Quality and Quantity, 17 (1983) 127–149
Elsevier Scientific Publishing Company, Amsterdam — Printed in The Netherlands

Measurement Models for Content Analysis

ROBERT PHILIP WEBER
Harvard University

Making inferences from a symbolic medium—usually text—is the essence of content analysis. The rules of this inferential process have not yet been formalized rigorously; however, the fundamental procedure is to make inferences about some characteristic of a message, its source and/or its audience from its content (Stone et al., 1966; Holsti, 1969; Markoff et al., 1974; Krippendorff, 1980) [1]. This paper addresses fundamental methodological problems in the analysis of text by computer-aided content analysis based on word counts [2]. Operationally, word-count content analysis entails the mapping of the many words in documents or other texts into much fewer content categories. Scores representing the relative frequencies [3] of these categories in each document are usually the basic variables in subsequent analyses. The coding rules for mapping words are frequently contained in a thesaurus-like dictionary which can be read by computer. General-purpose dictionaries, such as the Harvard IV (Kelly and Stone, 1975; Dunphy et al., 1974) and the Lasswell Value Dictionary (Lasswell and Namenwirth, 1968; Namenwirth and Weber, 1974), consist of a list of several thousand words and the categories to which they have been assigned. Details and problems of dictionary construction are discussed later in this paper.

An important use of content analysis is the generation of reliable and valid cultural indicators (e.g., Rosengren, 1981; Namenwirth, 1969a,b, 1970, 1973; Namenwirth and Bibbee, 1973, 1976; Namenwirth and Weber, 1980; Weber, 1981, 1982a; Mohler, 1978; Klingemann et al., 1982a,b). In literate societies, a large portion of culture is represented in texts such as newspapers, political documents, books, and scripts from radio, television and film. Cultural indicators generated from such texts constitute an essential set of variables for the study of cultural dynamics: for example, changes in ideology or political agenda. Furthermore, generating cultural indicators from large amounts of text has been made significantly easier and less costly because of recent innovations in optical character reading [4] and in capturing text from electronic media such as newswires, newspaper editing systems (De Weese, 1977), word-processing systems and text-format cable and television broadcasting (teletext) systems. Content-analytic cultural indicators

0033-5177/83/0000–0000/$03.00

have recently been discussed at length (Weber, 1982b), and are not addressed in detail here.

Compared with work on social indicators (e.g., Bauer, 1966; Sheldon and Moore, 1968; Land and Spillerman, 1975; Wilcox, 1972; Carley, 1981), research has lagged not only in cultural indicators, but in computer-aided content analysis as well. After a promising start in the 1960s (e.g., Stone et al., 1966; Gerbner et al., 1969), computer-based content analysis—indeed, almost all content analysis—was virtually abandoned by American social science. Many factors contributed to this decline [5]. One was the premature conclusion that artificial intelligence [6] would quickly render obsolete the General Inquirer (Stone et al., 1966; Kelly and Stone, 1975) and other word-count approaches (e.g., Iker, 1969; Cleveland et al., 1974). Artificial intelligence (AI) has important uses such as the representation and investigation of human cognition. Given current limitations, however, AI approaches to content analysis are not yet feasible for generating a diverse set of cultural indicators from very large amounts of text. Hence, AI should be viewed as a set of models and techniques that complement rather than supplant word-count content analysis.

A second problem was that the General Inquirer (hereafter GI) and similar strategies were often applied to substantive questions for which, in my opinion, they were not well suited (and for which AI approaches may well be the preferred strategy). These questions involved psychological and social-psychological problems in which the investigator wished to make inferences concerning the subjective emotional and cognitive states of individuals (cf. Ogilvie, 1966) [7]. Consequently, there did not develop a large body of content-analytic results integrated with both theory and results from other methods. As a result, researchers abandoned content analysis for more productive grounds.

One of the most important problems, however, was the failure to approach disputes concerning rival techniques or strategies for content analysis within an integrated framework. For example, questions involving the analysis of word counts rather than category counts, the use of dictionaries with a priori category schemes rather than categories inferred via factor analysis from the covariation of high-frequency words, and the classification of words into single rather than multiple categories were debated. These and other disputes were seldom resolved. Without a methodological framework, the ad hoc intellectual arguments that were proposed were often viewed as arcane by nonspecialists, who left the debate and the field to the experts.

