying readily available association, correlation, or consistency measures to them; and interpreting these as measures of reliability. Correlation coefficients—Pearson's product-moment r_{ij}, for example—measure the extent to which two logically separate interval variables, say X and Y, covary in a linear relationship of the form $Y = a + bX$. They indicate the degree to which the values of one variable predict the values of the other. Agreement coefficients, in contrast, must measure the extent to which $Y = X$. High correlation means that data approximate *any* regression line, whereas high agreement means that they approximate the 45-degree line. If one observer is consistently, say, two points behind the other, or they follow a regression line as suggested by the gray dots in the right of the following two contingency matrices, correlation is perfect, but agreement is not. The same holds for cross-tabulations of nominal data. In the contingency matrix on the left, ● signifies a nonzero frequency of co-occurring categories. There is a 1-to-1 relationship between the two sets of categories, association is perfect, and the use of categories by one observer is perfectly predictable from that by the other. But as categories do not match, there is no agreement whatsoever.





Although correlation coefficients have been used as reliability measures, even recommended (Potter & Levine-Donnerstein, 1999, p. 277), especially in the literature on psychological testing (see the otherwise informative exposition of reliability and validity issues by Carmines & Zeller, 1979; and a review by Salgado & Moscoso, 1996), the above should make clear that in content analysis their use is seriously misleading.

Regarding agreement indices, I have already noted criticism of %-agreement—or "crude" agreement, as this measure is sometimes called—in section 11.3.1. Yet its relatives creep into the literature in different guises. Holsti (1969, p. 140), for example, describes Osgood's (1959, p. 44) reliability index as $2M/(N_1 + N_2)$, wherein M is the number of units on whose categorizations two readers agree, N_1 is the number of units identified by one reader, and N_2 is that number identified by another. Although Holsti presents pertinent criticism of %-like agreement indices, citing Bennett, Alpert, and Goldstein's (1954) arguments and building up to his recommendation of Scott's (1955) π (pi), it is amazing how many content analysts still overlook the by now widely published objections to this uninterpretable agreement measure. For example, Neuendorf (2002) and Lombard et al. (2002), instead of discouraging %-agreement and its relatives, discuss it as a (perhaps too) liberal alternative.

Bennett et al. (1954) were probably the first to realize that %-agreement A_o was the more difficult to achieve the more categories were available for coding. They proposed a coefficient $S = (A_o - 1/K)/(1 - 1/K)$ that corrects for this effect, where K is the number of categories available for coding. It is remarkable that this coefficient has been reinvented at least five times since it was originally proposed: as Guilford's G (Holley & Guilford, 1964), as the RE (random error) coefficient (Maxwell, 1970), as Janson and Vegelius's (1979) C, as κ_n (Brennan & Prediger, 1981), and, most recently, as intercoder reliability coefficient I_r (Perreault & Leigh, 1989). Perreault and Leigh (1989) were at least aware of S. Proponents of this coefficient cite reasons ranging from fairness to each category and consistency with the research traditions of their disciplines to the absence of hard knowledge about the true distribution of categories in the population from which reliability data were sampled. In treating all categories as equally likely, S inflates agreement when used unevenly, especially when some are not used at all. The latter enables researchers to manipulate reliability in their favor by adding unused or rarely used categories to the set. Perreault and Leigh (1989) argue that chance-corrected agreement coefficients, such as κ, are too conservative, whereas S (or their I_r), they say, is not. For arguments against this assessment, see Krippendorff (in press-b).

In response to S's shortcomings, Scott (1955) proposed his reliability index π (pi), which is of the aforementioned form ${}_{\text{nominal}}\alpha = (A_o - A_e)/(A_{\max} - A_e)$:

$$\pi = \frac{A_o - P_e}{1 - P_e},$$

where A_o is the proportion of units with matching categories (%-agreement), $P_e = \Sigma_k p_k^2$ is the proportion of pairs of values that are expected to match by chance,

and p_k is the proportion of values k in the reliability data jointly identified by two observers.

In effect, p_k estimates the proportion of values k in the *population of units* that the observers are facing. P_e treats observers as interchangeable and considers their collective judgment as the best estimates of the population proportions, assuming (as is customary) that differences among observers wash out in their average. Thus P_e becomes the agreement that can be expected to occur in the population when chance prevails.

Subsequently, Cohen (1960) introduced an unfortunate modification of Scott's π into the literature, trying to bring agreement measurement closer to conventional contingency approaches, as he said, and calling it κ (kappa). This coefficient is popular in biomedical and educational research, where most of its proponents work, but inappropriate in assessing reliability in content analysis, as we shall see. Cohen merely replaced Scott's expected agreement P_e with a proportion that conforms to the tradition of association statistics, which I call P_c. κ is defined by

$$\kappa = \frac{A_o - P_c}{1 - P_c},$$

where A_o is the proportion of units with matching categories (%-agreement) (as in Scott's π), $P_c = \Sigma_k\, p_{Ak} \cdot p_{Bk}$ (unlike the P_e in π), p_{Ak} is the proportion of the value k used by observer A, and p_{Bk} is the proportion of value k used by the other observer, B. Here, P_c is the agreement that can be expected when the two observers' proclivity to use their categories differently is assumed and taken for granted.

A numerical example may demonstrate how π and κ differ in their results. Consider two contingency tables containing the frequencies of units recorded by two observers:

		Observer A			
Categories:		a	b	c	
Observer B	a	12	9	9	30
	b	9	14	9	32
	c	9	9	20	38
		30	32	38	100

$A_o = .460$
$\pi = .186$
$\kappa = .186$

		Observer A			
Categories:		a	b	c	
Observer B	a	12	18	18	48
	b	0	14	18	32
	c	0	0	20	20
		12	32	56	100

$A_o = .460$
$\pi = .186$
$\kappa = .258$

Both tables show reliability data to have the same %-agreement A_o, 46 out of 100, as can be seen in their identical diagonal entries. But they differ in how disagreements are distributed, which is also manifest in the two observers' marginal

frequencies. In the left-hand table, observers agree on these frequencies and Scott's π and Cohen's κ are the same, as they should be. But when they disagree on these frequencies, as is apparent in the table on the right, κ exceeds π, suggesting that there is more agreement. Evidently, this is far from so. There are still only 46 units in the diagonal cells. How can κ be so mistaken? Note that the 54 mismatches, initially populating both off-diagonal triangles, have now become unevenly distributed, occupying only one. What has increased thereby is not agreement but the predictability of the categories used by one coder from the categories used by the other. Unlike κ, π is not affected by where the mismatching values occur. In content analysis, it indeed should not matter who contributed which disagreements and, when data are nominal, which categories are confused. Moreover, predictability has nothing to do with reliability. Thus, when mismatches in a contingency table are unequally distributed, κ adds a measure of the uneven distribution of mismatching categories to the coefficient, π does not. κ overestimates reliability and cannot serve as a reliability index in content analysis and similar coding tasks.

It should be pointed out that Cohen (1960), in his original proposal of κ, falsely criticized π for ignoring "one source of disagreement between a pair of judges [due to] their proclivity to distribute their judgments differently over the categories" (p. 41). His proposal to modify π achieved just the opposite. κ counts disagreements among observer preferences for available categories as agreements, not as disagreements, as Cohen claimed it would. This is a major conceptual flaw. Brennan and Prediger (1981) describe this property of κ by pointing out that "two judges who independently, and without prior knowledge, produce similar marginal distributions must obtain a much higher agreement rate to obtain a given value of kappa, than two judges who produce radically different marginals." The first two judges "are in a sense penalized" for agreeing on marginal frequencies (p. 692). Many proponents of κ reproduce Cohen's false claim without verification. Zwick (1988), citing others, mentions this flaw as well and suggests testing for marginal homogeneity before computing κ, but this merely patches up κ's obvious inadequacies.

The structural differences between the most popular agreement coefficients (Kolbe & Burnett, 1991; Lombard et al., 2002; Neuendorf, 2002) can be seen most clearly when reduced to their simplest binary or dichotomous forms. We will state these in terms of a 2-by-2 contingency matrix, containing proportions *a, b, c,* and *d* of the $n = 2r$ values contributed by the two observers.

Observer A's Values:		0	1	
Observer B's Values:	0	a	b	p_B
	1	c	d	q_B
		p_A	q_A	1

Population Estimates

from $n = 2r$ = the number of values used jointly by both observers

$\overline{p} = (p_A + p_B)/2$

$\overline{q} = (q_A + q_B)/2 = 1 - \overline{p}$

	Agreement = 1 –	Observed / Expected Disagreement
%-agreement	A_o = 1 –	$(b + c)$ /
Bennett et al. (1954)	S = 1 –	$(b + c)$ / $2 \cdot ½ \cdot ½$,
Scott (1955)	π = 1 –	$(b + c)$ / $2\bar{p}\,\bar{q}$,
Krippendorff (1970a)	α = 1 – $\frac{n-1}{n}$	$(b + c)$ / $2\bar{p}\,\bar{q}$,
Cohen (1960)	κ = 1 –	$(b + c)$ / $p_A q_B + p_B q_A$.

where ½ is the logical probability of 0 or 1; $\bar{p}$ and $\bar{q} = (1-\bar{p})$ are population estimates; $n = 2r$ = the total number of values used jointly by both observers; and $(n - 1)/n$ corrects α for small sample sizes.

Evidently, all of these measures contain the proportion of mismatches $(b + c)$. The measure of %-agreement A_o stops there, making no allowances for expected disagreements and saying nothing about the categories that are available for coding and about the population of data being categorized.

S acknowledges expectations but states them relative to the number of categories in the coding instrument. In its binary form, with the two categories being equally likely, the expected disagreement is $2 \cdot ½ \cdot ½ = ½$, or 50%. *S* is sensitive to the number of categories available but says nothing about the population of data whose reliability is at stake.

In both π and α, the expected disagreement in the two cells b and c is $2\bar{p}\,\bar{q}$, which is obtained from the population estimates $\bar{p}$ for 0s and its complement $\bar{q}$ for 1s. π and α are alike except for the factor $(n - 1)/n$, which corrects α for small sample sizes. With rising sample sizes, π and the nominal α become asymptotically indistinguishable.

Cohen's κ, by contrast, reveals itself as a hybrid coefficient (Krippendorff, 1978). Its observed disagreement $(b + c)$, conforms to all the other agreement coefficients, but its expected disagreement, $p_A q_B + p_B q_A$, resembles that of correlation and association measures. In fact, it calculates expected disagreements in the off-diagonal cells just as the familiar χ^2 statistic does. Yet in assessments of agreements, association and predictability are not at issue, as already suggested. Evidently, κ is concerned with the two individual observers, not with the population of data they are observing, which ultimately is the focus of reliability concerns.

To relate the structure of these agreement coefficients to our conception of reliability, I want to be clear: Reliability concerns arise when the trustworthiness of data is unknown and there are doubts about how reliably the phenomena of interest have been observed and are described by these data. It is the population of phenomena that is of ultimate interest to researchers and that interchangeable observers face in the form of samples and record, hopefully without disagreement.

Having no privileged access to the whole population of phenomena, researchers must estimate its composition from whatever they can reasonably trust. It is a fundamental assumption of reliability concerns that the perceptions of many are trusted more than the perception of any one. Consequently, we must estimate the distribution of categories in the population of phenomena from the judgments of as many observers as possible (at least two), making the common assumption that observer differences wash out in their average. Evidently, by estimating the proportions of categories in a population of phenomena, π and α refer to this population and build the above reliability conception into their definitions; κ does not.

