Klaus Krippendorff

Validity in Content Analysis.

Chapter 3, Pages 69-112 in Ekkehard Mochmann (Ed.)
Computerstrategien für die Kommunikationsanalyse
Frankfurt/New York: Campus, 1980.

Klaus Krippendorff

VALIDITY IN CONTENT ANALYSIS (1)

1. Introduction

Content analysis involves replicable and valid methods for making inferences from observed communications to their context (2). As such, content analysis is at least 75 years old. It emerged during journalistic debates over the quality of the mass media but has now become part of the repertoire of social science research methods. In the course of its evolution one notes not only an increasing level of technical sophistication and scientific rigor, and an increasing spread to different disciplines using any kind of data for analysis, but also a growing social impact of its findings. Let me consider a few examples:

At least since LASWELL's (3) study of foreign propaganda in the U.S. during the early 40's, content analyses have been accepted as evidence in court. Content analyses have been prepared for plagiarism and copy-right infringement cases, and for arguments in the renewal of broadcasting licenses before the U.S. Federal Communication Commission.

GEORGE (4) showed how inferences from foreign domestic broadcasts aided political and military decision making during World War II. Since then expensive monitoring agencies and elaborate information systems are maintained by various governmental agencies to gather and extract intelligence from the mass media of informationally non-penetrable countries, of China (5) for example. Similarly, SINGER (6) discussed the use of content analysis techniques for monitoring the adherence of foreign powers to the nuclear test ban treaty.

Although the idea of monitoring the "symbolic climate" of a culture is not new (7) content analyses have again been suggested as ways of assessing changes in "the quality of life" (8) and to establish cultural indicators (9). With public policy implications in mind, the U.S. Surgeon General commissioned a large scale content analysis of violence on television (10) and a recent U.S. Senate hearing considered additional quantitative accounts of televised violence.

Though of less public concern but with serious consequences to individuals is the increased use of content analysis for various diagnostic purposes, for example, for identifying psychopathologies from a patient's verbal records or for selecting jurors on the basis of their free answers to test questions (11).

Obviously, many content analyses are undertaken to satisfy scholarly curiosities only, and issues that are as important as war and peace are not decided on the basis of content analyses alone. But the increased use of such methods for preparing social actions does affect the public life, the social wellbeing and the mental health of more and more people, hence errors in content inferences become more and more noticeable and costly.

Decision makers are not entirely to blame for their frequent inability to judge the evidential status of content analytic findings. And scientific institutions do not encourage that scholars pay in cash for the consequences of misleading research results. But it is entirely within reason to demand that, whenever a content analysis provides the basis of public policy decisions or feeds into any kind of social practice, evidence about the applicability and validity of the method used must accompany its findings.

As obvious as it seems, this demand is rarely met. Content analysts are notorious for accepting their results by face validity, i.e., on account of the consistency of findings with intuitions existing concurrently with the analysis (12). If results are judged in this manner only, non-obvious findings are likely to be rejected even though they may be correct and obvious findings tend to be accepted even when wrong. In this regard, the methodological status of content analysis must be likened to that of psychological and educational testing some fifty years ago. At that time, psychologists recognized the need for validating their measuring instruments but lacked agreement on standards. It was not until 1954 that the American Psychological Association published its Technical Recommendations for Psychological Tests and Diagnostic Techniques which discourage the use of the general term validity - unless it is clear from the context - and suggests instead to refer to the kind of information used to establish what a test actually measures. Accordingly, the Technical Recommendations distinguish between several types of validity which have been elaborated and refined in various subsequent publications (13).

However, owing to marked differences in methods and aims, the validation of content analyses poses somewhat different problems; and concepts acquired in psychological testing are not simply extendable to the former's domain. For

example, in psychological testing, individuals constitute the natural units of analysis. In content analysis no such convenient units exist. Messages are almost always embedded and derive their meanings from the context of other messages. Or, inferences from psychological tests tend to be based on known distributions of traits over a population of individuals and concern an indiviual's relative position regarding that distribution. Content analyses are more often unique and designed on and ad hoc basis, with statistical techniques employed to aggregate large volumes of data into inferentially productive representations. While individuals are generally available for the possible validation of a psychological test, it is the reason d'etre of content analysis that the sources of their data are only partially observable (14).

Perhaps it is because of these methodological obstacles that most texts in content analysis avoid systematic treatments of validity or follow at best the Technical Recommendations. Notable exceptions are JANIS' (15) chapter in LASSWELL et al.'s book, GEORGE's (16) post facto evaluation of predictions made during World War II and a few specific studies like STEWART's (17) attempt to validate measures of importance, FLESCH's (18) attempt to establish readability yard sticks, HOLSTI and NORTH's (19) attempt to validate inferences regarding the emotional state of the Kaiser in 1914, etc., most of which preceeded the Technical Recommendations in time. There have been no recent systematic attempts to cope with problems of validity in content analysis.

For these reasons, a more systematic presentation of types of validity in content analysis is timely and important. It could provide users of the method with a terminology for talking about the quality of findings and ultimately with a way of assessing whether, to what extent, and on which grounds the results of a content analysis are to be accepted or rejected as evidence.

2. A Typology for Validation Efforts

Generally, "validity" designates that quality which compels one to accept scientific results as evidence. Its closest relative is "empirical truth". As such, this definition is too broad to be useful and finer differentiations are called for.

Following CAMPBELL (20) I will distinguish between internal and external validity. Internal validity is best designated by the term "reliability" while external validity may be considered "validity" proper. When assessing the reliability of a method of analysis one assesses the degree to which variations in results reflect true variations in data as opposed to extraneous variations stemming from the circum-

stances of the analysis. Examples of extraneous variations that may reduce the reliability of a method are ambiguous recording instructions, observer's fatigue, changes in scale, punching and computing errors. Obviously, reliability is a prerequisite but no guarantee of achieving valid research results. Any content analysis must assure a high level of reliability and, in the absence of hard evidence about the validity of findings, information about the reliability of the methods used should be an indispensible part of any research report.

Three types of reliability may be distinguished: stability, reproducibility, and accuracy. Stability measures the degree to which a method of analysis yields identical results when applied to the same data at different points in time. Reproducibility measures the extent of agreement between the results of different methods that follow the same principles of construction (e.g. a common recording instructions to different coders) and are applied to the same data. And, accuracy measures the correspondence of the performance of a method with a given or known standard. Both, stability and reproducibility contribute to the replicability that a content analysis requires, the former and weaker notion by assuring that a method does not change over time (intra-individual consistency) and the latter and stronger notion by assuring that a method is communicable among researchers (inter-coder agreement). Having discussed these distinctions elsewhere (21) I am stating them here only for the sake of completing the typology. The main focus of this section is on types of validity proper.

Validity proper may be distinguished according to the kind of information utilized in the process of validation. Data oriented validity requires validating information about the way data are generated by a source. Information about the semantics of the indigenous symbolic qualities leads to considerations of semantical validity and information about the processes that bring the data into the hands of the analyst lead to sampling validity. Process oriented validity relies on information about the empirical connection between available data and what is intended to be inferred about the otherwise inaccessible context of these data. I speak of construct validity here because it validates the procedure as a whole relative to the system under consideration. Product oriented validity or pragmatical validity relies on information about what the analytical results claim. Depending on how these claims are compared with available evidence, we speak of correlational validity or predictive validity.

I will define these types briefly, relate them to the Technical Recommendations and other pertinent work and then proceed to discuss them in detail.

Data oriented validity assesses how well a method of analysis accounts for the information inherent in available data. It justifies the initial steps of a content analysis from knowledge about the source's idiosyncracies in making that information available for analysis.

Semantical validity is the degree to which a method is sensitive to relevant semantical distinctions in the data being analysed. It is the degree to which a content analysis recognizes and correctly represents the symbolic qualities, meanings and conceptualizations in the system of interest.

Sampling validity is the degree to which a collection of data are either statistically representative of a given universe or in some specific respect similar to another sample from the same universe so that the sample can be analysed in place of the universe of interest. In content analysis, it is the degree to which the collection of data contains with a minimum of bias a maximum of relevant information about the universe, correcting particularly for the bias in their selective availability.

Product oriented validity or pragmatical validity assesses how well a method "works" under a variety of circumstances. It justifies the results of a content analysis from past predictive or correlational successes without references to the structure of the underlying process.

Correlational validity is the degree to which findings obtained by one method correlate with findings obtained by another and justifies in a sense their substitutability. Here, correlational validity means high correlations between the inferences provided by a content analysis and other measures of the same phenomena (convergent validity) and low correlations between such inferences and measures of different phenomena (discriminant validity) in the context of available data.

Predictive validity is the degree to which predictions obtained by a method agree with directly observable facts. In content analysis, predictive validity requires both high agreement between what these inferences claim and the (past, present or future) states, attributes, events or properties in the context of interest and low agreement between inferences and contextual phenomena excluded by these claims.

Process oriented validity or construct validity is the degree to which the inferences of a content analysis must be accepted as evidence because of the demonstrated structural correspondence of the process and categories of analysis with accepted theories or models of the source.

The distinction between predictive and correlational validity corresponds to JANIS' (22) distinction between direct and indirect methods of validation respectively. A direct method of validation involves showing that the results of a content analysis describe what they purport to describe. According to JANIS, since the meanings of messages mediate between perceptions and responses and are not as such observable, the indirect method of validating a semantical content analysis "consists of inferring validity from productivity" (23) and "a content analysis procedure is productive insofar as the results it yields are found to be correlated with other variables" (24).

In the Technical Recommendations the distinction between predictive and concurrent validity depends on whether a test leads to inferences about an individual's future performance or about his present status on some coexisting variable external to the test. In the former case, inferences are confirmed by evidence at some time after the test is administered, in the latter case by evidence exising concurrently. In both cases, the Technical Recommendations suggest that findings and criterion variables be shown to correlate with each other. To me, the time dimension appears secondary to the method used to relate analytical results with validating information, hence, our typology does not distinguish the two types and subsumes both under correlational validity.

The Technical Recommendations' "content validity", which is established "by showing how well ... the test samples the class of situations or subject matter about which conclusions are to be drawn" (25), is identical to "sampling validity" as defined above. The choice of different labels is motivated merely by the possible confusion that the term "content" may precipitate in the context of this paper and by the specific demands that content analyses impose on sampling which tend to be absent in psychological test situations.

The term "pragmatical validity" has been taken from SELLTIZ et al. who add to their definition that "(the researcher then) does not need to know why the test performance is an efficient indicator of the characteristic in which he is interested" (26). The distinction between construct validity and pragmatical validity has also been implied by differentiating between two types of justification that FIEGL (27) termed 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 or theories 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 validation while the relation between means and particular ends provide the basis for vindication (28).

The typology proposed here is presented graphically in Figure 1 together with the distinctions made in the Technical Recommendations. It might be noted that this typology does not include the term face validity mentioned earlier because this form of accepting analytical results as evidence does not require any method of testing and is entirely governed by intuition. While intuition cannot be ignored in any step of a content analysis, it defies systematic accounts by definition of the term. The following concerns only validity proper.

3. General Considerations

In this section I wish to make four points that apply to validity in content analysis generally. First, validation is essentially a process that confers validity to a method from evidence obtained by independent means. Second, the proposed types of validity are not to be considered substitutable alternatives. Third, validation presupposes variability in the method and in the evidence brought to bear upon that method. Fourth, validation tabs only one of two kinds of errors to which content analysis is susceptible. Let me take up these points one by one:

First, validation is essentially a process of justifying the transfer of validity from established theories, from research findings that one knows to be true, or from processes that actually exist to other theories, findings or processes whose validity is in doubt. Thus, if one finds that the proportion of foreign words combined with a measure for sentence length and punctuation correlates highly with observed reading ease, then one might be justified to call it a readability index. Here the transfer of validity is accomplished by the empirical demonstration of a correlation and thus establishes the correlational validity of the index. Other bases for justifying such transfers are agreements which predictive validity requires, logical deductions which construct validity requires, statistical representation on which sampling validity rests and similarity in partition along which semantic validity is transferred.