This brings us to the *fourth interpretation.* According to the above, α's reliability scale is anchored at two hypothetical points: the condition of all observers applying the same conceptualizations to the same set of phenomena, one at a time and without disagreement, yielding—sampling considerations aside— individually identical and collectively accurate accounts of the distribution of phenomena in the population; and the condition of observers applying the same conceptualizations to the same set of phenomena, but without coordination as to what they are recording, yielding individually randomized but still collectively accurate accounts of this distribution. The latter is the best estimate of the categories in the population of phenomena. On this scale, α *is the degree to which independent observers, using the categories of a population of phenomena, respond identically to each individual phenomenon.* Thus α can be interpreted as measuring the reliability of data relative to an estimated population of the very phenomena that these data are to represent. κ does not estimate such a population and cannot speak about the reliability of data in their capacity to represent the phenomena of interest.

Probably because the aforementioned Cronbach's (1951) alpha has also been called a measure of reliability, it has found its way into the content analysis literature as well. However, this coefficient was never intended to assess coding efforts and in fact it cannot. In its binary form, it is Kuder and Richardson's (1937, p. 158) KR-20 and measures what in our context could be called the consistency of individual coders' judgments. It takes the variances of individual observers ($\Sigma_i p_i q_i$) as the variance of the "true" scores and expresses this as a proportion of the total variances (σ_T^2), which is the sum of the true score and the measurement error (Carmines & Zeller, 1979). In its familiar terms, and in the above terms, it is defined as follows:

$$\text{Cronbach's alpha} = \frac{m}{m-1}\left(1 - \frac{\sum_i p_i q_i}{\sigma_T^2}\right)$$

$$= 2\left(1 - \frac{p_A q_A + p_B q_B}{a(p_A + p_B)^2 + (b+c)(p_A - p_B)^2 + d(q_A + q_B)^2}\right).$$

It belongs to the family of correlation coefficients and must not be construed as an agreement measure. Their popularity and use in other empirical domains notwithstanding, %-agreement, Cohen's κ, and Cronbach's alpha are simply not appropriate for assessing the reliability of coding.

With the exception of Krippendorff's α, the above-listed coefficients were all conceived for nominal data generated by just two observers. α is appropriate to all common metrics, not just the nominal metric, and applicable to any number of observers, not just two. Moreover, α copes with missing data and is corrected for small sample sizes. As sample sizes become large, the nominal α approaches Scott's π, as noted above. The ordinal α then becomes identical to Spearman's rank correlation coefficient ρ (rho) without ties in ranks, and the interval α turns out to equal Pearson et al.'s (1901) intraclass correlation coefficient R_I, which is the correlation coefficient r_{ij} applied to coincidence rather than contingency matrices (Krippendorff, 1970a). These correspondences attest to α's generality, demonstrate its connections to well-established statistics, and enable researchers to apply uniform reliability standards to a variety of data.

There have been other proposals to extend agreement coefficients to several observers (Fleiss, 1971; Krippendorff, 1970b, 1971). Because these issues can become very complex, most researchers consider special cases. Landis and Koch (1977) have considered κ-type agreements in terms of majority opinion. Hubert (1977) has taken the approach of accepting only perfect consensus. Craig (1981) has proposed a modification of Scott's π to account for majorities among observer judgments. My extension of α to many observers was initially guided by Spiegelman, Terwilliger, and Fearing's (1953b) effort to rank patterns of disagreement in nominal data by subjective judgments. The pairwise sum of differences $\Sigma_i\Sigma_j\delta_{ij}^2$ in D_o and in D_e approximates their subjective ranks nearly perfectly, which gave me the confidence to apply this function to any number of observers. α demands neither majority judgments nor consensus and privileges no particular number of observers. For multiple observers, the interval α is compatible with variance analysis (Krippendorff, 1970b). Its handling of multiple observers is consistent with Siegel and Castellan's (1988, p. 286) recent extension of Scott's π to many observers (although reluctantly named κ there, causing much confusion). The ability to cope with missing data is a natural by-product of α's extension to many interchangeable observers (Krippendorff, 1992).

In the first edition of *Content Analysis,* I sketched several diagnostic devices—devices for computing the reliability of individual units, for identifying unreliable observers within a group of observers, for determining the metric in use by observers, for tracing the flow of disagreement through coding decision hierarchies—and ways to trade information for increased reliability, for example, by lumping unreliable categories or using variables conditionally (Krippendorff, 1980b, pp. 148–154). Recent advances include the ability to evaluate the reliability of multiple interpretations (the above is limited to assigning at most one value to units of analysis; see section 11.2.3), the use of standards to determine accuracy, and the bootstrapping of α's distribution. A presentation of these analytical capabilities must await another publication. A recent breakthrough was α's extension to calculating the reliability of unitizing (Krippendorff, 1995a). As this is a frequent problem, including in computer-aided qualitative text analysis, I outline its steps in the following section, being aware that developing it further would go beyond the needs of most readers.

α-AGREEMENT FOR UNITIZING 11.6

In most content analyses, units are not given or natural. They must be identified within an otherwise undifferentiated continuum, for example, of linearly ordered text, time records, tape recordings, or flows—within any continuous and quantifiable dimension. I have already mentioned the example of clipping newspaper articles, to which can be added highlighting and coding text segments (as in qualitative text analysis software) pertaining to a research question, identifying episodes of a certain kind in video recordings of social interactions, and generating data by having subjects push buttons to mark periods of interest, disinterest, or emotional arousal while watching TV shows. Analysis of the reliability of unitizing has been largely ignored, mostly because the development of adequate reliability measures has lagged far behind the development of coding methods. Guetzkow (1956) was the first to address the reliability of unitizing. Unfortunately, his coefficient measures the extent to which two observers agree on the *number of identified units,* not on the actual units counted, leaving totally open the question of what, if anything, the observers had in fact agreed on. Osgood's (1959, p. 44) and Holsti's (1969, p. 140) %-like indices have the same problem but moreover fail to consider chance. α_U (Krippendorff, 1995a) overcomes these deficiencies while taking its place in the family of α coefficients, sharing its essential properties. Below, I sketch the basic idea and offer a simple computational example (I do not expect that readers will perform calculations on larger data sets by hand).

Units of length. Unitizing starts with an initially undifferentiated *continuum,* about which we need to know only its beginning, *B,* and length, *L.* The unit for measuring these lengths is *the smallest distinguishable length, duration, or number*—for example, the characters in text, the frames of a video, the smallest measurable length on a ruler, or the smallest time interval one can distinguish. Lengths are expressed in full integers, not in decimal points, and not in units of varying size (such as fractions of inches for small lengths and feet or miles for larger ones).

Reliability data. Unitizing means partitioning a given continuum into sections. Reliability data for unitizing (see Figure 11.2) require that at least two observers or methods unitize the same continuum. These sections are numbered consecutively for each individual observer or unitizer. Each section is characterized by the following:

- Its consecutive number *g* or *h,* separately for each observer
- The observer *i* or *j* who identified it
- The category *c* or *k* to which units are assigned

- Its beginning b, subscripted by <*cig*>, <*cjh*>, <*kig*>, <*kjh*>, and so on, locating it on the continuum
- Its length ℓ, also subscripted by <*cig*>, <*cjh*>, <*kig*>, <*kjh*>, and so on, expressing its extent
- A binary value w, also subscripted by <*cig*>, <*cjh*>, <*kig*>, <*kjh*>, and so on, indicating whether it is an identified unit or an empty stretch between two units:

$$w_{cig} = \begin{cases} 0 \text{ iff section } \langle cig \rangle \text{ is not a unit} & \langle cig \rangle \\ 1 \text{ iff section } \langle cig \rangle \text{ is a unit} & \langle cig \rangle \end{cases}$$

These terms enable us to specify each observer's unitization of the same continuum as diagrammed in Figure 11.5.

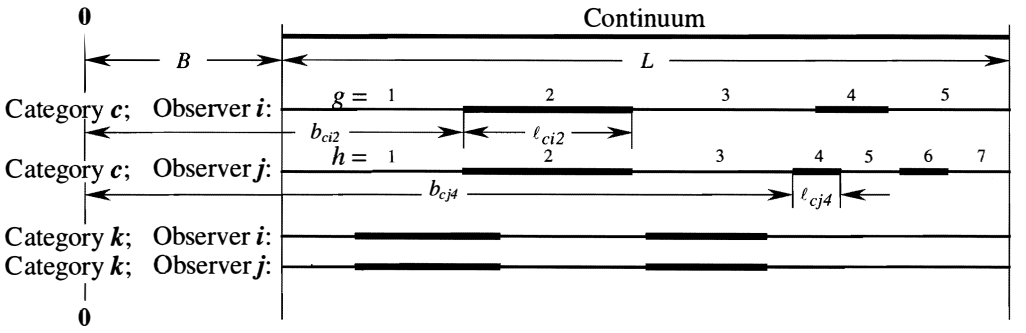


Figure 11.5 Unitizing Terms

Difference function δ^2_{cigjh}. For reliability to be perfect, units must be of the same category and occupy the same stretch of the continuum. Deviations from this ideal give rise to differences. The difference δ^2_{cigjh} between any two sections <*cig*> and <*cjh*> of the same category c and identified by two observers, i and j, is a function of their failure to overlap perfectly. In the above terms, this function is

$$\delta^2_{cigjh} \begin{cases} (b_{cig} - b_{cjh})^2 + (b_{cig} + \ell_{cig} - b_{cjh} - \ell_{cjh})^2 & \text{iff } w_{cig} = w_{cjh} = 1 \text{ and } +\ell_{cjg} < b_{cig} - b_{cjh} < \ell_{cjh}, \\ \ell^2_{cig} & \text{iff } w_{cig} = 1,\ w_{cjh} = 0 \text{ and } \ell_{cjh} - \ell_{cig} \geq b_{cig} - b_{cjh} \geq 0. \\ \ell^2_{cjh} & \text{iff } w_{cig} = 0,\ w_{cjh} = 1 \text{ and } \ell_{cjh} - \ell_{cig} \leq b_{cig} - b_{cjh} \leq 0. \\ 0 & \text{Otherwise.} \end{cases}$$

The first condition pertains to pairs of overlapping units. Here δ^2 is the sum of the squares of the two nonoverlapping lengths. The second condition applies when

observer *i*'s unit *g* is fully contained in observer *j*'s gap *h*. The third condition is the converse of the second and applies when observer *i*'s gap *g* fully contains observer *j*'s unit *h*. The fourth condition applies when two sections of the continuum overlap perfectly, are both gaps (not units), or have nothing in common in the continuum. To see how this function behaves in response to different degrees of overlap between two observers' unitizations, consider the examples in Figure 11.6.

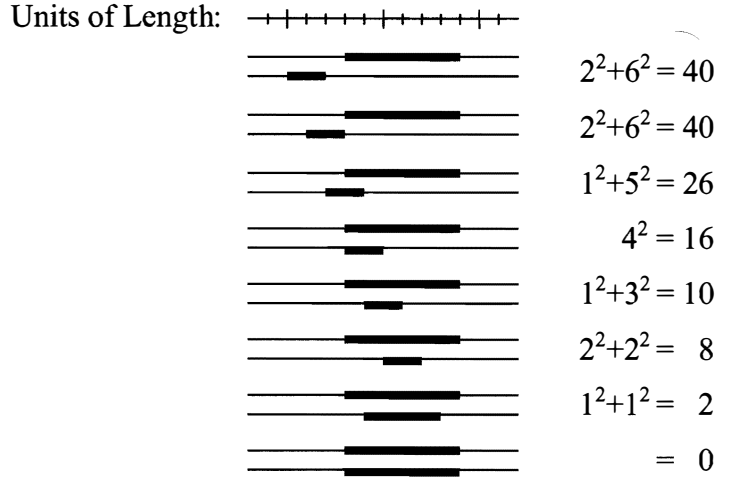


Figure 11.6 Examples of Differences for Unitizing

Observed disagreement D_{oc}. The observed disagreement D_{oc} for unitizing a continuum and assigning its units to category *c* is obtained—much as the observed disagreement D_o of the α-measures for coding is obtained—through comparison of each observer's sections with all other observers' sections on the continuum. Summing the observed differences and dividing the sum by its maximum yields the measure of the observed disagreement D_{oc}:

$$D_{oc} = \frac{\sum_{i=1}^{m}\sum_{g}\sum_{j\neq i}^{m}\sum_{h}\delta^2_{cigjh}}{m(m-1)L^2},$$

where *m* is the number of observers that unitize the continuum, $m(m-1)$ is the number of pairs of observers whose units are being compared, *L* is the length of the continuum, and δ^2_{cigjh} is the difference between two sections <*cig*> and <*cjh*>. Incidentally, $\delta^2_{cigjh} = \delta^2_{cjhig}$. Note that the four sums pair all observers, *i* and *j*, with each other but not with themselves and run through all pairs of sections of any one category *c*.

Expected disagreement D_{ec}. The expected disagreement measures the differences between $mL(mL-1)$ virtual comparisons of all possible unitizations, combining the actually identified units and gaps between them in all possible ways, and

applying the above disagreement measure D_{oc} to each pair. Actually making these comparisons would be a transcomputational task, hence the need for a simpler formula for D_{ec}. With $N_c = \Sigma_{i=1}^{m} \Sigma_g w_{cig}$ = the total number of units of category c identified by all m observers, the expected disagreement for units in category c is

$$D_{ec} = \frac{\frac{2}{L}\sum_{i=1}^{m}\sum_g w_{cig}\left[\frac{N_c-1}{3}\left(2\ell_{cig}^3 - 3\ell_{cig}^2 + \ell_{cig}\right) + \ell_{cig}^2 \sum_{j=1}^{m}\sum_h \left(1-w_{cjh}\right)\left(\ell_{cjh} - \ell_{cig} + 1\right) \text{ iff } \ell_{cjh} \geq \ell_{cig}\right]}{mL\,(mL-1) - \sum_{i=1}^{m}\sum_g w_{cig}\ell_{cig}(\ell_{cig}-1)}.$$

Its proof is lengthy; I provide it elsewhere (see Krippendorff, 1995a). Note that, just as the expected disagreement for coding makes no reference to observers and recording units, the expected disagreement for unitizing makes no reference to the original location of the sections on the continuum or to who identified them.

Let me merely point out the principal components of the above equation. The first double summation in its numerator goes through all observers' segments, with w_{cig} separating these into units proper and the omitted gaps between two units. The first expression in the square brackets accounts for the differences between one unit and all other units overlapping with that unit in all possible ways. The double summation in the square brackets checks whether the one unit would fit into any of the gaps between units and adds the differences due to each. In the denominator, mL is the number of possible locations for a unit to occur in the continuum, and $mL(mL - 1)$ is the number of pair comparisons of such units that the disagreement measure calculates virtually. The second expression in the denominator discounts the results of comparing sections with themselves.

α_U-agreement for all categories. Similar to coding,

$$\alpha_U = 1 - \frac{\sum_c D_{oc}}{\sum_c D_{ec}}.$$

For a computational example, we assign numerical values to the beginnings and lengths of the units in Figure 11.5

Continuum	**B**	**L**	
	150	300	
Sections	***b***	***ℓ***	***w***
*ci*1	150	75	0
*ci*2	225	70	1
*ci*3	295	75	0
*ci*4	370	30	1
*ci*5	400	50	0
*cj*1	150	70	0
*cj*2	220	80	1

*cj*3	300	55	0
*cj*4	355	20	1
*cj*5	375	25	0
*cj*6	400	20	1
*cj*7	420	30	0
*ki*1	150	30	0
*ki*2	180	60	1
*ki*3	240	60	0
*ki*4	300	50	1
*ki*5	350	100	0
*kj*1	150	30	0
*kj*2	180	60	1
*kj*3	240	60	0
*kj*4	300	50	1
*kj*5	350	100	0

The nonzero *differences* between the two observers' sections in category *c* are

$$\delta^2_{ci2j2} = (225 - 220)^2 + (225 + 70 - 220 - 80)^2 = 5^2 + 5^2 = 50 = \delta^2_{cj2i2},$$
$$\delta^2_{ci4j4} = (370 - 355)^2 + (370 + 30 - 355 - 20)^2 = 15^2 + 25^2 = 850 = \delta^2_{cj4i4}$$
$$\delta^2_{ci5j6} = 20^2 = 400 = \delta^2_{cj6i5}$$

Evidently, the first pair of units in category *c*, showing observer *i* as merely a bit more conservative than *j* is, contributes very little by comparison to the remaining three units, which appear more scattered on the continuum. In category *k* all differences are zero:

$$\delta^2_{ki2j2} = 0 = \delta^2_{ki4j4}.$$

With these differences in hand, the *observed disagreement* in category *c* becomes:

$$D_{oc} = \frac{\delta^2_{ci2j2} + \delta^2_{ci4j4} + \delta^2_{ci5j6} + \delta^2_{cj2i2} + \delta^2_{cj4i4} + \delta^2_{cj6i5}}{m(m-1)L^2} = \frac{2(50 + 850 + 400)}{2(2-1)\,300^2} = .0144.$$

In category *k*, the observed disagreement $D_{ok} = .0000$, of course.

Calculating the *expected disagreement* with the above formula requires a few more steps. In category *c*, with a total of $N_c = 2 + 3 = 5$ identified units, the expected disagreement is obtained as follows:

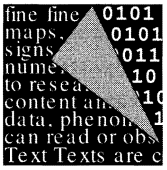
$$D_{ec} = \frac{\frac{2}{300}\left\{\begin{array}{l}\left[\frac{5-1}{3}(2\cdot 70^3 - 3\cdot 70^2 + 70) + 70^2 \begin{pmatrix} 75-70+1 \\ +75-70+1 \\ +70-70+1 \end{pmatrix}\right] \\ + \left[\frac{5-1}{3}(2\cdot 30^3 - 3\cdot 30^2 + 30) + 30^2 \begin{pmatrix} 75-30+1 \\ +75-30+1 \\ +50-30+1 \\ +70-30+1 \\ +55-30+1 \\ +30-30+1 \end{pmatrix}\right] \\ + \left[\frac{5-1}{3}(2\cdot 80^3 - 3\cdot 80^2 + 80)\right] \\ + \left[\frac{5-1}{3}(2\cdot 20^3 - 3\cdot 20^2 + 20) + 20^2 \begin{pmatrix} 75-20+1 \\ +75-20+1 \\ +50-20+1 \\ +70-20+1 \\ +55-20+1 \\ +25-20+1 \\ +30-20+1 \end{pmatrix}\right] \\ + \left[\frac{5-1}{3}(2\cdot 20^3 - 3\cdot 20^2 + 20) + 20^2 \begin{pmatrix} 75-20+1 \\ +75-20+1 \\ +50-20+1 \\ +70-20+1 \\ +55-20+1 \\ +25-20+1 \\ +30-20+1 \end{pmatrix}\right]\end{array}\right\}}{2\cdot 300(2\cdot 300-1) - \begin{pmatrix} 70(70-1) \\ +30(30-1) \\ +80(80-1) \\ +20(20-1) \\ +20(20-1) \end{pmatrix}} = .0532.$$

And in category k, with a total of $N_k = 2 + 2 = 4$ identified units, the expected disagreement turns out to be $D_{ek} = .0490$.

Finally, the α_U-agreement for unitizing with the two categories is

$$\alpha_U = 1 - \frac{D_{oc} + D_{ok}}{D_{ec} + D_{ek}} = 1 - \frac{.0144 + .0000}{.0532 + .0490} = .8591,$$

which concludes this illustration.



CHAPTER 13

Validity

> Validation provides compelling reasons for taking the results of scientific research seriously. It can serve as the ground for developing theories and the basis of successful interventions. This chapter develops a typology of validation efforts that content analysts may utilize in justifying their research. It also shows ways in which analysts can quantitatively assess at least some of these efforts.

VALIDITY DEFINED 13.1

Validity is that quality of research results that leads us to accept them as true, as speaking about the real world of people, phenomena, events, experiences, and actions. A measuring instrument is considered valid if it measures what its user claims it measures. A content analysis is valid if the inferences drawn from the available texts withstand the test of independently available evidence, of new observations, of competing theories or interpretations, or of being able to inform successful actions.

Riffe, Lacy, and Fico (1998) suggest, "The essence of the validity problem in content analysis as well as in other research . . . is that research should speak as truthfully as possible to as many as possible" (p. 150). The meaning of *truthful* is the central focus of this chapter. The idea that research should speak "to as many people as possible" leads to useful distinctions among face validity, social validity, and empirical validity.

Face validity is "obvious" or "common truth." We appeal to face validity when we accept research findings because they "make sense"—that is, they are plausible and believable "on their face"—usually without having to give or expecting to hear detailed reasons. It makes sense, indeed, to measure public attention to an issue by the relative frequency with which the issue is mentioned

in mass media. It makes sense to measure the quality of political deliberations by the number of alternative reasons brought into a discussion. After subsequent empirical scrutiny, face validity may prove untenable, but it appears just right at the time the research is accepted. Face validity does not equal expectations. For example, it did not occur to anyone in the 1970s that members of minority groups were targets of jokes in U.S. television fiction more often than were members of the majority until content analysts thought to pursue this topic and found that correlation. Findings like these make sense in retrospect. Although face validity has its roots in common sense, in widely shared consensus, it is fundamentally an individual's judgment with the assumption that everyone else would agree with it.

The reason content analysts rely on face validity perhaps more than do researchers who use other methods of inquiry is that content analysis is fundamentally concerned with readings of texts, with what symbols mean, and with how images are seen, all of which are largely rooted in common sense, in the shared culture in which such interpretations are made, which is difficult to measure but often highly reliable at a particular time. This is not to say that the reliance on face validity is absent in other research endeavors. In fact, even the most rigorous researchers would not use methods or publish findings that violate their common sense. Face validity is the gatekeeper for all other kinds of validity. It is difficult to explain how face validity works, yet it is omnipresent.

Social validity is that quality of research findings that leads us to accept them on account of their contribution to the public discussion of important social concerns, such as violence on television, antisocial messages in rap music, racism in sermons, hate talk on radio talk shows, and lack of civility in political campaign advertisements. Research examining such public issues is socially validated by proponents and antagonists who worry about these issues and are eager to translate research findings into actions. In Riffe et al.'s (1998) terms, social validity is "the degree to which the content analysis categories created by the researchers have relevance and meaning beyond an academic audience" (p. 137). Unlike face validity, the social validity of content analysis studies is often debated, negotiated, and a matter of public concern. A content analysis that is socially valid can attract public attention, propose practical solutions, and generate funding. Publicly acknowledged authorities on the subject of research are key to the social validity of the findings. Arguing from the privileged position of scientists, content analysts may well inadvertently become such authorities, especially when they explain their findings in seemingly irrefutable quantitative terms at congressional hearings, for example, or to advocacy groups working in support of particular public agendas. The line between accepting research because of the reputation of the researcher and accepting it because of the evidence it brought forth is often blurred. Even empirically oriented test psychologists have started to take social issues increasingly seriously. A significant part of the latest edition of the *Standards for Educational and Psychological Testing* established by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1999) is

concerned with fairness, with the responsibilities of the researcher for the test taker, and with testing and public policy. In these standards, social validity concerns appear as concerns about the possible social or psychological consequences of testing. Although most researchers enjoy this kind of validity, it is distinct from empirical validity, which is the focus of the remainder of this chapter.