In the tradition of psychological testing, validating information is largely obtained in the form of experimental evidence such as in the hypothetical case of the above readability index, or in OSGOOD's (29) validation attempts

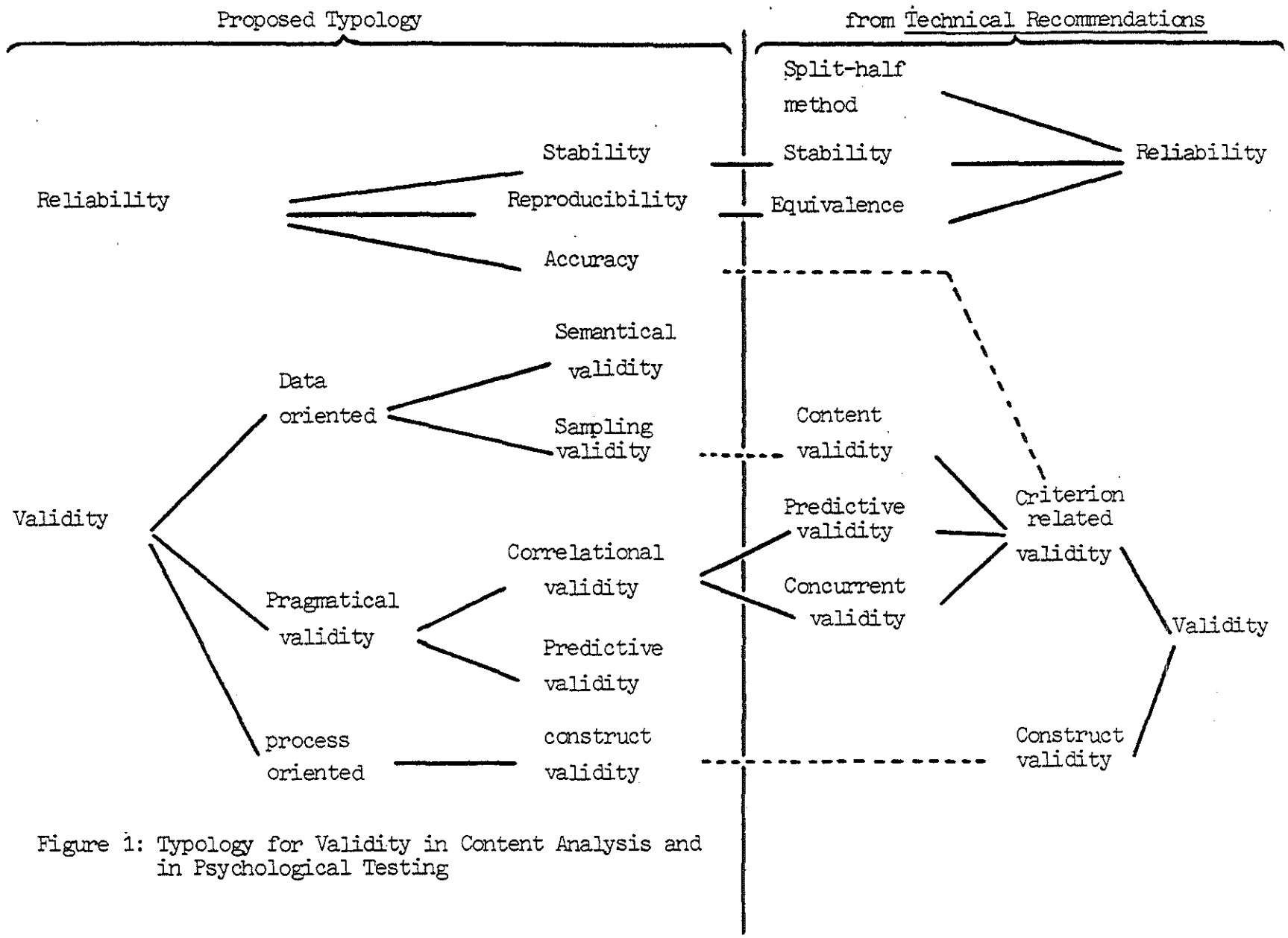


Figure 1: Typology for Validity in Content Analysis and in Psychological Testing

of contingency analysis which involved special experiments that showed contingencies in text to be causally linked to associations in audience member's cognition. In content analysis this is not always possible. Historians use corroborating documents to validate their findings just as GEORGE (30) did to evaluate predictions derived from domestic propaganda in World War II. Construct validity - as will be shown - relies heavily on established theories, tested hypotheses or other undisputable knowledge about the source. And in the absence of hard evidence, content analysts have often resorted to validating their findings against the aggregated judgements by experts on the subject. Most inferences of a patient's psychopathologies are validated in this way partly because there is no other hard evidence against which research findings may be tested.

The validity transferrable in the process of validation is absolutely limited by the validity of the information that can be brought to bear on the situation. Experts can err too and all the more so when they are a closely knit and highly idiosyncratic group, when they have interests in the outcome, or when their values are at stake. But the use of experts is still better than the mere reliance on a single researcher's intuition (face validity) or on the prestige of the person who claims to have the truth. Naturally, the harder the evidence the more validity may be conferred upon a method.

Second, all inferences from content analysis should be valid for the right reasons and for all intended applications. An ideally valid content analysis therefore simultaneously meets all validity criteria. Naturally, this is hardly achievable in practice. The complexity of the world of symbols and communications is not alone to blame for it. Deviations from this ideal result largely from the scarcity of validating information available. Thus, the proposed typology is intended to reflect the kind of information that may become available to a content analyst.

I suppose, one could construct a scale for the validating strength of validating information and consequently for types of validity. For example, where good theories about a source are not yet available, high pragmatic validity may nevertheless suggest that there is in the analytical procedure that corresponds to the structure of the context of the data which may function as a weak theory about the system of interest. Hence, construct validity is potentially stronger than pragmatical validity because the former implies the latter. An analytical construct also implies a semantical mapping which must be valid if the construct is. In these comparisons construct validity turns out to be the strongest form of validity that content analysis can satisfy before predictive validity can be demonstrated. The proposed typology does therefore not provide substitutable

alternatives rather it provides a battery of tests for making use of validating information that may become available, however, incomplete and partial this may be.

Third, validation requires variability of both the method to be validated and the data from which validity flows. The fact that a broken watch shows correct time every twelve hours provides no justification for the transfer of validity precisely because variability is absent. In psychological testing, variability is assured by relying on many units of analysis, on a large sample of individuals, on a battery of tests, or on different stimuli and responses, all of which are expected to yield statistical distributions. If they would not, nothing significant could be said about them.

By analogy, a content analysis that is designed ad hoc (without the wider context of its application and without a history of its use) and results in a singular finding (a frequency, a profile, a point in a semantic space, a correlation coefficient, etc.) is much like a test that is designed to be applied to only one specific individual. Unless content analysis results are shown to vary under different circumstances, little can be said about their potential validity. This fundamental fact is all the more disheartening as most content analyses are indeed tailored to unique situations, are applied only once and then forgotten, leaving their findings as weak and uncertain as they have come about. POOL (31) and HOLSTI's (32) observation that attempts to standardize categories have by and large failed is probably based on the fear that the absence of observed variability in content analysis brings their validity in to serious question.

However, there are a few approaches to validation in content analysis that are essantially unique in construction. The first is to make use of variability where it is available, namely in the initial steps of a content analysis, leaving unjustified only the final steps at which data are condensed into a single figure. Data oriented validation procedures - semantical validity and sampling validity - are cases in point which allow many data reduction techniques (e.g. sampling, clustering, multidimensional scaling, factor analysis) to be validated. A second approach uses the freedom of choices that a content analyst exhausts when assembling his analytical construct according to established theories of the source. He can establish construct validity by showing which logical alternatives in the analytical process were discarded and why. A third approach might be mentioned. It relies on extraneous evidence of variability in the source. This is best illustrated by HOLSTI and NORTH's (33) attempt to validate inferences made from political documents exchanged during the crisis preceeding World War I. The authors analysed these exchanges on a day-to-day basis regarding expressed hostilities, tension and the like. The

measures showed variation but evidence about corresponding variation in reality was lacking. HOLSTI and NORTH then searched for the validating information in diaries and memoires of those who took part in the 1914 decision. Reportedly, in one instance, the quantitative analysis of the Kaiser's messages and marginal comments on other documents indicated that he was under considerable stress during the final days prior to the outbreak of war. Eyewitness accounts of his closest aides apparently supported this inference. For example, according to HOLSTI and NORTH, Admiral TIRPITZ wrote of the Kaiser during this critical period: "I have never seen a more tragic, more ravaged face than that of our Emperor during those days." The fact that TIRPITZ found this observation noteworthy and at variance with the Kaiser's usual expressions is clearly an indication of variability of the criterion. Although this account is entirely anecdotal and unsystematic in nature, the absence of any evidence about the variability in the Kaiser's manifest stress would have left the validity of the content analysis measure entirely uncertain.

The fourth and last point I wish to make in this section is that validation does not resolve all uncertainties in content analysis. To start out with, validation is essentially a process by which those analytical procedures are weeded out whose inferences do not correspond with existing evidence. But by eliminating those procedures that conflict with reality, the remaining ones are not necessarily valid. This is so because all inferences that can be obtained from content analysis are inductive in nature which suggests, among other things, that past successes may not hold in the future. Actually, were it not for the ergodicity assumption required in induction, one should relable the process "invalidation" for the negative proof it provides.

Besides this uncertainty in induction, it is useful to distinguish between two errors in content analysis. One is revealed when inferences are shown to be wrong. Because the identification of such errors is all that validation can accomplish, I call this the error of validity. The other is perhaps less conspicious but potentially more serious and is easily committed when inferences from content analysis extend beyond what can be validated in principle or in the near future. I call this the error of extension.

Errors of extension are exemplified by a content analysis that claims to make inferences about the cognitive structure of an author but restricts itself to contingency analysis only. While there is validating evidence available for the existence of associational connections on which contingency analysis is built (34), it is highly unlikely that associations explain all patterns of cognition as implied in the analyst's claim (35). Unfortunately, content analysts often leave the target of their inferences vague, making it

difficult to ascertain either error. For example, a content analysis that merely claims to describe violence in television fiction leaves open to question which kind of violence (as commonly understood or as operationally defined) is described and by what kind of data it could be validated (viewers ability to identify it in the same way, violent behaviors caused from exposure to media violence, release of aggression while viewing, etc.). With the target of intended inferences left uncertain, apparent evidence about the invalidity of some aspects of such findings then allows the content analyst to withdraw into a niche in which validating information cannot readily be brought to bear on the situation. He is then likely to commit errors of extension which for no good reason seem to be feared far less than errors of validity.

In practical applications of content analysis results, errors of extension might be considered more severe than errors of validity. Predictions that are known to have been true in only 60% of all cases allow a decision maker to ascertain at least the risk of failures. In the absence of validating information such risks are simply unknown. While speculations and hypotheses undoubtedly extend man's understanding into yet unknown domains, decision makers must be concerned especially with errors of validity for only if content analyses can be upheld in the face of validating information can inferences they provide serve as a justification for practical action.