Empirical validity is the degree to which available evidence and established theory support various stages of a research process, the degree to which specific inferences withstand the challenges of additional data, of the findings of other research efforts, of evidence encountered in the domain of the researcher's research question, or of criticism based on observations, experiments, or measurements as opposed to logic or process. Campbell (1957) calls the latter "internal validity." Empirical validity cannot deny intuition (face validity), nor can it divorce itself entirely from social, political, and cultural factors (social validity)—after all, scientific research is reviewed by the researchers' peers, who may have their own theoretical agendas and are hardly immune to social and political concerns. However, in the following I separate empirical validity from the face and social validities and consider it to be established largely within the scientific community and to be based on rational arguments that bring empirical evidence to bear on the research results, the research process, and the conditions under which data were acquired.

In discussing empirical validity, several content analysis textbooks follow the American Psychological Association's *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (1954), a landmark publication that defined the kinds of validity concerns that psychologists face when they are developing tests of individual characteristics or abilities. In addition to face validity, the chief types of validities distinguished in the 1954 *Recommendations* were content validity, construct validity, criterion-related validity, and predictive validity. These *Recommendations* focused narrowly on evidence, and so did not mention social validity, which concerns questions regarding the larger context of psychological testing.

Content validity is the extent to which a psychological test captures all the features that define the concept that the test claims to measure. For example, measurement of an applicant's aptitude for a job would require a complete inventory of job requirements, not just IQ tests or motivational measures.

Construct validity acknowledges that many concepts in the social sciences—such as self-esteem, alienation, and ethnic prejudice—are abstract and cannot be observed directly. To validate a measure of self-esteem, for example, one would first have to spell out the observable behaviors and verbal responses that the self-esteem concept entails, then measure and correlate these with the proposed measure, and finally examine whether or not each correlation supports what a theory of self-esteem predicts.

Criterion-related validity, sometimes called *instrumental validity,* is the degree to which a measure correlates with or estimates something external to it; for example, IQ may be shown to correlate with grade point average. Customarily, criterion-related validity is divided into two kinds, concurrent and predictive.

Concurrent validity is demonstrated by correlations that concur with the test in question, and *predictive validity* concerns variables that estimate features that may become available sometime in the future.

These classical distinctions have undergone various transformations. The above-mentioned *Recommendations* gave way to the first version of the *Standards for Educational and Psychological Testing* in 1985, and these were followed by the 1999 *Standards,* which no longer support distinctions between validity types, referring instead to types of "validity evidence" (see American Educational Research Association et al., 1999, p. 11). They also recognize that theoretical constructs underlie all measurements, and this recognition led to a classification of validating evidence based on *test content, response process, internal structure, relations to other variables,* and, as mentioned above, the *consequences of testing* (pp. 11–17). The debate about these conceptions of validity is ongoing, but it is important to recognize that the above focuses narrowly on only one theory of scientific inquiry, measurement theory, and is informed mainly by one disciplinary orientation, the psychological testing of individuals. Measurements are only part of what content analyses can provide, and inferences about psychological characteristics of individuals or populations of individuals are rare in content analysis, although not excluded.

To understand the conceptions of validity that are useful in the conduct of content analysis, one must keep in mind that all empirical validation efforts enlist evidence and established theories to ensure that research results are taken seriously. When the goal is merely the construction of theory, a research project may matter only to a small scientific community. But when research is intended to have policy implications—when findings are to aid business decisions, provide evidence in court, categorize people, or affect the lives of individual human beings in other ways—wrong conclusions may have costly consequences. Validation reduces the risk of making decisions based on misleading research findings. Content analysts and psychologists concerned with testing have to cope with different risks.

Content analysts face at least three kinds of obstacles when they try to apply traditional methods of validation: substantive, conceptual, and methodological. We return to our conceptual framework of content analysis (as depicted in Figure 2.1) to understand all three.

Substantively, probably most important to content analysts is the acknowledgment that texts are used because of their meanings to people other than the analysts, starting with producing, reading, and interpreting text and proceeding to constructing, maintaining, or undoing social realities. The object of content analysis is far more complicated than analyzing how individuals respond to test questions with preformulated answers. Highlighting some of these differences, Potter and Levine-Donnerstein (1999) have introduced useful distinctions among three kinds of content analyses, liberally rephrased as content analyses that aim to describe manifest content, content analyses that provide inferences about latent patterns, and content analyses that provide interpretations (or "make projections," as they say; p. 261). The first kind conforms to a measurement

conception of content analysis and is unproblematic as far as the application of the above kinds of validity is concerned, although it conflicts with the conclusions from Chapter 2. The second refers to a context—of experts, as Potter and Levine-Donnerstein suggest, and established theory in relation to which validity needs to be established. Figure 2.1 depicts just this, the scientific community or any chosen stakeholder group providing this context. The third kind of content analysis these scholars describe allows more freedom of imagination on the part of content analysts but is constrained by cognitive schemata and inferential rules that are embodied in a designated population of text users whose conceptions are both the focus of analysis and the source of validity. This ethnographic and interpretivist-sounding conception does not differ from that outlined in Chapter 2 as a definition of content analysis of the second kind, except for the authors' allowing research results to be more freewheeling. In this kind of content analysis, Potter and Levine-Donnerstein affirm, validity standards cannot be divorced from chosen contexts, but they limit analysis to the conceptions of individuals for whom the analyzed texts have the meanings they have. Sensitivity to a context distinguishes content analysis from other methods of inquiry and provides the criteria for acceptance of its results.

The *conceptual obstacle* to the validation of content analyses derives from an inadequate definition of *content as inherent in text* (see Chapter 2) and the attendant commitment merely to describe *it*. To be sure, all descriptions are abstract and as such arbitrary, and conceiving content as inherent to texts has enabled content analysts to apply any category schemes they please, provided the schemes are reliable. Content analysts with such a conception of content in mind confuse their descriptions of content with how others may read or use the same texts, seemingly needing no further validation. For example, suppose a content analysis of mass-media entertainment concludes that there has been a shift in the United States from material to spiritual values. There is no way to validate this finding unless the analysts are willing to take responsibility for their definitions and make clear where this claimed shift should be observable as well (other than in the analyzed media), in whose lives this shift is expected to make a difference, or which variables are expected to correlate with this abstraction in order for it to be considered as describing something real. As mentioned in Chapter 2, one reason for the popularity of the conceptions of content as inherent to text and of content analysis as merely descriptive of this content is the virtual impossibility of empirically (in)validating such findings. In the absence of specificity about what could validate or invalidate the findings of a content analysis, appeals to face and social validity, which researchers can more easily control through their own rhetoric, public testimony, and publication, seem to be the only recourse.

The *methodological obstacle* to validation is more difficult to overcome. Consider the following trilemma: (a) If content analysts happen to have no independent evidence about what they are inferring, then validity or invalidity cannot be established, at least not until pertinent data show up. Content analysis shares this epistemological constraint with all predictive efforts. (b) If these analysts had evidence about the context of their texts—that is, about phenomena

related to the analyzed texts—but used it in the development of their analytical constructs, then this evidence would no longer be independent of the research results and hence cannot be used to validate the findings. And finally, (c) if these analysts had concurrent evidence that could validate their inferences but kept it away from their analysis, there would be no point in their conducting the content analysis. It would at best add one incident to vindicate the analysis. Content analysts can resolve this trilemma, at least in part, by relying on various forms of imperfect and indirect validating evidence about the phenomena of interest. I address how this could happen in the following section.

13.2 A TYPOLOGY FOR VALIDATING EVIDENCE

A fundamental difference between psychological testing and content analysis is that the latter is concerned with bodies of text that are meaningful in relation to a chosen context (see Chapter 2), whereas the former does not acknowledge that relationship and the inferential step it entails. It follows that content analysts must empirically demonstrate the context sensitivity of their research. In addition, it is important to bear in mind that content analysis data, texts, and findings, although unquestionably mediated by human individuals as language-using beings, readers/writers, conceptualizers, and actors, are not necessarily about individuals. This means that psychological measurement theoretical conceptions of validity have to be expanded by model theoretical conceptions, as graphically outlined in Figure 9.1. This epistemological shift calls for validating evidence that differs, albeit in only some categories, from the evidence defined by the above-mentioned *Standards,* whose conceptions content analysis can adopt only in parts. To start, three kinds of validating evidence may enter a content analysis:

- Evidence that justifies the treatment of text, what it is, what it means, and how it represents what (This is loosely related to what the *Standards* refer to as "evidence based on test content"; American Educational Research Association et al., 1999, p. 11.)
- Evidence that justifies the abductive inferences that a content analysis is making (Here, analysts are concerned with the validity of the analytical constructs on which they rely. This is loosely related to what the *Standards* call "evidence based on [the] internal structure" of a test; p. 13.)
- Evidence that justifies the results, whether a content analysis contributes answers to the research questions of other researchers or is borne out in fact (This is loosely related to the older "criterion-related validity," which the *Standards* discuss in terms of "evidence based on relations to other variables.")

These and the following distinctions are depicted in Figure 13.1.

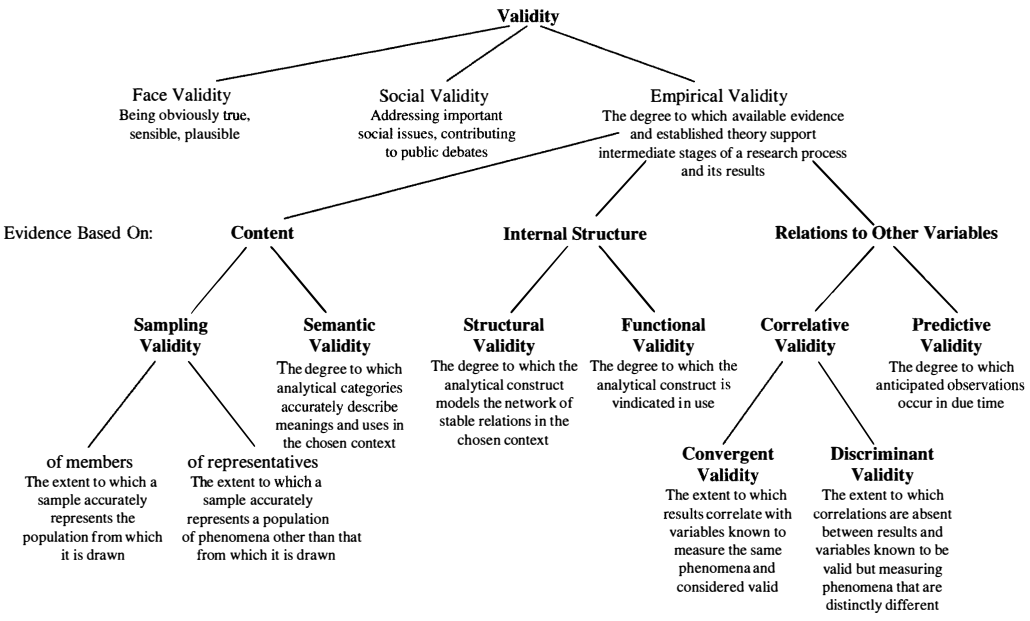


Figure 13.1 A Typology of Validation Efforts in Content Analysis

Evidence that justifies the treatment of texts concerns largely the sampling and recording phases of a content analysis. Such evidence may be divided into two kinds:

- Evidence on *sampling validity* concerns the degree to which a sample of texts accurately represents the population of phenomena in whose place it is analyzed. Ideally, content analysts actively sample a population, using sampling plans that ensure representativeness. But in many practical situations, texts become available by their sources' choice and contain intentional biases in representing the phenomena of their interest. Content analysts may not be able to control such biases but want to know whether and how much such samples can be trusted.

- Evidence on *semantic validity* ascertains the extent to which the categories of an analysis of texts correspond to the meanings these texts have within the chosen context. The anthropological preference for emic or indigenous rather than etic or imposed categories of analysis and ethnographers' efforts to verify their interpretations with their informers demonstrate concern for semantic validity, here with reference to the lives of populations of interviewees. In content analysis, other contexts are considered as well, but contexts they must be. Other kinds of research efforts (psychological testing for one, but also survey research) tend to avoid semantic validity by controlling the range of permissible answers to questions and not really exploring what these questions could mean to their subjects. Semantic validity is allied largely with content analysis.

Evidence that justifies the abductive inferences of a content analysis sheds light on how well the analytical construct in use actually does model what it claims to model. Again, two kinds of such evidence may be distinguished:

- Evidence on *structural validity* demonstrates the structural correspondence between available data or established theory and the modeled relationships or the rules of inference that a content analysis is using.
- Evidence on *functional validity* demonstrates a functional correspondence between what a content analysis does and what successful analyses have done, including how the chosen context is known to behave. If these behaviors covary repeatedly and over a variety of situations, one can suspect that they share an underlying construct.

This distinction between structural validity and functional validity is motivated by Feigl's (1952) distinction between

> two types of justification . . . validation and vindication. In this context, *validation* is a mode of justification according to which the acceptability of a particular analytical procedure is established by showing it to be derivable from general principles, . . . theories (or data) that are accepted quite independently of the procedure to be justified. On the other hand, *vindication* may render an analytical method acceptable on the grounds that it leads to accurate predictions (to a degree better than chance) regardless of the details of that method. The rules of induction and deduction are essential to (construct) validation while the relation between means and particular ends provide the basis for (construct) vindication. (Krippendorff, 1969b, p. 12)

In the 1980 edition of *Content Analysis,* I called these two kinds of validity *construct validation* and *construct vindication* (Krippendorff, 1980b). However, because construct validity has a slightly different definition in the psychological test literature, my use of this label caused some confusion, hence the current change in terms. Incidentally, Selltiz, Jahoda, Deutsch, and Cook (1964) call vindication *pragmatic validity,* as "the researcher then does not need to know *why* the test performance is an efficient indication of the characteristics in which he is interested" (p. 157).

Finally, one could be concerned with the validity of the results of a content analysis, criterion-based validity or instrumental validity, and consider two ways of supporting the results:

- Evidence on *correlative validity* ascertains the degree to which the findings obtained by one method correlate with findings obtained by other variables that are considered more valid than the method in question. To be correlated, all variables must be presently and simultaneously available. The result, therefore, is also called *concurrent validity.* Correlative validity requires a demonstration of both *convergent validity,* or high correlation

with measures of the contextual characteristics it claims to indicate, and *discriminant validity,* or low correlation with measures of contextual characteristics it intends to exclude.

- Evidence for *predictive validity* establishes the degree to which the answers of a content analysis accurately anticipate events, identify properties, or describe states of affairs, knowledge of which is absent or did not enter that analysis. Analogous to selecting among the possible answers to a research question, predictions spell out what can be anticipated and what is ruled out. Predictions may concern phenomena that precede, are concurrent to, or follow the texts used in making them.

Sampling Validity 13.2.1

Sampling validity becomes an issue whenever a sample of texts differs from the population of phenomena of interest, not just in size, which often is convenient, but also in composition, which can bias the content analysis of the sampled texts. As already stated, sampling validity is the degree to which a population is accurately represented in the sample. To begin with, two situations need to be distinguished:

(1) The sample consists of a *subset* of members of the population of interest.

(2) The sample consists of *representations* of phenomena that lie outside the sample and the population from which the sample is drawn.

Evidence in the first situation—drawing a sample from the very population of interest—is well understood by means of *statistical sampling theory,* as discussed in Chapter 6. Whether one is interested in a sample's medial, proportional, variational, or distributional accuracy, statistical theory provides measures of the sampling error. This error is a measure of a sample's *in*validity. For purposes of this discussion:

$$\text{Sampling Validity (1)} = 1 - \text{sampling error} = 1 - \frac{\sigma}{\sqrt{N}}\sqrt{\frac{n-N}{n-1}},$$

where σ is the standard deviation of the population, which is a measure of its diversity of categories; n is the size of the population; and N is the size of the sample.

In traditional sampling theory, sampling errors are a function of three factors. First and most important is the sample size, N. The larger the sample, the smaller the sampling error and the larger the sampling validity (1). Second is the diversity of categories within the population, represented by the standard deviation σ. Given two samples of equal size, the sample drawn from a more diverse population has larger sampling errors and is less likely to be valid than a sample drawn from a less diverse population. Third is the proportion of the

population sampled. As samples become more inclusive, $(n - N) \rightarrow 0$, the sampling error shrinks and sampling validity (1) grows.

Evidence in the second situation—drawing samples of representations in view of what they represent—is not easy to obtain, yet it is a frequent concern for content analysts. To be clear, as I noted in Chapter 2, content analysts do not study texts, images, or distributional characteristics for their own sake or to generalize to other texts, images, or distributional characteristics; rather, they use texts as a means to get to what the texts' users have in mind, what the texts are about, what they mean or do and to whom. There rarely exists a one-to-one correspondence between meanings, references, uses, or contents and units of texts. Sampling theory offers no simple test to establish whether a sample of textual units fairly represents the phenomena that a content analyst intends to access through these texts. Additionally, the texts that reach the content analyst usually are *presampled* by others—by organizations following institutionalized rules; by individuals with particular intentions, who highlight some kinds of information and downplay others; or by media of communication that have their own built-in technological or economical filters. Communication researchers have long studied how reality is constructed, represented, and misrepresented in the mass media, but they have rarely used these findings in validation efforts. Concepts such as gatekeeping in news flows, ideological/racial/gender biases in writing, the positive spin that affected parties put on politically embarrassing stories, and the attention paid by institutions of journalism to particular stories (i.e., to whom journalistic institutions grant a voice and, by implication, what or whom they ignore) are well established and often quantified.

If the phenomena of interest need to be accurately represented in the texts that researchers are analyzing, then sampling must undo the biases that result from the selective ways texts are made available. Validating evidence for sampling validity (2) can be of two kinds:

- Knowledge of the population of phenomena with which one of its samples is to be compared
- Knowledge of the self-sampling practices of the source of the available texts

To measure the degree to which a population of phenomena is fairly represented in a sample of textual units, a simple percentage measure that derives from the well-known coefficient of contingency C is useful. In its incarnation as a validity measure, $1 - C^2$, it concerns proportions only and has two versions that correspond to the above two kinds of validating evidence:

$$\text{Sampling Validity (2)} = 1 - C^2 = \frac{1}{1 + \sum_i \frac{(P_i - p_i)^2}{p_i}} = \frac{1}{1 + \sum_i P_i \frac{b_i^2}{1 - b_i}}$$

where P_i is the proportion that is observed to represent phenomena of category i in the sample and p_i is the proportion representing phenomena of category i in

the population from which the sample is drawn. When known or ascertainable, the proportions p_i serve as validating evidence in the first version of the formula for sampling validity (2). When the biases b_i of representing categories i are known, algebraically equivalent to $(1 - p_i/P_i)$, b_i serves as validating evidence in the second version of the formula for sampling validity (2). Studies of biases, assessing or estimating b_i, are more common, easier to conduct, and designed differently than those assessing the proportions p_i, hence the two versions.

The first version of the above-stated sampling validity (2) is a function of the observed proportion P_i in the sample and the proportion p_i in the population, which is the validating evidence and must be obtained independent of the sample. When $P_i = p_i$ for all categories i, sampling validity is unity. It declines with increasing differences between the two proportions.

The second version of sampling validity (2) is a function of the observed proportion P_i in the sample and the self-sampling bias b_i $(= 1 - p_i/P_i)$, which is the extent to which the source of the sampled text over- or underrepresents categories i. Here, the bias b_i serves as validating evidence, which must be obtained independent of the sample as well. If this bias $b_i = 0$ for all categories i, then sampling can proceed as usual. If this bias deviates from zero in either direction, then sampling validity (2) is reduced. If the source biases b_i are known, one can approximate a valid sample either by using the technique of varying probability sampling (Chapter 6, section 6.2.4), which compensates for the known biases, or by transforming the proportions P_i in a biased sample by $P_i' = (1 - b_i)P_i$, which corrects for the biases in representing the phenomena in question.

Semantic Validity 13.2.2

Semantic validity is the degree to which the analytical categories of texts correspond to the meanings these texts have for particular readers or the roles they play within a chosen context. Virtually all content analyses respond to texts according to their meanings: denotations, connotations, insinuations, implications, associations, metaphors, frames, uses, symbolic qualities, and so on. Users of the texts could serve as sources of validating evidence for the categories that a content analysis employs. In older definitions of content analysis, accurate descriptions of these meanings were the only aim mentioned, whether they referred to classifications of sign-vehicles (Janis, 1943/1965), descriptions of the "manifest content of communication" (Berelson, 1952, p. 16; Berelson & Lazarsfeld, 1948, p. 6), coding (Cartwright, 1953, p. 424), or "putting a variety of word patterns into [the categories] of a classification scheme" (Miller, 1951, p. 96). Although it is widely recognized that accurate descriptions of these meanings are the key to the success of content analyses, despite their ultimately inferential aims, what counts as accurate and particularly whose meanings are taken to be valid depend on the chosen context of an analysis.

In Chapter 2, I noted that even analysts involved in purely descriptive efforts must acknowledge a context that they or others could consult to validate those

efforts. If content analysts claimed to describe meanings without reference to any context of specific uses and users—authors, readers, newspaper editors, professional experts, professionals with specialized perspectives, social institutions, standard thesauri or dictionaries, even analysts' own discourse communities—there would be no way to know what could validate or invalidate these descriptions, and analysts would be left to appeal to face validity or to play on their scientific (social or political) authority. Although semantic validity is an issue that most content analysts take seriously, it is rarely formally tested.

It is easy for researchers to take an objectivist stance and consider meanings as universal and as defined in general dictionaries, or to take an ethnographic stance and delegate decisions on meanings to the authors of given texts. However, both of these extremes deny the fact that all descriptions simplify or abstract, usually in the interest of the describer's questions. Categories are always more general than the objects they categorize. In fact, content analysts rarely take the unique meanings of the analyzed texts as the basis for their inferences; instead, they operate on a level of abstraction above that of ordinary talk. Concepts such as speech acts, monologue, self-esteem, ethnic prejudices, sexual harassment, and libel, as well as such distinctions as between pro- and antisocial behavior, are all fairly abstract and not necessarily shared with the sources of texts being analyzed. Distinctions among the functions of political campaign discourse in terms of acclaiming, attacking, and defending (Benoit, Blaney, & Pier, 1998) are analytically useful but may not help political candidates to organize their campaigns. It is the use of abstract categories that makes semantic validation the content analytic analogue to content validation of psychological tests. For example, the content validity of a test designed to determine the aptitude of a job candidate for a particular kind of employment is the extent to which the test includes all demands of that job, not just a few outstanding qualifications. Analogously, the semantic validity of the categories "acclaiming," "attacking," and "defending" should be the extent to which these categories embrace all functions of political campaign discourse and clearly distinguish among the three categories.