4. Semantical Validity

The first step of almost all content analyses involves some recognition of the meaning, references or other semantic features in the data at hand. In fact, older definitions made "the classification of sign-vehicles" (36), the "description of the manifest content of communication" (37), the "coding" (38), the "putting (of) a variety of word pattern into ... (the categories) ... of a classification scheme" (39) a definitional requirement of content analysis. And, indeed, many content analyses are intended to render nothing other than a quantitative account of the semantical features that trained observers recognize.

By and large, semantical validation is not a problem in psychological testing although I cannot claim it to be unique to content analysis either. Let me start with a few simple examples: Suppose a content analysis is designed to determine whether the proportion of commercials aired by a certain station exceeds legally prescribed limits. Even so the task may be regarded as a "purely descriptive" one, this does not free the analyst from an examination of whether the classification does correspond to the legal conceptions. Or, when the frequency of foreign vs. domestic

news items is at stake, the count is preceeded by a distinction that may or may not correpond to journalistic, political or common distinctions. Data oriented types of validity are particularly important when content analyses are descriptive in intent whereby semantical validity evaluates whether the distinctions made by the descriptive language conform to some given standard, knowledge or expert judgment about the source. In these simple examples, semantic validation would involve respectively whether commercials are identified as stipulated by the law, or whether the distinctions between foreign and domestic news items conform to the distinctions of some reference group.

Entirely descriptive tasks in content analyses should not be belittled, neither in their scientific importance nor in the methodological problems posed by them. On the one hand, even the identification of evidence about "achievement motives" in popular literature, for example, may be regarded as a sort of description although of an extremely complex sort, possibly involving procedures that may have to be subjected to construct validation. On the other hand, any classification, however simple, must be regarded as a form of inference leaving open to question why different units of analysis are put into the same category. Finally, descriptions are often an initial part, of a larger analytical effort in the context of which semantical validation might lend some initial certainty to the data subsequently used. An example of the latter may be found in the following:

In the course of a larger project aimed at analyzing values in political documents, I was once confronted with the problem of developing a procedure that would allow one to identify what we called "value laden sentences" in a given text. A panel of experts could pick them out fairly reliably but coders varied greatly in this ability and computer programs we had hoped to employ turned out to be virtually powerless in this case. I will not describe the history of this work except to say that we started out by distinguishing between sentences that did or did not contain established political symbols such as democracy, freedom, victory and ended up by testing each sentence for its conformity to any one of a set of structural definitions of the way values are expressed (40). In attempting to increase the approximation between the set of sentences that coders identified by our method and the set of sentences considered value laden by experts, we were in fact engaged in testing and iteratively improving the semantical validity of the identification procedure.

In these simple examples, semantic validity is manifest in the identity of two distinctions, one made by the method, M, and one considered to be valid, V. The intersections can then be interpreted as follows:

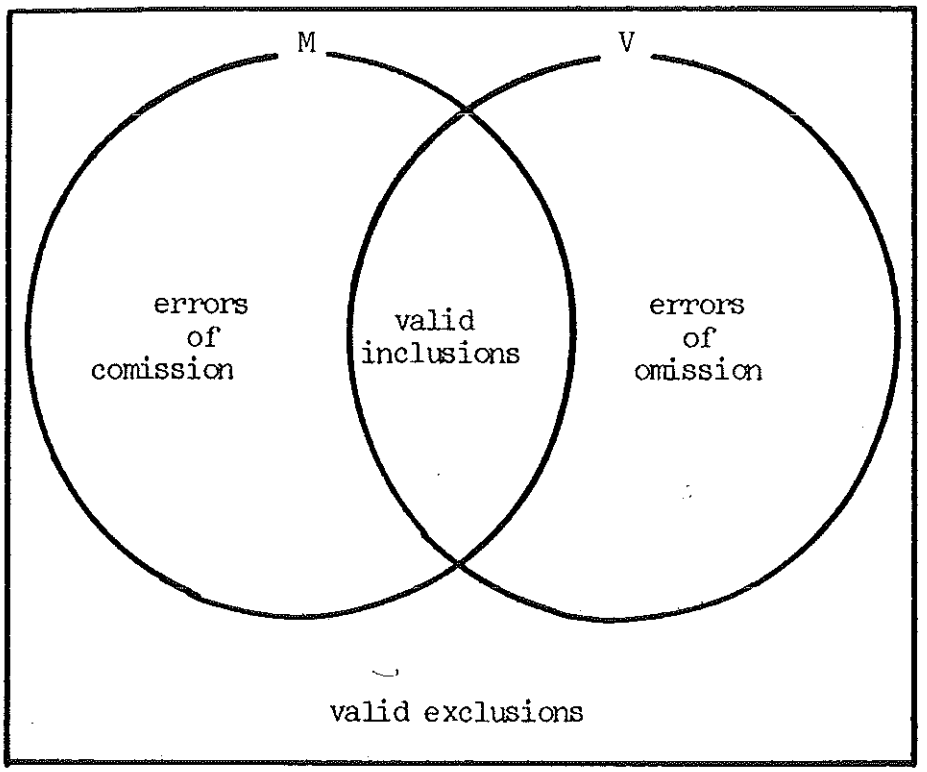


Figure 2

But the examples so far considered are too simple. Distinctions often involve many categories and errors of comission of one category may be errors of omission of the other. To begin with the generalization of the diagram, let me state that any unambiguious description of events, any classification of signvehicles, any reliable coding of messages, in fact any proper measurement procedure defines a mapping of a set of units of analysis into the terms of an analytical language. Accordingly, some units of analysis are assigned the same terms or categories and are hence considered equivalent with respect to the analytical procedure to be evaluated. Units of analysis that are described in identical terms thus form an equivalence class and any procedure embodying the mapping of units into analytical terms effectively partitions the set of all units (whether it coincides with the sample of units actually obtained or with a universe of combinationally possible units) into a set of mutually exclusive equivalence classes. Graphically the situation is depicted as follows with the three tags for personal pronouns of the Harward III Social-Psychological dictionary of the General Inquirer taken as example:

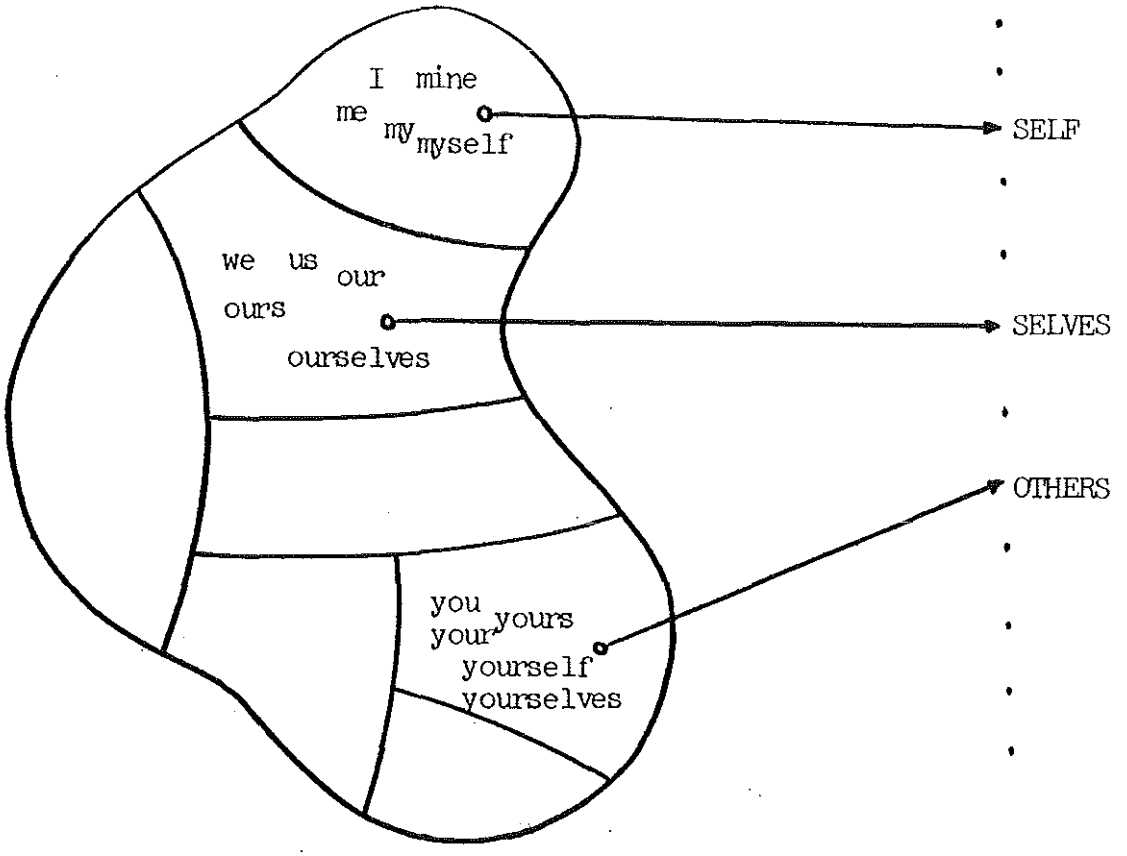


Figure 3

The tags, SELF, SELVES and OTHER, can be regarded as labels of different equivalence classes of pronouns and are as such part of the partition that all tags of the computer dictionary induce.

Now, all evaluations of the semantical validity of a classification (measurement, coding identification, etc.) involves a comparison of two partitions, the partition induced by the method to be evaluated and an independently obtained valid method. This comparison may be depicted in form of a lattice of partitions with the least upper bound containing the largest number of distinctions on which both partitions agree, the largest lower bound containing the smallest number of distinctions (all distinctions) occurring in both partitions, and letters a, b, ..., g denoting recording units or classes thereof whether they be words as in the General Inquirer dictionary, sentences or other symbolic units.

The extent of agreement between the two partitions then serves as a measure of the semantical validity. In terms of the above lattice, perfect agreement exists when all four

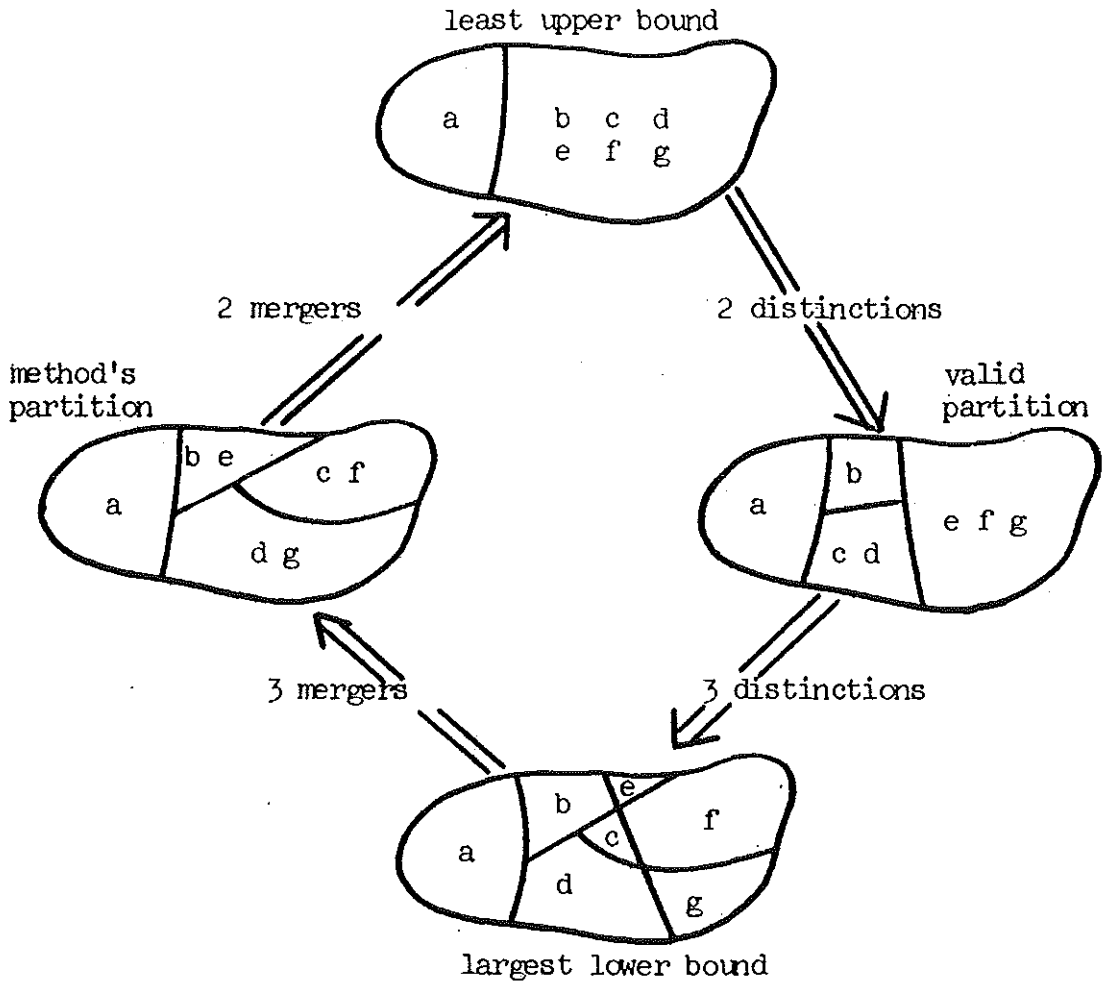


Figure 4

partitions are identical. Deviations from this ideal can be measured by the number of steps required to obtain the least upper bound and the largest lower bound of the two partitions from each other by stepwise merging or partitioning its elements. In the above figure, the valid partition and the method's partition may be said to be five units apart. (Suitable forms of standardization of this measure are possible but are immaterial for the purpose of this paper.)

While this method of establishing the semantical validity of a procedure is stated in quite straight forward terms, there are often practical obstacles against obtaining a valid partition of units by independent means. Nevertheless, it is almost always possible to obtain a listing

of all units of analysis that find themselves in the same category (and are, hence, treated by the analytical procedure as semantically equivalent) and to inspect such a listing in some detail for whether the members of the equivalence classes thereby formed can be regarded as synonymous, whether noticeable semantical differences can be ignored and distinctions are indeed meaningful. OSGOOD, SUCI and TANNENBAUM (41) provide examples of establishing the validity of their semantic differential scores by comparing the word clusters obtained by the technique with subject's judgments of word similarities without using the differential.

In computer approaches to content analysis, semantical validity is a particularly important problem. Often, such approaches amount to nothing but counting words without consideration of their meanings. Where lexical differences among words coincide with differences in meaning, problems of semantical validity are then absent indeed. But this is rarely the case. An example is provided by DUNPHY who presents a sample of the Key-word-in-context printout for the word "play" (42) to show that a computer program that merely identifies occurences of the word "play" ignores its many different senses. Thus, if it were significant for an analysis to distinguish between the meanings of "play" in:

A PLAY	A theatrical performance
To PLAY an instrument	To manipulate
To PLAY a large role in...	To contribute
To PLAY around	To do no serious work
To PLAY baseball	To be involved in a game
To PLAY music	To be able to reproduce music using an instrument
A PLAY boy	A particular individual
To PLAY with the other children	To interact with others in an undirected way

Mere lexical identifications, without consideration of differences in contexts are clearly insensitive to the multiple meanings of the word. To establish the semantical validity of a whole tagging dictionary procedure would involve examining all occurrences in a text that the procedure regards as identical and then matching these equivalence classes with semantical distinctions obtained by a different method of unquestionable validity.

The problem of semantical validation arises most naturally at the developmental stage of an analytical procedure, or when the applicability of an existing instrument is in

question. At this stage the body of data for which the procedure is intended tends to be somewhat unknown and the analyst will have to resort to creating artificial data for the purpose of validation, data that contain all of the expected semantical peculiarities. A strategy for generating such data is combinatorial or logical extension. The former is exemplified by generating all possible data (e.g. words) from basic elements (e.g. letters) and the latter by creating all conceivable counter examples.

The latter method is well founded in linguistics where a procedure that claims to embody a theory, say, of the English language should recognize or generate all and only sentences that native speakers judge to be grammatical English sentences. I will not dwell on the controversy that such a demant has created for the task of linguistics except to point to the fact that this criterion makes references to a potentially infinite universe consisting of all English sentences that have been uttered in the past, that will be uttered in the future and that may for whatever reasons never be uttered but are proper English sentences nevertheless. Faced with such a vast universe, linguists tend to consider hypothetical examples, that are often constructed with great ingenuity, to ascertain whether the procedure would properly dinstinguish among their syntactical or semantical features, and thereby locate the syntactical or semantical features that account for errors. The search for linguistic "counter examples" is an effort at semantical invalidation and as such an established method of science.

Content analyses tend not to have such general aims but may nevertheless be validated by similar methods. In the critique of his contingency analysis, OSGOOD (43) employs the same mode of reasoning. In effect he observes that when a psychoanalytic patient states:

1) "I loved my mother."

A contingency analysis would add this incident to the association between LOVE and MOTHER, and so do the following statements:

2) I have always loved my mother more than anyone else.

3) Mother loved me dearly.

4) I never loved my mother.

5) "I have always loved my mother?" Ha! Ha!

6) My (be)loved father hated mother.

Since the two critical words co-occur in all six statements, a contingency analysis would cast them into the same equivalence class. However, relative to 1), 2) shows contingency analysis to be insensitive to the strength of an expressed association, 3) shows contingency analysis to be insensitive to active-passive distinctions, 4) shows con-

tingency analysis to be insensitive to negation, 5) shows contingency analysis to be insensitive to irony, and 6) shows contingency analysis to be insensitive to grammatical considerations. OSGOOD makes the additional point that contingency analysis is incapable of responding to instrumental uses of language, for example, when the patient did not love his mother but wants his psychoanalyst to believe that he did. If some or all of the differences among the above statements are analytically significant, then contingency analysis, counting co-occurences only, would have to be jugded semantically invalid. However, inasmuch as OSGOOD has demonstrated some correlational validity of contingency analysis, some psychological processes might well be indicated by the technique. A critical examination of the semantical distinctions that an analysis makes or discards may thus give valuable insights into the nature of a procedure and provides perhaps sufficient reasons for accepting or rejecting its results.

5. Sampling Validity

Generally, sampling validity assesses the degree to which a collection of data can be regarded as representative of a given universe or as in some specific respect similar to another sample from the same universe obtained by the same method. In content analysis, the sampling validity criterion is intended to assure that the data contain with a minimum of bias a maximum of information about the data source.

The most familiar case of sampling validation and possibly the one that the Technical Recommendations refer to by the name "content validity" involves showing that two samples are similar in the sense that both are representative of the same universe, and, since the analysis of one yields valid inferences - the argument continues -, there is then no reason to suspect that the same analysis of the other does not. Hence, validity is transferred from one sample to the other on the basis of their being individually representative of a common universe. But this is only one case of sampling validity.

Another rather obvious case of sampling validity is invoked when a content analysis has purely descriptive aims. Such is the case when one is concerned with an author's vocabulary, with the frequency of dramatic violence on television, with whether or not a document exists, with the kind of references made in a body of text, etc. Descriptive aims are associated with content analysis since BERELSON's (44) definition and implied in the process of "identifying specified characteristics of messages" which HOLSTI (45) and STONE (46) consider a definitional requirement of the technique. Content analysts with purely descriptive intents

can avoid problems of sampling either by analyzing all data on a given phenomenon or by refusing to generalize their findings. The concordance of the complete works of an author is an example of the former "solution" while the examination of one solid week of television programming (without attempt to sample over a larger time span and without intent to interprete the findings beyond that one week) is an example of the latter "way out." But when the work of an author becomes two voluminous and choices need to be justified, or when one year's television programming is to be compared with another, using a week of programming only, then questions of sampling are inevitable. While there are many practical problems associated with sampling in content analysis, most of which have been ably discussed by KOPS (47),the theory that outlines how samples are to be drawn in such situations and how far findings are generalizable is essentially worked out. For purely descriptive intents sampling validity reduces to a measure of the degree to which a sample is an unbiased collection from the universe of possible data. As a criterion it assures that the sample's statistical properties are similar to those of the universe and in that sense represent the universe within analyzable magnitudes.

However, the relation between sample and universe is often confounded by other notions of "representation" that seems to be inherent in the message characteristics of communications and possibly constitute an essential ingredient of the symbolic nature of content analysis data. The distinction is dramatized in the difference between attempts to make inferences about an author's vocabulary (the universe) from a small sample of his works and attempts to make inferences about an author's cognitive structure which is merely manifest in, not part of that author's writings. In the first case, inferences are statistical generalizations from a sample to a universe of which the sample is a part, while inferences in the second case follow the paths of linguistical and psychological representations and perhaps of causal connections from a sample to its antecedent conditions, neither of which is contained in the other. Such examples demonstrate the need for a broader validity criterion, one that is applicable to other forms of representation as well.

Note that content analyses with purely descriptive aims equate the universe from which a sample is drawn with the target of an analysis in which case data must be sampled to assure that each datum has the same probability of inclusion in the sample. But, content analyses that aim at making specific (content) inferences about a source must distinguish between the universe of messages from which the sample is either drawn or made available and the target of the intended inferences which is the universe of meanings, consequences, causes, antecedent conditions, states or events not directly observable on the source. Processes

that mediate between the two universes are attributable to the real world of the source and are not under the control by the analyst.

It is well known in communication research that almost all social processes that originate in cognitive or real world events and yield communications, symbolic representations, indices, etc. - the data for content analysis - are selective, biased, and constitute in effect processes of self-sampling: Consider the over-representation of marrying age and well-to-do WASP's among television characters, consider the selective way witnesses in court recall events from memory, consider how few personality's private lives are considered news worthy, or the kind of individuals known to us from history and mythology, consider how social prejudices and tabus constrain the assertions being made, regardless of the facts, etc. Processes of self-sampling assign uneven probabilities of inclusion of events into symbolic forms. While stochastic in nature, such processes are likely to be systematic in the sense that they are describable in sociological or psychological terms and knowledge about them can be used in evaluating the representativeness of available data.

In content analyses with inferential aims, the choice of data must undo the statistical bias inherent in the way data are made available to the analyst. The sampling validity criterion is intended to evaluate the success of this effort and the key to it lies in the knowledge about the self-sampling characteristics of the data source, i.e. about the statistical relation between available data and the universe of possible data of interest. I will illustrate the two principal self-selecting processes by means of an example and then outline a method for evaluating sampling validity.

Suppose the task is to compare opinions held by the decision making elites in the United States and in the Peoples Republic of China on some important political issue, say, regarding acceptable forms of alliances between the U.S.S.R., the U.S. and China. Furthermore, suppose that the U.S. data are obtainable by personal interviews whereas the Chinese data must be obtained from mass communications. Techniques for making valid inferences from survey data are well developed so that the processing of the U.S. data presents no problem. However, the validity of inferences from content analyzing the Chinese data is in doubt. Without sampling validation of the Chinese data, comparisons may lead to unwarranted conclusions which are all the more undesirable as political actions might be dependent on these findings.

The first step is to delineate the universe of interest to the analyst, usually the target of the intended inferences.