The preparations for an analysis of values in political documents may serve as an example of an iterative use of semantic validity criteria. In this study, we started with a collection of what a panel of political scientists could easily identify as "value-laden statements" of the kinds we would find in the documents to be examined for the values their authors expressed. To reproduce these experts' distinctions, we formulated explicit recording instructions. The coders varied greatly in their ability to make the same distinctions, and a computer program we had hoped to employ turned out to be virtually useless. The whole history of this effort is too long to relate, but we began by developing a list of political values the documents contained—democracy, freedom, progress, and the like—and allowed others to be added. This turned out to be far from satisfactory. A large number of value-laden statements contained none of the values on our list. We then looked into various modes of reasoning that led us to implicit goals, preferences for processes, criteria for decision making, and so on, added them to our

emerging instructions, and reapplied them to our collection. Slowly, we narrowed the gap between the distinctions that our instructions suggested between value-laden and value-neutral statements and those that our panel of experts made (Krippendorff, 1970c). One might question our use of experts as providing the evidence for our semantic validation efforts. We could have used another population for reference, but because we were working under typical time constraints and our analysis was to make a contribution to the political science literature, the theoretical motivations seemed to justify the choice of this context, and we were satisfied that the instrument reasonably approximated the consensus reached by those we trusted to know what they were talking about.

Semantic validity arguments come in numerous guises. To validate the dictionary entries of the General Inquirer, a computer program for tagging texts and then analyzing the tags, Dunphy used a KWIC (keyword in context) list to explore the various meanings of tagged words. For example, the keyword *play* (see Figure 12.1) showed numerous senses that were not initially anticipated and that would have been confused by computer programs that tagged only single words (Dunphy, 1966, p. 159). To address such semantic problems, Stone, Dunphy, Smith, and Ogilvie (1966) developed so-called disambiguation procedures that looked into the linguistic environments of homonyms for clues to how their meanings could be identified more correctly. These rules improved the Inquirer's ability to distinguish between textual units according to how they would be read by ordinary English readers. In effect, these researchers' efforts were aimed at improving the semantic validity of the computer tagging of text.

Consider the distinction between *self* and *other* references made by a dictionary approach to computer text analysis (see Chapter 12, section 12.5.1). Suppose the words *self, I, my, myself, me, mine, we, our, ourselves,* and *us* are tagged *self,* and the words *other, you, your, yourself, he, his, him, himself, she, hers, her, herself, them,* and *themselves* are tagged *other.* Suppose sentences are the units of analysis. These two tags would identify two subsets of sentences within the set of all sentences in the sample. One could obtain evidence for the semantic validity of these dictionary entries by asking competent readers, working independently, to classify the sampled sentences into those that speak about the "self" of the authors and those that speak about what these authors consider "others." Figure 13.2 depicts two sets of tagged sentences surrounded by dashed lines and the sets that would serve as validating evidence within solid lines. The degree to which these two kinds of sets overlap is a qualitative indication of the semantic validity of the two tags that this dictionary is using. In computer-generated categorizations of text, the semantic validity is rarely as good as the figure suggests, but we are interested here only in what is meant by semantic validity: the complete overlap between a classification of uncertain validity with one we have reasons to trust.

A more traditional way for scholars to challenge the semantic validity of a content analysis is by creating counterexamples. This strategy is well established in linguistics, where as soon as one linguist makes a claim that a proposed grammar accounts for all grammatically correct sentences, another comes up with

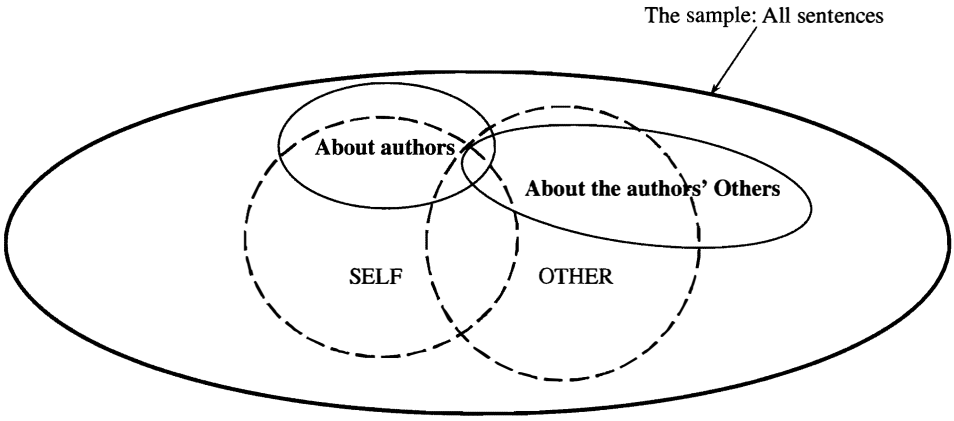


Figure 13.2 A Qualitative Illustration of the Semantic Validity of Two Categories

examples of sentences that this grammar will misidentify. Closer to content analysis, in the critique of his own contingency analysis, Osgood (1959, pp. 73–77) employs this strategy to discount exaggerated claims of what his analysis does. According to Osgood's contingency analysis, the following statement would be counted as an incident of association between *love* and *mother:*

1. I love my mother.

But so would these statements:

2. I loved my mother more than anyone else.
3. Mother loved me.
4. I don't love my mother.
5. Have I always loved my mother?—Hell no!
6. My beloved father hated mother.

Because *love* and *mother* co-occur in all six statements, contingency analysis would group them into the same equivalence class of *love-mother* co-occurrences. However, relative to statement 1, statement 2 shows contingency analysis to be insensitive to verbal qualifications of an expressed association—"more than anyone else" is not a frequency. Statement 3 shows this analysis to be unable to distinguish between active and passive constructions—that is, the target of love. Statement 4 shows it to be insensitive to negation, statement 5 shows it to be insensitive to irony, and statement 6 shows it to be insensitive to grammatical constructions. Osgood did not use these observations to argue against his method, of course; rather, he used them to clarify what it registers: statistical, not logical, associations among pairs of concepts, presumably in someone's mind. Here, even denying a relationship between two concepts would be considered evidence that they have something to do with each other.

A content analysis in a legal context may serve as a final example. In a libel case recorded at a Texas court as *Wood, et al. v. Andrews, et al.* (Dallas County, Cause No. 94–8888-C, 1997), the plaintiffs, 20 psychiatrists who worked for a mental health institution, hired a content analyst to establish objectively the libelous and defamatory nature of publicity they had received in the press. That publicity was attributed largely to one of the defendants, a lawyer who represented a number of former patients of the hospital where the plaintiffs worked (one of whom was a codefendant) who were suing that institution for patient abuse, malpractice, illegal profiteering, and insurance fraud. This content analyst retrieved all 52 newspaper articles published between 1993 and 1996 that contained references to that mental institution; 36 of these articles mentioned the lawyer defendant and 16 did not. She examined the 36 articles that mentioned the defendant and used for comparison the articles that did not. She then unitized all of the articles into 970 assertions, assigning each to one of 16 categories that emerged from her reading of the articles, focusing on kinds of bad publicity. She then drew her conclusions from a statistic of these categorizations.

In response, the defendant hired an expert to examine the content analyst's research and findings. The expert raised the issue of the semantic validity of the categories the analyst had used to make her point. Although she had conducted her analysis carefully, with categories traceable to the original assertions or their rephrases, her conclusions were irrelevant because her categories ignored the context that mattered in the case: the terms of the law, the legal definition of libel. Texas law defines libel in fairly specific terms. To be considered libelous, an assertion has to meet the following criteria:

i. Made with the intent to harm a person publicly or financially or in disregard of injurious consequences to another person

ii. Knowingly untrue

iii. Read and understood as stating facts

iv. Causing its readers to alter their speaking in ways that blacken a person's public image, impeaches that person's honesty, integrity, virtue, or reputation, and

v. Actually incurring financial injury to that person or expose that person to debilitating hatred, contempt, and ridicule.

In other words, assertions in the category of libel have to be (i) made with the intent to harm, (ii) known to be untrue, and (iii) read as stating facts. Evidence on criteria iv and v would presumably require observations or testimony. The content analyst's categories traced bad publicity about the plaintiffs to the defendant but failed to provide answers in the legally required categories. For example, accusations of insurance fraud, if true, are not libelous, regardless of how often they are mentioned. And assertions critical of the plaintiffs may not have been made with the intent to harm. In this context, a semantically valid content

analysis would have to let the articles answer the questions in categories to which applicable law could apply—not how an average reader might interpret the newspaper articles. One could conclude that this analyst's categories had no semantic validity in the prescribed context.

Semantic validity acknowledges that recording units, when placed in one category, may differ in all kinds of ways, but not regarding the meanings that are relevant to the analysis, and units that turn up in different categories must differ in relevant meanings. The emphasis on relevant meanings is important, as text interpretations can be endless, whereas content analysts are concerned only with specific research questions. In the above examples, deviations from this ideal signaled that the procedures of an analysis needed to be altered, that the categories needed to be redefined, or that the findings should be discounted.

I will now state a simple measure of the semantic validity of categorizations and then show how it can also be applied in evaluations of the semantic validity of text searches. To begin, it is important to recognize that assigning the units of a sample to any one of several mutually exclusive categories amounts to partitioning that sample into mutually exclusive sets of units. The semantic validity of one method is established through the comparison of its partition with the partition obtained by another method that serves as validating evidence. Ideally, these partitions are identical, but in practice they rarely are. A minimal measure of the semantic validity of categorizations can be defined in these terms: Let

j denote one of a set of categories of analysis, 1, 2, 3, . . . ;

n be the size of the sample of textual units being categorized in two ways;

$A_1, A_2, A_3, \ldots A_j, \ldots$ be mutually exclusive sets of units distinguished by the method in question;

$E_1, E_2, E_3, \ldots E_j, \ldots$ be the validating evidence, the mutually exclusive sets of units distinguished by another method that is considered valid;

$\cap$ be the intersection of two sets, denoting the units common to both (AND in Boolean terms); and

be an operator that enumerates (provides a count of) the members of a set.

In these terms, when the two partitions are identical, all A_j and E_j contain the same units, $A_j = E_j = A_j \cap Ej$ for all categories j, then the measure should become unity, indicating that semantic validity is perfect. Deviations from this ideal should produce values less than unity. A measure that satisfies these requirements is

$$\text{Semantic Validity} = \Sigma\#(Aj \cap Ej)/n.$$

One can apply more sophisticated statistics, Cohen's (1960), for example, or a coefficient that would extend the measure to different metrics or allow overlapping sets, as in Figure 13.2. However, we are interested here only in the simplest approach to semantic validity.

Regarding the evaluation of the semantic validity of text searches, recall from Chapter 12 that searching a textual database for relevant units of text starts with the formulation of a suitable query. Formulating such a query involves considerable linguistic insight, largely because the population of texts within which a query searches for matching character strings is different from the population of meanings that are represented in the searched texts. A semantically valid query will identify all and only those units of text, or documents, that are relevant. A query may fail to identify documents that are relevant to a research question or may fail to exclude documents that are irrelevant to that question. Relevance, it should be kept in mind, is an attribution made by content analysts based on their understanding of the purpose of a research project. Search results, in contrast, stem from the matching of character strings with a given query.

In the technical literature on information retrieval, which concerns largely whether particular queries fail to retrieve documents that do contain the desired character strings or retrieve documents without matches, scholars have reported on the use of two measures for assessing the quality of search engines: precision and recall. *Precision* is the degree to which a search engine lists documents that match a query. *Recall* is the degree to which a search engine returns all the matching documents of a collection (Rijsbergen, 1979).

Technical failures can affect the semantic validity of text searches, but here we are concerned with comparing the results of an electronic text search (retrieved or not retrieved units of text) with the validating evidence obtained by human judgment (relevant or irrelevant units of text). In effect, these define two bipartitions of the textual universe, which can be represented numerically in the form of a fourfold table of frequencies:

Units of Text	**Relevant**	**Irrelevant**	
Retrieved	*a* Correct inclulsions	*b* Commissions	*a* + *b*
Not Retrieved	*c* Omissions	*d* Correct exclusions	*c* + *d*
	a + c	*b* + *d*	*n*

In this table, n is the size of the textual universe searched. Applying the above-stated semantic validity measure to this far simpler situation yields

$$\text{Semantic Validity} = (a + d)/n.$$

Two errors distract from the semantic validity of text searches. The first is the Error of Commission = $b/(a + b)$ [or 1–Precision],

which is the proportion of the number b of irrelevant units of text that were mistakenly retrieved to the total number $(a + b)$ of retrieved units. In a search of articles containing self-references in the press, Bermejo (1997) found this error to be 16%, which is remarkably good. The other error is the

$$\text{Error of Omission} = c/(a + c) \text{ [or 1–Recall]},$$

which is the proportion of the number c of relevant units that the search failed to identify to the total number $(a + c)$ of relevant units in a textual universe. In a pilot study that involved retrieving articles on domestic violence from three newspapers, Wray and Hornik (1998) found errors of commission of 10%, 19%, and 29% and errors of omission of 12%, 20%, and 25%, although they cast their finding in terms of precision and recall. How the two measures reduce the semantic validity of text searches can be seen in this equation:

$$\begin{aligned}\text{Semantic Validity} = 1 &- (a + b)/n \text{ Error of Commission} \\ &- (a + c)/n \text{ Error of Omission.}\end{aligned}$$

In typical text searches, these two errors are rarely of equal significance, however. When a search result contains the answer to a question directly—that is, without further analysis—both errors weigh equally and the single measure of semantic validity is appropriate. But when a search aims at identifying documents for further examination, errors of commission merely create more work for coders, who usually do not have any problem eliminating irrelevant documents after reading, whereas errors of omission deprive content analysts of relevant data that could lead to different conclusions—hence the need to account for these errors separately.

An epistemological problem in assessing the semantic validity of text searches is that cells c and d are typically unknown. In fact, one cannot measure errors of omission unless one finds a way to examine or at least to estimate the number of unretrieved documents and the proportion of correct exclusions. For limited textual databases, the size n of the available textual universe may well be known, at least by approximation. Unfortunately, the size n of very large databases may be too large to yield meaningful calculations. However, such limitations do not apply to the more common semantic validations of content analysis categories for which samples tend to be finite and manageable in size.

13.2.3 Structural Validity

Structural validity is at issue when content analysts argue over whether the analytical constructs they have adopted accurately represent the known uses of available texts, the stable meanings, language habits, signifying practices, and behaviors in the chosen context. Thus structural validity assesses the backing

(see Chapter 2, section 2.4) of an analyst's abductive inferences primarily from categorized text and secondarily in processes of categorizations, provided the latter involves coders that serve as a backing for inferences or interpretations implicit in the coding/recording process. This evidence may consist of unquestionable incidences of the stable relationships between potentially available texts and the targets of content analysis and valid theories about them. When a content analysis is designed de novo, and thus has no history of successes or failures, structural validation is the only way to lend credibility to its inferences.

The work of historians is most clearly of this kind. Although it is said that history never repeats itself, it may well repeat certain patterns that can be accounted for through generalizations, especially about human/social nature. For historians to rely on such patterns, they must be conceived as relatively permanent within a particular historical context. Dibble (1963), who analyzed arguments by historians in support of and against inferences about the factual nature of events drawn from historical documents, distinguished four kinds of evidence or generalizations for the structural validity of historical accounts. One kind of evidence concerns the roles and practices of the social institutions that create the records to be validated, using their own codes of conduct and preserving certain documents and not others. These are sociological generalizations about what Dibble calls "social bookkeeping" practices. A second kind concerns the characteristics of witnesses who describe what they experienced or report on what they heard. These are psychological generalizations about the working of memory, the influence of interests, and the influence of emotional or ideological involvement with the events in question. A third kind of evidence concerns how the structure of narratives relates to the narrated events. These are linguistic or literary generalizations about how texts are organized and what they mean to readers at that time. Finally, there are physical generalizations of how documents tend to travel, who accesses or reproduces them, how they reach their destinations, how they are filtered or filed, and how they fade or drop out of circulation. Dibble suggests that historians use such generalizations to validate or invalidate their inferences. They exemplify what content analysts do as well, perhaps a bit more systematically.

The inferences that Leites, Bernaut, and Garthoff (1951) made from speeches delivered on the occasion of Stalin's birthday, discussed in Chapter 9, serve as another particularly transparent example of structural validation in an essentially unique situation. Once the researchers' analytical construct was in place, the inferences from available speeches followed. The validity of their construct was established by experienced Sovietologists who argued by references to generalizations about how political discourse functioned in the Soviet Union, especially how politburo members close to Stalin would have to avoid showing interpersonal closeness. With the structural validity of their construct demonstrated, the results of its application were accepted on this ground—and later proved to be correct.

Osgood, Saporta, and Nunnally (1956) fashioned their evaluative assertion analysis according to then-prevailing theories of affective cognition, cognitive dissonance theory in particular, which had been substantiated in numerous

controlled experiments with subjects. This was the only validating evidence used in this case, and through it the researchers sought to establish that the computations built into their analytical procedure structurally corresponded to what was known about individual cognition. Evaluative assertion analysis has been extended—for example, by Cuilenburg, Kleinnijenhuis, and De Ridder (1986) and Kleinnijenhuis, De Ridder, and Rietberg (1997)—based on the structural validity provided earlier and only occasionally reexamined.

13.2.4 Functional Validity

Functional validity is the degree to which analytical constructs are vindicated in use rather than in structure. A content analysis is vindicated by reference to its history of use, particularly by its absence of significant failures. *Usefulness* and *success* may mean many things, of course, and these concepts make sense only in the presence of alternative methods competing with each other in the same empirical contexts. To vindicate a content analysis, one must demonstrate that its analytical constructs, which account for the analyst's proceeding from available texts to the answers to given research questions, are useful over time and in many empirical situations. Whereas evidence for structural validity is based on a correspondence between what one knows about a context and how that knowledge is built into the analytical procedure, functional validity is grounded in whether or not or how well it works.

Functional validity has long been recognized, although it has been known by different names. Janis (1943/1965) suggested that because meanings unobservably mediate between texts (or "signs," as he preferred to call them) and observable behaviors, one can establish the validity of semantically oriented content analyses only indirectly, by "inferring validity from [their] productivity" (p. 65). He noted, "A content analysis procedure is productive insofar as the results it yields are found to correlate with other variables" (p. 70). In effect, Janis argued that because there is no validating evidence for how audience members understand given messages, accounting for references, attributions, assertions, and speculating about probable effects are justifiable only when the "categories . . . occur as variables in many true empirical propositions" (p. 65). He essentially bypassed answering the question of how or why an analysis produces the results it does as long as they connect with other phenomena that are considered interesting.

An example might be the use of neuronal network theory in the computation of word co-occurrences by the software system CatPac (Woelfel, 1993, 1997). The designers of this system incorporated several ideas into its operation: that concepts are represented by single words, that the textual proximity of pairs of words and the frequency of their co-occurrences affect the way they are stored/recalled in an author's or reader's brain, that the strength of their pairwise relationships is dynamically adjusted with use, that recent co-occurrences overshadow earlier ones, and so on. These propositions, individually convincing,

could be regarded as validating the procedure structurally, and its proponents claim as much when they call CatPac a system that performs a "semantic network" analysis. However, the computations that yield results are so complex, and so little is known about how the human brain develops concepts, that the connection between how people conceptualize and how the computational procedure gets to its results remains obscure.

However, CatPac has been used extensively and applied to a variety of data by researchers interested in mass communication, marketing, politics, bibliographic citations, and many more areas. Improvements have been introduced over time. Occasional lacks of face validity caused the developers to make a variety of adjustments, such as excluding function words and stemming, lemmatizing the vocabulary to reduce grammatical variations considered meaningless in the context of the system's use. CatPac applications naturally migrated into areas that seemed most promising, particularly where the results correlated with other phenomena of interest or aided practical decisions. It found niches in which it proved itself of practical value, useful, and successful. Does CatPac really compute semantic networks? Not the way linguists and researchers in the artificial intelligence community conceptualize them. Does it replicate what neurons do in the brain? Surely not structurally. However, the very fact that it finds users and uses in competition with other computational techniques can be regarded as vindicating evidence, demonstrating its functional validity.

Correlative Validity 13.2.5

It is an epistemological fact (but a fact not always recognized in the literature of psychological testing) that correlations between test results and other variables relate measures of phenomena to each other, not phenomena. Correlations cannot bridge the epistemological gap between measures and the phenomena they claim to measure. Correlations do not predict, either—as I shall demonstrate below. They merely weigh the extent to which one measure can substitute for another. If the measures in a battery of measures are correlated with each other, those that correlate perfectly are substitutable without question, and those that correlate less than perfectly are substitutable to the degree of their correlation. The basic idea of correlational validity is that validity travels along high correlations. Validity always comes from somewhere—one or more trusted variables whose validity is established prior to or outside of efforts to establish correlational validity. If the results of a content analysis of crime reports in newspapers and the results of public opinion polls about perceived crime correlate higher with each other than either of the two variables with official crime statistics, as found by Zucker (1978), then content analysis and public opinion polling might well be substitutable for each other, but neither can replace crime statistics. If none of these variables can be trusted to begin with, validity cannot be an issue. Figure 13.3 depicts the general idea of correlational validity schematically.