With the model of survey research in mind, the U.S. data may be collected by interviewing a random sample of individuals from a list that contains members of the U.S. Congress, high level officials in the State Department and in the White House plus certain influential personalities from business and industry. But, members of the Chinese elite while known in large categories are not individually accessable. Now, a simple minded content analyst might easily be lead into a methodological trap by drawing a sample from Chinese news print, domestic and foreign broadcasts, etc., and thereby contributing data for comparison that are representative of an entirely different universe: the universe of mass media expressions. To avoid invalid comparisons, a content analyst must therefore differentiate between the universe of messages and the universe that he considers the target of possible inferences, here the members of the decision making elite in China.

With the two separate universes in mind, the next step is to obtain information about the self-sampling characteristics of the source. Two processes must be distinguished for they result in rather different corrective actions. The first concerns the probability of an opinion on foreign policy to enter or not to enter a particular medium regardless of who's opinion it may be. Obviously, opinions on foreign policy are less likely to be found in typewriter manuals, commercial advertisements or in local news items. Prestige papers and official government organs might be more informative. And the knowledge of this probability allows the analyst to decide among the media to be considered relevant or irrelevant respectively. The first process amounts to an either-or distinction with probabilities indicating the relevance of the communication channel for analysis.

The second process concerns the probabilities with which members of the Chinese elite voice or are given preference to express their opinions on foreign matters. We know that the accessibility of mass communications to members of any decision making elite is rather unequally distributed. Someone in charge of propaganda and publicity has easier access to the media than others; someone who assumes a more public role is likely to make news more readily than someone who fills administrative posts only, though the opinions of both may be equally significant when it comes to foreign policy formulations. Additionally, some members of the elite may have preferences for or even political obligations to publish in one rather than in the other medium.

Samples drawn by a content analyst who ignores such processes of self-sampling might be representative of Chinese mass communications but it will be biased with respect to the Chinese political elite. To undo the

statistical bias inherent in the way political opinions are selectively published and communicated to the analyst, 1) media with a low probability of carrying political opinions may be ignored in favor of those that are more likely to contain relevant information and 2) those members of the elite that are overexposed should be sampled less than the underexposed members. In other words, in content analyses with inferential intents, high sampling validity can be achieved only when the analyst samples from available messages in such a way that he obtains a representative sample of the phenomena of interest rather than of what happens to be made available by some source.

Given, then, an estimate of the probability with which a phenomenon of interest, here, an opinion held by a member of the decision making elite, is represented in the stream of available data, an unbiased sample from these data must assure that the frequency of that phenomena is available for sampling with n_i as the frequency of phenomenon i in the sample, and P_i as the estimated probability that the phenomenon I will be made available to the analyst, the criterion against which sampling validity is to be measured is

$$n_i \text{ is proportional to } p_i^{-1}$$

when p_i is uniformly distributed and self-sampling is, hence, unbiased, sampling validity reduces to showing that sampling from available data was random. In the example, sampling validity would exist only if rare opinions by unusually invisible decision makers would be given a larger attention than common opinions associated with highly visible communicators (48).

The proposed condition for sampling validity in content analysis is stated here only for a minimal situation, one from which all complications are removed. Others will have to be developed following the spirit of the preceeding discussion. Regardless of the form such a condition may then take, the aim of sampling validity is to assure that data represent the universe of interest and surpass inevitable biases inherent in the way data are made available to the analyst.

6. Pragmatical Validity

A classical example of a pragmatical validation is provided in STONE and HUNT's (49) attempt to differentiate real and simulated suicide notes by computer content analysis. The first step of this demonstration involved an analysis of 15 real and 15 simulated notes by the General Inquirer (50). It revealed three discriminating factors:

a) References to concrete things, persons and places (higher for real notes)

b) Use of the actual word "Love" in the text (higher for real notes)

c) Total number of references to processes of thought and decision (higher for simulated notes)

These were incorporated in to a discriminate function. In a second step, this function was then applied to the remaining pairs of notes whose identity was unknown to the researchers. It turned out that 17 out of 18 pairs of notes were correctly identified as either real or simulated. Apparently the computer faired better than human judgement.

In this demonstration, the initial 15 pairs of notes of known identity constituted the validating information for the existence of an empirical link between text characteristics and attributes of the source. The discriminate function represented this link procedurally. And the subsequent success of the inferential procedure was cited as further evidence for the validity of this discriminant function. A total of 32 out of 33 correct inferences - so one would argue here - lends some if not considerable pragmatical validity to the inference that might be drawn from subsequent notes. The fact that the discriminate function as discovered and applied did not seem to be derivable from existing theory was apparently irrelevant to the reasoning.

The argument used is a fairly simple one:

- For a given sample (of pairs of letters from which inferences were to be made and actual conditions of their authors) the method was shown to be successful to a degree better than chance;
- The new data on hand are similar to or compatible with those that led to successful inferences in the past;
- Therefore, the inferences now drawn from these data by the same method may be accepted as evidence on the ground of that method's record of past successes.

While the absence of theoretical considerations in pragmatical validation is not regarded as a deficiency, a more serious problem is that content analyses are rarely sufficiently repeated in practice. POOL (51) and HOLSTI (52) have complained about the lack of standards in content analysis and that many studies are designed ad hoc and are unique. The lack of repetition renders knowledge of the degree of success of a method uncertain. Whether the units of analysis are suicide notes, single works, whole books, taped interviews or TV episodes, the sample size of past applications and the proportion of inferences known to have been correct must be large enough to lend pragmatic validity to a content analysis.

Heterogeneities in the population present another problem. An important step in the pragmatical validation of content analyses involves showing that "the data on hand are similar to or compatible with those that led to successful inferences." It assures the applicability of the assumption that the data - inference relation holds also outside the sample (within which successes and failures have been experienced), specifically in the case to be validated. As one goes outside this sample, pragmatical validation does therefore not allow for meanings to change and the symbolic manifestation of source attitudes and attributes to be different. The pragmatic validation of content analyses can thus proceed only within a relatively homogeneous population.

Thus, in content analysis, pragmatic validation refers to whether an analysis "works" and is measured by how successful content inferences are in a variety of circumstances regardless of the nature of the process involved. It involves an inductive argument amounting in fact to a generalization from a sample of inference-evidence pairs to a larger population of such pairs with the law of large numbers providing the primary basis of the justification.

Pragmatic validation can be accomplished in two ways: by correlational validity and by predictive validity. I will discuss these types in the following sections.

7. Correlational Validity

Correlational validation is most common in psychological testing, has a long history and its methodology is therefore highly developed. It has virtually coevolved with statistics in the behavioral sciences and is based on the idea that whenever a variable is to be interpreted as a measure of another quantity or magnitude, it must at least correlate with it. In psychological terms, a test is said to provide meaningful indices to the extent test scores and criterion scores are shown to correlate highly.

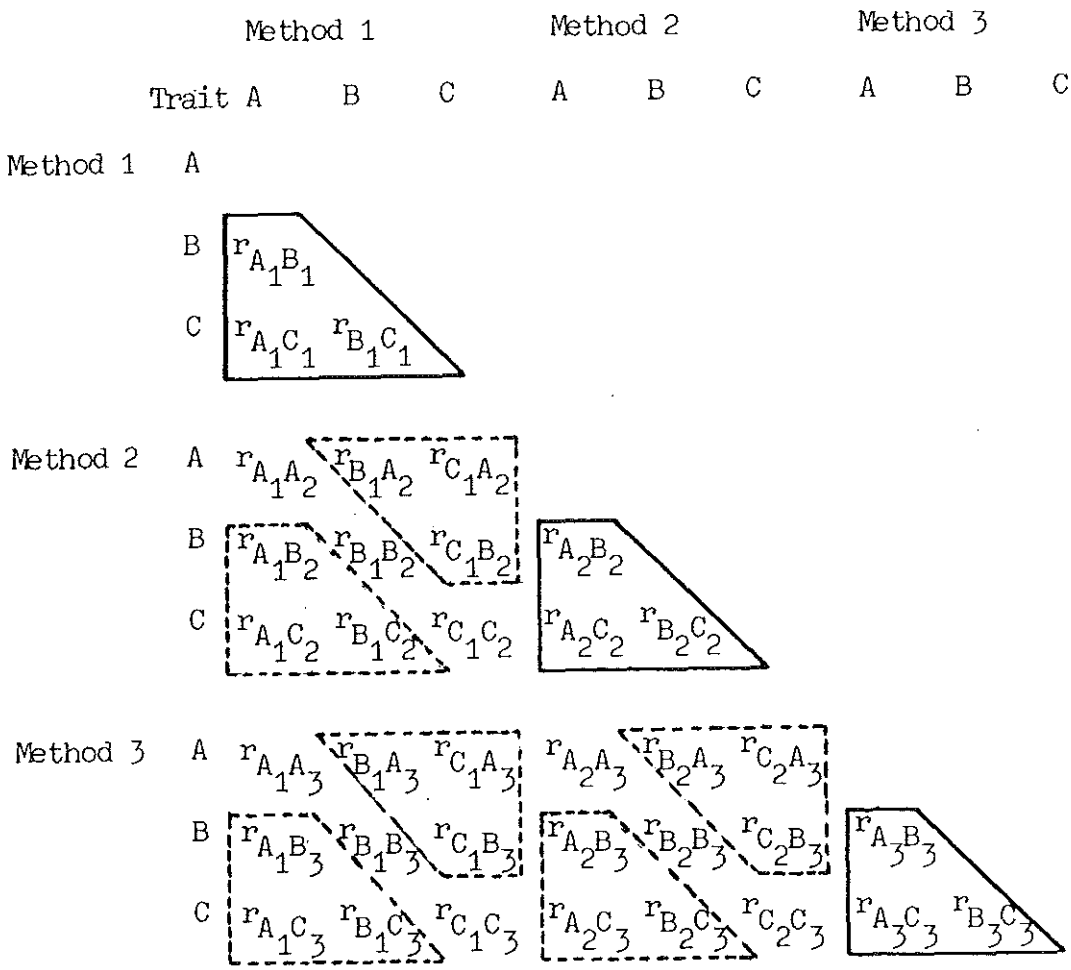
As mentioned above, the Technical Recommendations make rather unfortunate distinction between predictive and concurrent validity dependent on whether test results are intended to correlate either with measurements obtained at some subsequent point in time or with criteria available concurrently. Accordingly, aptitude tests would require predictive validation while tests that classify patients would have to be validated against concurrent criteria. The distinction solely relies on the difference in time between administering a test and obtaining validating information about its criterion. In content analysis we do not need to make this distinction but recognize that both types are established by demonstrating statistical correlation.

A fact well recognized in the psychological literature is that test results and criteria are both measures, neither of which should be confused with the phenomena either of which claims to represent. Since correlations do not predict but indicate the strength of a systematic (linear) relation between measures, the demonstration of high correlation between them therefore provides nothing but a justification for substituting one measure for another with one presumed to have certain practical advantages over the other. Among the practical advantages of content analysis is that it provides unobtrusive measures, permits inferences from symbolic as opposed to behavioral data, allows analyses of records that antecede interest in them, etc. To determine whether a content analysis might be used in place of a psychological test, a survey, or other more direct measures of phenomena of interest, a demonstration of high correlation between the content analysis and those measures needs to be demonstrated.

CAMPBELL and FISKE (53) were the first to develop the idea of validation by correlational techniques into a full-fledged methodology. They recognize that any justification for a novel measure requires not only a high correlation with established measures of the trait it intends to measure but also low or zero correlations with established measures of traits it intends to discriminate against. The former requirement is called convergent validity, the latter discriminant validity. Thus, a research result may be invalidated by either or both, low correlation with measures of the same trait and high correlation with measures of traits against which it intends to differ.

To show that a measure possesses both convergent and discriminant validity calls for correlations between measures of a variety of traits, each obtained by several independent methods. The matrix of correlations obtained for this purpose is called a Multitrait-Multimethod Matrix. Such a matrix is presented as Figure 5.

In this Multitrait-Multimethod Matrix, the heterotrait-monomethod correlations are found within the solid boundaries, the heterotrait-heteromethod correlations are found within the broken boundaries, leaving the diagonals to contain the monotrait-heteromethod correlations. Convergent validity is indicated by high monotrait-heteromethod correlations whereas discriminant validity is indicated by low heterotrait-monomethod correlations. More specifically according to CAMPBELL and FISKE (54):



Multitrait-Multimethod Correlation Matrix

Figure 5

Convergent validity is indicated when the monotrait-multimethod correlations differ significantly from zero and ideally approach one:

$$r_{A_1A_2} >> 0$$
$$r_{A_1A_3} >> 0$$
$$r_{A_2A_3} >> 0$$

and so on for B and C.

Discriminant validity is indicated when:

1) within each heteromethod block, the monotrait correlations are larger than the corresponding heterotrait correlations:

$$r_{A_1A_2} > r_{B_1A_2}$$
$$r_{A_1A_2} > r_{C_1A_2}$$
$$r_{A_1A_2} > r_{A_1B_2}$$
$$r_{A_1A_2} > r_{A_1C_2}$$

and so on for $r_{A_1A_3}$, $r_{A_2A_3}$, and for B and C.

2) for each method and for each trait, the monotrait-heteromethod correlations are larger than the corresponding heterotrait-monomethod correlations:

$$r_{A_1A_2} > r_{A_1B_1}$$
$$r_{A_1A_2} > r_{A_1C_1}$$
$$r_{A_1A_2} > r_{A_2B_2}$$
$$r_{A_1A_2} > r_{A_2C_2}$$

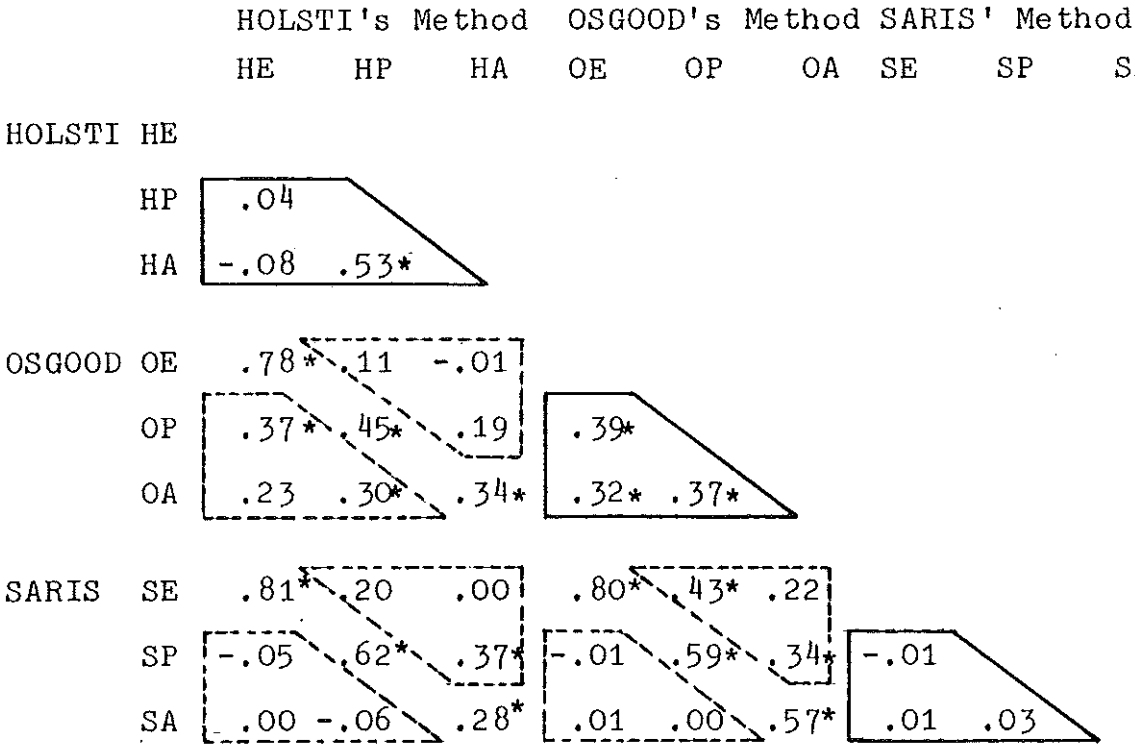
and so on for $r_{A_1A_3}$, $r_{A_2A_3}$, and for B and C.

3) and according to ALVIN (55), the rank ordering of the heterotrait-monomethod correlations should be repeated in each heterotrait-heteromethod triangle. For example, if

$$r_{A_1B_1} > r_{A_1C_1} > r_{B_1C_1} \quad \text{then} \quad r_{A_1B_2} > r_{A_1C_2} > r_{B_1C_2}$$
$$r_{A_1B_3} > r_{A_1C_3} > r_{B_1C_3}$$

and so on for all heterotrait-heteromethod correlations.

An example of evaluating the substitutability of several content analysis approaches to the three dimensions of OSGOOD's affective meaning is provided by MORTON, SARIS-GALLHOFER and SARIS (56). The authors correlate the results obtained by HOLSTI's computer dictionary, OSGOOD's method, and their own newly developed indices for OSGOOD's evaluative, potency, and activity dimension and obtained the correlations presented in Table 1:



Multitrait-Multimethod Matrix for three Different Content Analysis Procedures of Dictionary Construction for Affective Meaning Inferences

Table 1

(HE = Holsti evaluative, HP = Holsti potency, HA = Holsti activity, etc.)

Without examining the reason for these correlations, the entries in the monotrait-heteromethod diagonals would lend support to the contention that the three methods posses convergent validity and that this validity is higher in the case of the evaluative dimension and lower in the case of the activity dimension.

However, the pattern of reasoning is more difficult in the case of their respective discriminant validities. Of the twenty-seven inequalities stipulated in criterion 2) above, four do not hold of which three are caused by the high correlation between the potency and activity dimension in HOLSTI's monomethod triangle and one by the high correlation between the same two dimensions in OSGOOD's monomethod triangle. While suitable tests of the significance of these inequalities are not reported in this research report, the pattern of partial failure to satisfy them would speak in favor of substituting the latter for any one of the two former methods. This result is evidently clearer for the evaluative dimension than for the measured activity. For lack of space we omit the analysis of the inequalities in the discriminant validity criterion 3).

In the domain of content analysis, correlational validity is of particular importance when the phenomena of interest mediate between the reception and production of messages. This refers most obviously to all mediational concepts of meaning which underly a large number of research designs and are explicit in, among others, OSGOOD's affective meaning system evoked in the above example. The first one to recognize this is JANIS who in 1943 suggested that the content analyst's job is to "estimate the significations attributed to signs by an audience" (57). He thought of significations as internally represented meanings that come immediately to mind whenever someone is confronted with some sign, verbal assertation or symbol and that will effect the verbal or nonverbal behavior of audience members. JANIS points out that significations cannot be observed directly. But because of their presumed effect on message receivers, in order for a content analysis to be valid, its results must at least correlate with some aspect of audience behavior.

Continuing with JANIS, where the criterion variables are directly observable, inferences from content analysis should agree rather than correlate (a difference that will be discussed under predictive validity). But whenever inferences refer to phenomena or events that are only indirectly observable, i.e., when validating information is merely related to the phenomena of interest, correlation is the only key to evidence about validity. JANIS called the latter the indirect method of validation and discusses some typical sources of errors when attempting to validate content analyses by this method. While I do not feel that content analysis is limited

to JANIS' conception, mediational phenomena are common targets of content analyses indeed. And when mediational phenomena provide the focus of attention, indirect methods of validation with their necessary reliance on correlational techniques are indispensible.

8. Predictive Validity

Prediction is a process by which available knowledge is extended into an unknown domain. The predicted phenomena may have existed somewhere in the past - as for historical events or the antecedent conditions of available messages - may be concurrent with the data being analysed - as for inferences about attitudes, psychopathologies or personalities of interviewees - or are anticipated to be observable sometime in the future.

While substitutions are justifiable by demonstrating high correlation, predictions must exhibit high agreement with the phenomena, event or attributes being predicted. Ideally, predictions and facts stand in a one-to-one semantical correspondence. This difference between correlations and agreement is crucial here: A slow watch will correlate highly with time but is systematically wrong and therefore useless, unless one knows the bias. The famous body-count during the Vietnam war may have correlated highly with military activity but its numerical value turned out to have no meaning. Political decision makers can hardly be satisfied with the assurance that content analysis estimates of "war mood" from enemy propaganda correlate highly with other indicators when it can not be known whether these inferences systematically over or under estimate the facts. The examples serve to show that high correlation is a necessary condition for predictive validity but it is not sufficient. Instead, it is required that inference from content analysis and known facts agree.

Digressing into epistemology: facts too are accessable only through the medium of described observations, i.e. measurement. Thus, if the difference between predictions and substitutions would merely rely on the difference between establishing agreement and establishing correlation, then predictive validity could be equated to a kind of strong substitutability. But the difference depends also on the observational status of the criterion chosen: Employers can hardly be impressed by how well their applicant's scores in one test correlate with those of another. But they are eager to know how well they will actually perform on a job. The network of correlations among readability scores of school textbook, say, may provide further insights into instrument design. But what ultimately decides among them is high agreement with observed reading ease, speed, inter-

est, comprehension, etc. Measures of dramatic violence on television are to be regarded similarly. High correlations among them may justify substitution but say nothing about their predictive value. Predictive validity is demonstrated only in high agreements with directly observable facts (audience behavior, crime rate, public fear, etc.) that matter to someone with vested interest in the reality, so observed, by policy makers, for example. Substitutability may accept any variable as a criterion but predictive validity accepts only those that are important to someone because of their factual status.

Qualitatively, predictive validity is assessed by entering each of a set of possible events in the following four-fold table:

	events predicted	events excluded by prediction
events that did occur	A	B
events that did not occur	C	D

Events Counting for and Against Predictive Validity

Figure 6

Obviously, when all events fall into the A and D cell predictive validity is perfect (except for a sampling error where applicable). Content analyses can make two kinds of errors. They may say too much including being always correct and thereby commit errors of comission which appear in cell C. Or they say too little without necessarily being wrong and commit errors of omission which appear in cell B. For predictions to be meaningful and validatable, neither row nor column must be empty, that is, there must be evidence for discrimination and convergence, to borrow CAMPBELL and FISKE's terms.

A classical, though not quite perfect example of this form of predictive validation is GEORGE's (58) attempt to evaluate the FCC inferences made from enemy domestic propaganda during World War II. All inferences were available in form of reports by the propaganda analysts and could be

matched one by one with documents that became available after the war. Those inferences for which validating information was available were judged either correct, nearly so, or wrong. The results showed the analysis effort to have had considerable predictive validity. The validation is not quite perfect because, by putting inferences in these three (or similar) categories, cells B and C are not differentiated and D is probably discarded.

In a more quantitative mode, predictive validation can follow CAMBELL and FISKE's criteria with one important difference, that the entries in the Multitrait-Multimethod Matrix are not correlations but agreement coefficients. Such coefficients have been proposed by KRIPPENDORFF (59) and are not further considered here except that these agreement coefficients measure the degree to which two variables match or, conversely, deviate from perfect matching.