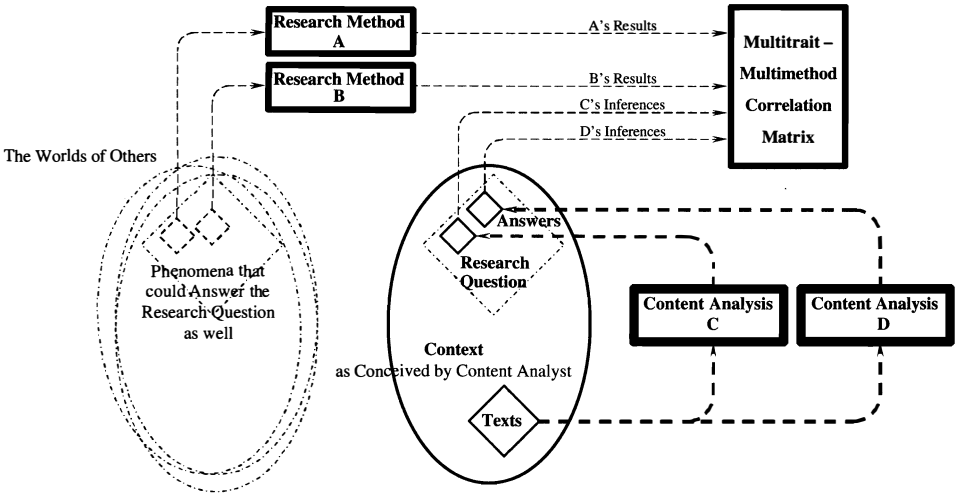


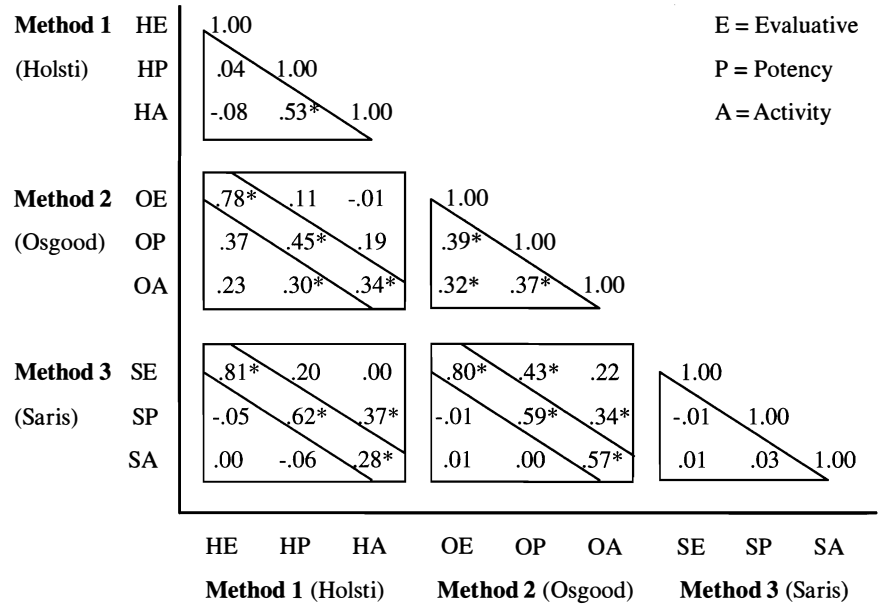
Figure 13.3 Correlative Validation

Campbell and Fiske (1959) develop the idea of validation by correlation statistics into a full-fledged methodology. Taking Popper's idea of falsification to heart, they argue that correlative validity of a new method requires not only high correlation with established measures of the trait it intends to measure but also low or zero correlation with established measures of traits it intends to distinguish. They call the former *convergent validity* and the latter *discriminant validity*. It follows that a research result can fail to be correlatively valid in two ways: by low correlations with measures that are known to measure the phenomena of interest and by high correlations with measures that are known to measure distinctly different phenomena or phenomena independent of the one intended to be measured.

To show that a measure possesses both convergent and discriminant validity, one must compute a battery of correlation coefficients between measures of a number of traits, each obtained by several independent methods. These are tabulated in what Campbell and Fiske call a multitrait-multimethod matrix (see also Alwin, 1974). A detailed discussion of this method is beyond the scope of this chapter, but Table 13.1 provides an example of such a matrix (for more on this method, see Krippendorff, 1980b). It compares three computer implementations of Osgood's main semantic differential scales of affective meaning, evaluative (E), potency (P), and activity (A) (see Chapter 7, section 7.4.4), by three researchers, Holsti (H), Osgood (O), and Saris (S) (see Saris-Gallhofer & Morton, 1978).

In this table, all correlations with themselves are listed on the main diagonal. These are unities, of course, and uninformative as such. Convergent validity would be demonstrated by high correlations, ideally unities, in the diagonals of

Table 13.1 Multitrait-Multimethod Matrix for Three Content Analyses



the three heteromethod blocks. In these submatrices, they are all significant, as indicated by asterisks, but they differ regarding the affects being compared. The measures of the evaluative dimension correlate highest with each other, followed by the measures of potency; the measures of the activity dimension are lowest in all three instances. This is consistent with the findings of other research, suggesting that judgments of good and bad—heroes and villains, successes and failures, beauty and ugliness—yield better scales, turn out to account for more variance in semantic differential studies than the other two, and are also more reliable in content analyses generally. Discriminant validity is indicated when the off-diagonal correlations in the heteromethod blocks are lower than the correlations in their diagonals, ideally zero. Here they are lower, but still far from the ideal. In fact, the surprisingly significant correlations between OA and HP, OA and SP, and OP and SE suggest that the three dimensions of affective meanings are not clearly differentiated across these methods. However, the culprit in this lack of discrimination is found in the three monomethod triangles. Within Osgood's method, all three dimensions correlate significantly—that is, it does not discriminate too well among the three kinds of meanings. In Holsti's method, it is the correlation between the activity and potency dimensions that signals a lack of discrimination, whereas in Saris's method, off-diagonal correlations are near the ideal of zero, independent of each other, and show high discriminant validity.

This discussion is not intended to generalize about the three methods. They may differ for a variety of reasons (see Saris-Gallhofer & Morton, 1978). My only aim here is to show how convergent and discriminant validity can play out in correlative validations.

Another example is the validation of an index of argument quality by Cappella, Price, and Nir (2002). These researchers developed their index in the course of a study of online deliberations during the U.S. presidential elections in the year 2000. It counts the number of arguments that participants in this study could give in support of their own positions and, what is perhaps more interesting, the number of arguments that these participants could imagine others would have against their positions. Cappella et al. wisely refrain from claiming to measure opinion quality, as a single dimension would probably fail a semantic validity test. The name they give their measure reflects more closely what it actually measures, "argument repertoire." To test its convergent validity, they show that it correlates highly with numerous variables that could be construed as assessing aspects of a common construct, including political knowledge, political interest, flexibility (subjects' willingness to participate in various discussion groups), and mass-media exposure. They note that

> those with the capacity to write out reasons for their opinions and to identify relevant reasons for opposed opinions also express interest in politics, are more accurate in their factual political knowledge, and use the print and broadcast media as sources of their political news. Even their personal communication is more political and diverse. Coupled with . . . data . . . indicating higher argument repertoire for those with more education and more commitment to their ideology and party, [we] have good evidence of convergent validity. (pp. 83–84).

Cappella et al. do not show discriminant validity, however, and it is therefore not so clear what their measure distinguishes. It may well embrace general communication and social skills as well as intelligence, which would be far beyond the intention to define a measure that adds to the vocabulary of public opinion researchers.

13.2.6 Predictive Validity

Predictions extend available knowledge to as yet unobserved domains. The predicted phenomena may have existed somewhere in the past (e.g., historical events, characteristics of late authors, antecedent conditions of received communications), may be concurrent with the texts being analyzed (e.g., attitudes, psychopathologies, individual aptitudes, the extent to which someone is plagued by problems, the makeup of cultural climates), or may occur in near or distant futures (e.g., the consequences of persuasive messages, the success of future employees, the continuations of trends). I am suggesting two defining criteria for predictive validity. The first emphasizes the nature of evidence as in the *Standards* (American Educational Research Association et al., 1999): For correlational validity, validating evidence must be concurrent, whereas for predictive

validity it need not be and in fact typically is not. The second defining criterion requires predictions to be specific, to select a set of observations that is smaller than the set of all conceivable ones—just as any answer to a research question must exclude some of the logically possible answers. Eventually, predictions are validated when the validating evidence stays within the set of predicted observations.

To draw a clear line between correlational and predictive validity, I return to Cappella et al.'s (2002) argument repertoire measure. These researchers found high correlations of argument repertoire not only with variables that belong to the same construct (as noted above), but also with variables that they conceptualized as caused by what argument repertoire measures: participation. They observed two kinds of participation: willingness to attend group deliberations about political topics and willingness to get involved in substantive exchanges while attending (p. 89). Both correlated highly with argument repertoire. Because all data of this study were concurrent and these variables correlated, the validity thereby established is correlational, even if one could argue, and hence conceptualize, that participation is an effect and not a cause.

However, conceptions of causality aside, once researchers state findings so as to be selective among conceivable alternatives and open to the consideration of evidence concerning these alternatives, predictive validation can take place. In fact, when Cappella et al. report that those with greater argument repertoires are more willing to participate in political deliberations, they make rather specific predictions that could be checked against future data. Establishing the validity of their predictions would require agreement with subsequent observations—not correlations, but observations concerning whether and how often people with high argument repertoires do indeed participate in one or both ways.

A classic example of predictive validation is George's (1959a) attempt to evaluate the Federal Communications Commission's predictions made from German domestic propaganda during World War II. All of the FCC analysts' inferences were available in the form of reports they had written. After the war, George was able to match the inferences, one by one, with documents that had then become available. He judged each of the inferences for which validating evidence was available as correct, nearly so, or wrong. He demonstrated that the FCC analysts' predictions were accurate to a degree better than chance. George's research (which was not as simple as this brief description makes it seem) suggests how analysts can bring subsequent evidence to bear on predictions.

To recap: Predictions cannot be validated by correlation. A watch that runs slow correlates highly with standard time but is incorrect nevertheless. Unless one knows the bias of the watch, one cannot tell the correct time with it. The infamous body counts disseminated by the U.S. government through the mass media during the Vietnam War may have correlated highly with military activity, but after a while nobody could trust the exaggerated numbers. To predict the author of an unsigned document, it is not enough to show that signed documents show a correlation between word choices and the identities of authors; the unsigned document must be traced to one author, ideally excluding all others.

Predictions of past, present, or future happenings from texts must also avow exclusions, happenings that are not expected. If a content analysis said yes to all possible answers to a research question, it would be as worthless as if it said no to all of them. The more selective a content analysis is, the more information it provides. Subsequent observations validate predictions when they occur within the set of observations that had been predicted, not outside that set (always or at least to a degree better than chance).

To quantify predictive validity, the measures that are appropriate are the same as those used to assess semantic validity. Both concern valid representations—in the case of semantic validity, of the meanings, referents, or uses of texts; and in the case of predictive validity, of whether the answers to research questions are borne out in fact. The appropriate measure of predictive validity is not correlation but agreement.

Figure 13.4, which is an overlay of Figure 2.1, locates the validation efforts discussed in this chapter within the components of content analysis (as discussed in Chapter 4).

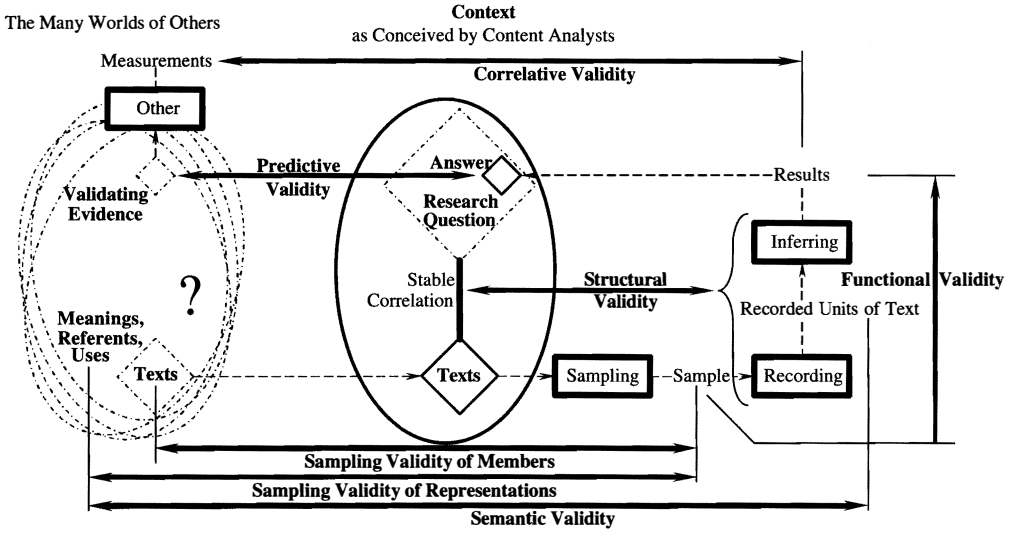


Figure 13.4 Comparisons in Different Kinds of Validity