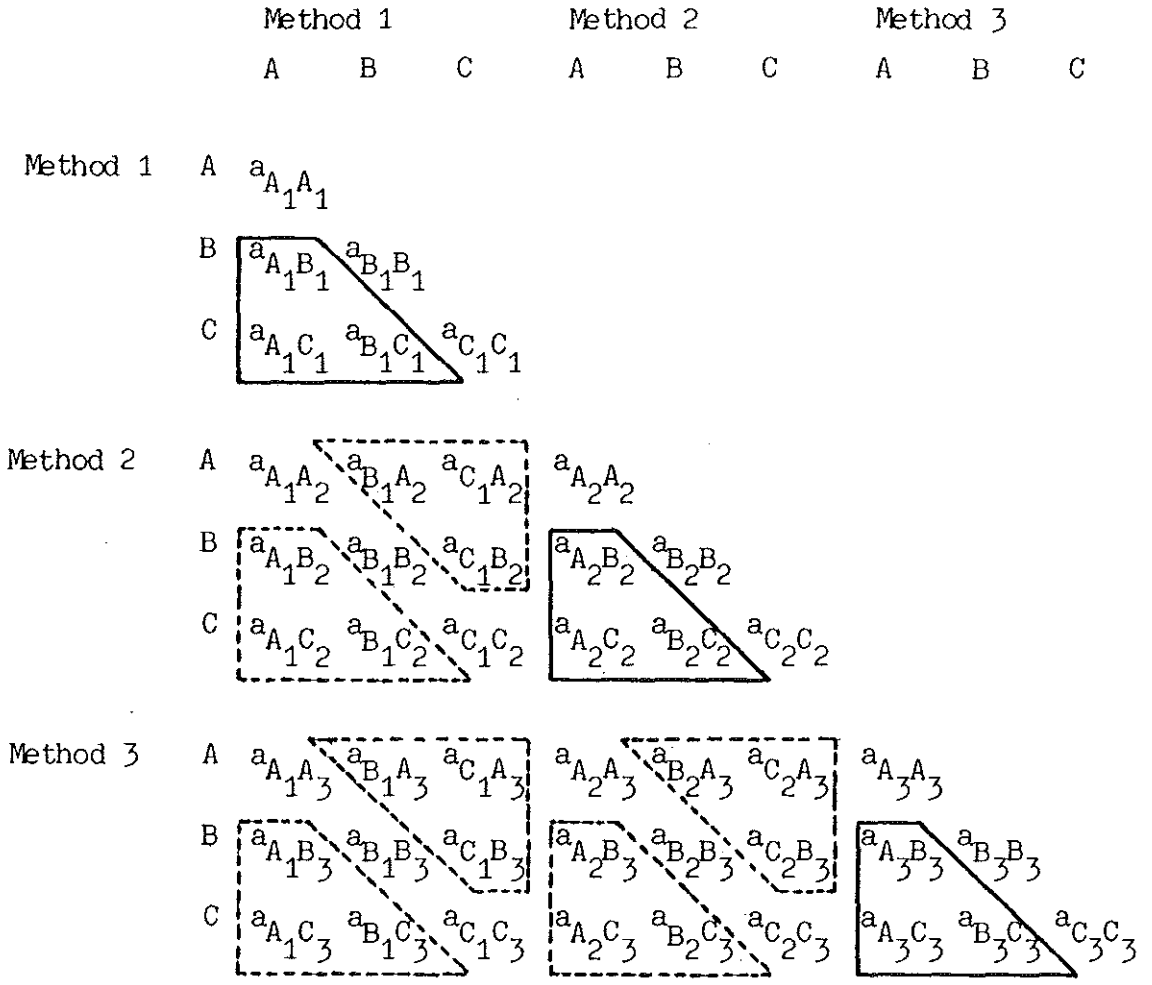
An additional advantage of the use of agreement measures in Multitrait-Multimethod Matrices is that the entries in the monotrait monomethod diagonale $a_{A_1A_1}$, $a_{B_1B_1}$,..., $a_{C_3C_3}$ are to be interpreted as internal consistency measures which have been identified above as the weakest form of reliability assessment: stability.

Thus, generally, predictive validity can be characterized as the degree to which findings obtained by one method conform to known facts of empirical significance as obtained by another method. Specifically, a content analysis may be said to be predictively valid, if its inferences can be shown to exhibit both, high agreement with the (past, present, or future) states or properties of the source claimed to be true and low agreements with characteristics of the source excluded by the same inferences.

9. Construct Validity

When content analysis procedures are designed de novo and are unique to a particular set of data or situation, pragmatic validation becomes impossible. Pragmatical validation requires at least some evidence about past successes and relies on sample sizes much larger than one. And yet, content analysts are quite often confronted with the need to provide valid inferences from a given body of text in unique and in a sense unpresidented situations.

The work of historians is most typically of this nature. Whether the statement "history never repeats itself" reflects a philosophical position or a historical fact, it is a position that many content analysts assume as well, and in assuming this to be the case statistical validation



Multitrait-Multimethod Agreement Matrix

Figure 7

procedures are practically ruled out. DIBBLE (60) who analysed the arguments made by historians in favor or against inferences drawn from documents came to the conclusion that they all involved assumptions about psychological characteristics of the observers, rules of the social system keeping the records, physical-social conditions surrounding the writing of the documents, etc. While historical documents and the inferences drawn from them are thought to be unique and outstanding, the assumptions linking a text with some event have the logical status of generalizations regarding documentary evidence.

GEORGE (61) who analysed FCC efforts during World War II to extract intelligence from foreign broadcasts came much to the same conclusion. While the areas of interest to the propaganda analyst are essentially variable and uncertain (why else would he want to know about it), inferences made in such apparently unique situations relied on patterns (linguistic, personality, social structure) that were known to be or assumed to be stable characteristics of the context of the data and either underlying or governing the variable events in question.

A simple and therefore most instructive example of construct validation in content analysis is LEITES, BERNANT and GARTLOFF's (62) analysis of speeches made by members of the Soviet politburo at the occasion of STALIN's 70th birthday. While all of the published speeches appeared to express the same adulation of STALIN, LEITES et al. hoped that a careful analysis of nuances in style and emphasis would shed some light on the power relations existing in the Kremlin. The problem of succession was of some interest to political analysts at that time, particularly since absence of formal rules for the transitions of power presented considerable uncertainty.

In this (statistically)unique situation, LEITES et al. could neither rely on past content analyses nor on generalizations from past power transfer. Instead they had to develop and justify an analytical construct that would link the politburo member's relative power position (nearness to STALIN) within the Kremlin with the way they addressed both STALIN and the public. The clue to such a construct was found in the Soviet use of language to express nearness. LEITES and his collaborators, all experienced sovietologists, discovered that Soviet political discourse provides two distinct approaches. One set of "symbols of nearness and intimacy(father, solicitude, etc.) appear most frequent in popular image of STALIN and (is) stressed for the audience which is far removed from him." The other set of symbols derives from the prevailing "depreciation of such nearness in political relationships. The ideal party member does not stress any gratification he may derive from intimacy for political ends. ... Those close to STALIN politically are

permitted to speak of him in terms of lesser personal intimacy (leader of the party, etc.) and are privileged to refrain from the crudest form of adulation. The relative emphasis on the Bolshevik image or on the popular image of STALIN (they conclude), therefore not only reflects the Bolshevik evaluation of the party as distinguised from and superior to, the masses at large, but also indicates the relative distance of the speaker from STALIN." (63) Compared with the lengthy logical derivation of the construct from existing theory, from literature, and more so from experiences, the task of counting the speakers' relative emphasis on the Bolshevik as opposed to the popular image, and the subsequent ranking of politburo members according to this emphasis was a minor task. The resulting picture with MOLOTOV, BERIA and MALENKOV closest to STALIN and a group including KHRUSHCHEV most distant to him, was supported by the by now well known struggle after STALIN's death.

The argument underlying the example and construct validation generally is again straight forward:

- a valid theory, established hypotheses or at least some defendable generalizations about the source are given,
- the construction of the analytical procedure (method) is logically derivable from that theory so that the analysis is in fact a valid operationalization of that theory,
- therefore the inferences now drawn from data by the method may be accepted on account of the underlying theory's independently established validity.

Thus, in construct validation of content analyses, validity derives entirely from established theory, tested hypotheses and generalizations about the source, whatever the evidential status of this knowledge might be at the time. It is these generalizations plus the logical derivation of the process and categories of analysis (operationalization of the construct) that are laid in the open to be challenged. Once this is accepted the findings cannot be doubted (at least not with the validating information going into construct validity). The validity of the findings from a content analysis can not exceed the validity of the theory underlying its analytical construction.

Obviously, when a content analysis is essentially unique, construct validation - the validation of the process of analysis rather than its input or result - is the only form of validation available to the analyst. While the events following STALIN's death corroborated LEITES et al.'s inferences, in my terms, lent some predictive validity to it, these events were not available at the time. All that was known went into the analytical construct. Construct validation is also the most productive way of developing

novel forms of analysis. Any analytical procedure, whether in form of a computer program or in form of instructions for manual data processing, might be said to be an operational model of the source under consideration. The more accurate this model is in representing the source the more accurate will the inferences be. Unless one can test this model repeatedly against available data (data oriented and pragmatic validation), validating information can only come from existing theory.

At this point I do not wish to further exemplify attempts to operationalize theories of meanings or attempts to justify existing computer programs in terms of available knowledge of cognition, etc., all of which are incidents of construct validation. But I do want to mention that the failure of content analysis procedures and constructs to correspond to or to be justifiable by existing theories and models of symbolic behavior presents a serious limitation to the validity of content analyses. For example, when a content analysis cuts a text into separate units while the reader responds to the connections between such units, valid inferences are not likely forthcoming on account of the procedure's lack of construct validity (64). Construct validity is an answer to why an analysis must be successful, pragmatical validity merely assesses whether it was.

I should like to add that my use of the term construct validity has to be somewhat more limited than in the Technical Recommendations. In applications that these recommendations consider, "construct validity is evaluated ... by demonstrating that certain explanatory constructs account to some degree for (the individual's) performance on the test." The Technical Recommendations conceptualize the validation as a two-way process: "First the investigator inquires: From (the) theory (underlying the test), what predictions would he make regarding the variation of scores from person to person or occasion to occasion? Second, he gathers data to confirm these predictions" (65). My emphasis on a one way process of validation is merely born out of the nature of content analysis as a method for making specific inferences from symbolic data. When such inference attempts are unique only the Technical Recommendations' first step can be completed: The justification of the procedure and categories of analysis (test construction) from a valid theory. While it would be undeniably desirable to proceed to the recommendation's second step and validate the underlying theory in return, most content analysts are not given the opportunity to do so. It is the absence of this opportunity that accounts for the heavy emphasis on the logical part of construct validation in content analysis.

10. Conclusion

Let me conclude by saying that problems of validation have become a major stumbling block for content analyses to have practical implications. Unlike their colleagues in the natural sciences and even in economics and in experimental psychology, content analysts have to cope with meanings, contextual dependencies and symbolic qualities of their data which makes validation different from where such phenomena are absent. Hopefully, the proposed typology and rudimentary measures for different kinds of validity provide a means by which at least part of an analytical effort can be channeled more successfully than in the past into making content analyses more acceptable in practice. But, this effort will also serve scientific purposes for any science advances with an increased awareness of how its methods contribute to and constitute the understanding of reality and, ultimately, of science itself.

Notes

1 One of "Top three Papers in Mass Communication" of those presented at the 1978 International Communication Association Meeting in Chicago, Illinois.

2 KRIPPENDORFF, Content Analysis; An Introduction to its Methodology, in preparation.

3 LASSWELL 1965.

4 GEORGE 1959.

5 KATZ et al. 1973.

6 SINGER 1963.

7 see TENNEY 1912.

8 ZAPF 1974.

9 GERBNER 1969.

10 BROUWER et al. 1969.

11 see contributions from GRÜNZIG and also KLINGEMANN and SCHÖNBACH in this volume.

12 KRIPPENDORFF 1967, pp. 313-316.

13 e.g. CRONBACH and MEEHL 1955; CAMPBELL and FISKE, 1959.

14 KRIPPENDORFF 1967, pp. 130-167; 1969a, p.8.

15 JANIS, The Problem of Validating Content Analysis, 1965.

16 GEORGE, Propaganda Analysis, 1959.

17 STEWART, Importance in Content Analysis: A Validity Problem, 1943.

18 FLESCH 1951.

19 HOLSTI and NORTH 1966.

20 CAMPBELL, Factors Relevant to the Validity of Experiments in Social Settings, 1957.

21 KRIPPENDORFF, Reliability in Message Analysis, 1973.

22 JANIS, The Problem of Validating Content Analysis, 1965.

23 JANIS, op. cit., p. 70.

24 JANIS, op. cit., p. 65.

25 American Psychological Association, 1954, p. 13.

26 SELLTIZ, JAHODA, DEUTSCH and COOK, 1963, p. 157.

27 FEIGL 1952.

28 KRIPPENDORFF, 1969 a, p. 12.

29 OSGOOD 1959.

30 GEORGE 1959.

31 POOL 1959, pp. 212-216.

32 HOLSTI 1969, pp. 114-116.

33 HOLSTI and NORTH 1966.

34 OSGOOD 1959, pp. 58-61.

35 see discussion in ALLPORT, 1965.

36 JANIS 1943.

37 BERELSON and LAZARSFELD 1948, p. 6.

38 CARTWRIGHT 1953, p. 424.

39 MILLER 1963, p. 96.

40 KRIPPENDORFF 1970.

41 OSGOOD, SUCI, TANNENBAUM 1967, p. 141.

42 DUNPHY 1966, p. 159, Figure 4.4

43 OSGOOD 1959, pp. 73-77.

44 BERELSON 1952, p. 14.

45 HOLSTI 1969, p. 14.

46 STONE et al. 1966.

47 KOPS, Auswahlverfahren in der Inhaltsanalyse, 1977.

48 KOPS, op. cit., pp. 230-240.

49 STONE and HUNT 1963.

50 STONE et al. 1966.

51 POOL 1959.

52 HOLSTI 1969.

53 CAMPBELL and FISKE 1959.

54 CAMPBELL and FISKE 1959.

55 ALWIN 1974.

56 MORTON, SARIS-GALLHOFER and SARIS 1974.

57 JANIS 1965, p. 61.

58 GEORGE 1959.

59 KRIPPENDORFF 1973.

60 DIBBLE, Four Types of Inferences from Documents to Events, 1963.

61 GEORGE 1959.

62 BERNANT and GARTLOFF, 1950-51.

63 LEITES, BERNANT, GARTLOFF 1950-51, pp. 317-339.

64 KRIPPENDORFF 1967.

65 Technical Recommendations 1954, p. 14.

References

American Psychological Association:
1954 "Technical Recommendations for Psychological and Diagnostic Techniques." Psychological Bulletin Supplement, 51, 2: 200-238

ALLPORT, Gordon W.:
1965 Letters from Jenny. New York: Harcourt, Brace & World

ALWIN, Duane F.:
1974 "Approaches to the Interpretation of Relationships in the Multitrait-Multimethod Matrix." Pp. 79-105 in Herbert L. COSTNER (Ed.) Sociological Methodology 1973-1974, San Francisco, California: Jossey-Bass

BERELSON, Bernard:
1952 Content Analysis in Communication Research. Glencoe, Illinois: Free Press

BERELSON, Bernard and Paul F. LAZARSFELD:
1948 The Analysis of Communication Content, Preliminary Draft. Chicago and New York: University of Chicago and Columbia University

BROUWER, Marten, Cederic C. CLARK, George GERBNER, and Klaus KRIPPENDORFF:
1969 "The Television World of Violence." Pp. 311-339, 519-591 in Robert K. BAKER and Sandra J. BALL (Eds.), Mass Media and Violence, Vol. IX, A Report to the National Commission on the Causes and Prevention of Violence. Washington, D.C.: U.S. Government Printing Office

CAMPBELL, Donald T.:
1957 "Factors Relevant to the Validity of Experiments in Social Settings." Psychological Bulletin, 54,4: 297-311

CAMPBELL, Donald T. and Donald W. FISKE:
1959 "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." Psychological Bulletin, 56,2: 81-105

CARTWRIGHT, Darwin P.:
1953 "Analysis of Qualitative Material." Pp. 421-470 in Leon FESTINGER and Daniel KATZ (Eds.), Research Methods in the Behavioral Sciences, New York: Holt, Rinehart & Winston

CRONBACH, Lee L. and Paul E. MEEHL:
1955 "Construct Validity in Psychological Tests." Psychological Bulletin, 52,2: 281-302

DIBBLE, Vernon K.:
1963 "Four Types of Inferences from Documents to Events." History and Theory, 3,2: 203-221

DUNPHY, Dexter C.:
1966 "The Construction of Categories for Content Analysis Dictionaries." Pp. 134-168 in Philip J. STONE, Dexter C. DUNPHY, Marshall S. SMITH, and Daniel M. OGILVIE, The General Inquirer: A Computer Approach to Content Analysis. Cambridge, Mass. MIT Press

FEIGL, Herbert:
1952 "Validation and Vindication: An Analysis of the Nature and the Limits of Ethical Arguments," in Wilfrid SELLARS and John HOSPERS (Eds.), Readings in Ethical Theory. New York: Appleton-Century-Crofts, Inc., pp. 667-680

FLESCH, Rudolf:
1951 How to Test Readability. New York: Harper & Brothers

FLETCHER, Rudolf:
1948 "A New Readability Yardstick." Journal of Applied Psychology, 32: 221-237

GEORGE, Alexander L.:
1959 Propaganda Analysis: A Study of Inferences Made from Nazi Propaganda in World War II. Evanston, Illinois: Row, Peterson

GERBNER, George:
1969 "Toward 'Cultural Indicators': The Analysis of Mass Mediated Public Message Systems." Pp. 123-132 in George GERBNER, Ole R. HOLSTI, Klaus KRIPPENDORFF, William J. PAISLEY, and Philip J. STONE (Eds.), The Analysis of Communication Content. New York: Wiley & Sons

HOLSTI, Ole R.:
1969 Content Analysis for the Social Sciences and Humanities. Reading, Mass: Addison Wesley

HOLSTI, Ole R. and Robert C. NORTH:
1966 "Perceptions of Hostility and Economic Variables." Pp. 169-190 in Richard MERRITT and S. ROKKAN (Eds.), Comparing Nations. New Haven: Yale University Press

JANIS, Irwin L.:
1943 "Meaning and the Study of Symbolic Behavior." Psychiatry, 6: 424-439

JANIS, Irwin L.:
1965 "The Problem of Validating Content Analysis." Chapter 4, pp. 55-82 in Harold D. LASSWELL, Nathan LEITES et al. (Eds.), Language of Politics, Cambridge, Massachusetts: MIT Press

KATZ, Phillip, Michael M. LENT and Eric J. NOVOTNY:
1973 Survey of Chinese Mass Media Content in 1972: A Quantitative Analysis. Kensington MD: American Instituts for Research, November

KOPS, Manfred:
1977 Auswahlverfahren in der Inhaltsanalyse. Meisenheim/Glan, Germany: Anton Hain

KRIPPENDORFF, Klaus:
1967 "The Goal of Message Analysis." Pp. 130-166 in his An Examination of Content Analysis, Urbana: University of Illinois, Ph.D. dissertation

KRIPPENDORFF, Klaus:
1969a "Theories and Analytical Constructs, Introduction." Pp. 3-16 in George GERBNER, Ole R. HOLSTI, Klaus KRIPPENDORFF, William J. PAISLEY, and Philip J. STONE (Eds.), The Analysis of Communication Content. New York: Wiley & Sons

KRIPPENDORFF, Klaus:
1969b "Models of Messages: Three Prototypes." Pp. 69-106 in George GERBNER, Ole R. HOLSTI, Klaus KRIPPENDORFF, William J. PAISLEY, and Philip J. STONE (Eds.), The Analysis of Communication Content, New York: Wiley & Sons

KRIPPENDORFF, Klaus:
1970 "The Expression of Value in Political Documents." Journalism Quartlerly, 47, 3: 510-518

KRIPPENDORFF, Klaus:
1973 Reliability in Message Analysis. Philadelphia: University of Pennsylvania, the Annenberg School of Communications, Mimeo

KRIPPENDORF, Klaus:
Content Analysis; An Introduction to its Methodology. San Francisco: Sage, in preparation

LASSWELL, Harold D.:
1965 "Propaganda Detection and the Courts." Pp. 173-232 in Harold D. LASSWELL, Nathan LEITES et al., Language of Politics. Cambridge, Massachusetts: MIT Press

LEITES, Nathan, Elsa BERNANT, and R.L. GARTLOFF:
1950-51 "Politburo Images of Stalin." World Politics, 3: 317-339.

MAHL, George F.:
1959 "Exploring Emotional States by Content Analysis." Pp. 89-130 in Ithiel de Sola POOL (Ed.), Trends in Content Analysis. Urbana: University of Illinois Press

MILLER, George A.:
1963 Language and Communication. New York: Mc Graw-Hill.

MORTON, Elaine L., Irmtraud SARIS-GALLHOFER, and William E. SARIS:
1974 "A Validation Study of HOLSTI's Content Analysis Procedure." Mimeo

OSGOOD, Charles E.:
1959 "The Representational Model and Relevant Research Methods." Pp. 30-88 in Ithiel de Sola POOL (Ed.), Trends in Content Analysis. Urbana: University of Illinois Press

OSGOOD, Charles E., George J. SUCI, and Percy H. TANNENBAUM:
1967 The Measurement of Meaning. Urbana: University of Illinois Press

POOL, Ithiel de Sola:
1959 "Trends in Content Analysis: A Summary." Pp. 189-233 in his (Ed.), Trends in Content Analysis. Urbana: University of Illinois Press

SELLTIZ, Claire, Marie JAHODA, Morton DEUTSCH, and Stewart W. COOK:
1963 Research Methods in Social Relations. New York: Holt, Rinehart and Winston

SINGER, J. David:
1963 "Media Analysis in Inspection for Disarmament." The Journal of Arms Control, 1: 248-260

STEWART, Milton D.:
1943 "Importance in Content Analysis: A Validity Problem." Journalism Quarterly, 20: 286-293

STONE, Philip J. and Earl B. HUNT:
1963 "Computer Approach to Content Analysis: Studies Using the General Inquirer System." Proceedings, Spring Joint Computer Conference: 241-256

STONE, Philip J., Dexter C. DUNPHY, M.S. SMITH, and D.M. OGILIVIE:
1966 The General Inquirer: A Computer Approach to Content Analysis in the Behavioral Sciences. Cambridge: MIT Press

TENNEY, Alran A.:
1912 "The Scientific Analysis of the Press." The Independent, 73: 895-898

United States Senate:
1972 Hearings before the Subcommittee on Communications of the Committee on Commerce, March 21-24, 1972, Serial No. 92-52. Washington, D.C.: U.S. Government Printing Office

ZAPF, Wolfgang:
1974 "The Polity as a Monitor of the Quality of Life." American Behavioral Scientist, 17,5: 651-675