An equally serious difficulty was the failure to investigate empirically methodological problems of content analysis. Unlike well-known efforts in survey research, for example, to study the problems of alternative question wordings, scale construction, sampling techniques or telephone interviews,

there was little or no systematic attempt to study the consequences of alternative operational procedures for content classification and analysis. Thus, it was never ascertained whether analyzing inferred or a priori content categories leads to different or similar interpretations. Similarly, there was no systematic effort to determine whether different a priori content classification schemes lead to similar or conflicting results [8]. Consequently, there was little empirical evidence to support the rival claims made by various factions.

This paper reconceptualizes several methodological problems of word-count content analysis in terms of structural-equation models and suggests a framework for future research based on measurement models for content analysis. To achieve these ends, several neglected problems in word-count content analysis are recast, including: (1) category reliability and validity; (2) single versus multiple classification dictionaries; (3) a priori versus inferred categories; (4) measurement models and levels of aggregation; and (5) the consequences of different dictionaries for the substantive results. These problems are discussed in detail below.

The principal frame of reference is the analysis of covariance structures as developed by Jöreskog and embodied in the LISREL computer program [9]. The LISREL model serves as a metalanguage in whose terms the above issues are defined and investigated. The LISREL approach to the analysis of covariance structures is employed because it is quite general and can handle a variety of models. In addition, the LISREL model is becoming increasingly well known within the social-science community, thus making it easier to communicate these models and the issues at stake. A brief overview of the LISREL model is presented next.

Measurement Models for Content Analysis

The general LISREL model (Jöreskog and Sörbom, 1979, 1980, 1981) consists of two parts: the measurement model and the structural-equation model. The primary emphasis here is on the measurement model, which indicates how latent or unobserved variables are related to observed variables. The measurement model specifies the measurement properties of the observed and latent variables. The structural-equation model specifies the causal relationships among the latent variables and indicates the causal effects and unexplained variance. The structural-equation aspect of the LISREL model is used later in this paper to define second-order confirmatory factor-analysis models. In the models defined below, the observed variables are usually words or word senses [10], while the latent variables are content categories and/or themes in texts. The covariance matrices to be

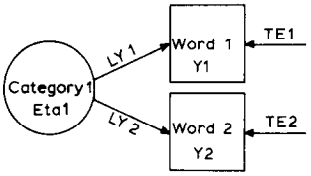

Fig. 1. Measurement Model for Two Words and One Category.

analyzed are formed with the number of observations equal to the number of documents or other units of text (e.g., paragraphs), and with each variable in the matrix representing the relative frequency of a word in each document.

The measurement model for the endogenous (dependent) variables is given in matrix notation by

$$\mathbf{Y} = \Lambda_y \eta + \epsilon$$

where $\mathbf{Y}$ is a vector of observed variables, Λ_y are the factor loadings of the observed variables on the latent variables η, and ϵ are the measurement errors. Figure 1 illustrates a simple measurement model of this type. In a similar fashion, the measurement model for the X's, or exogenous (independent) variables, is given by

$$\mathbf{X} = \Lambda_x \xi + \delta$$

where $\mathbf{X}$ is a vector of observed variables, Λ_x are the factor loadings of the observed variables on the latent independent variables ξ, and δ are the measurement errors for the X variables.

The structural model is given by

$$\beta\eta = \Gamma\xi + \zeta$$

where β is the matrix of causal coefficients among the dependent variables η, Γ a matrix of the causal coefficients linking the independent and dependent variables, ζ a matrix of the residuals of the dependent variables, and ξ and η are as noted before. Figure 2 illustrates a simple structural-equation–measurement model with two measured X variables, two measured Y variables, and one latent independent and dependent variable.

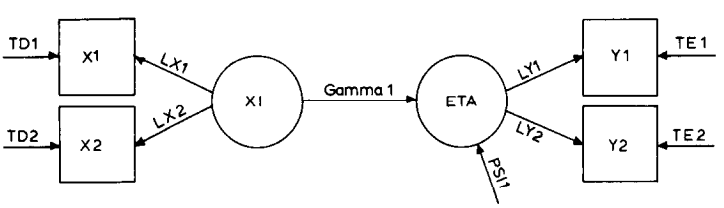

Fig. 2. Elementary Causal Model, with Two Latent and Four Observed Variables.

Let **S** be the covariance matrix among all the X's and Y's, that is, among all observed variables. The LISREL program estimates a predicted covariance matrix $\boldsymbol{\Sigma}$ as a (complicated) function of eight parameter matrices which may contain both fixed or predetermined elements and unknown coefficients to be estimated. LISREL produces maximum-likelihood estimates of these parameters under the assumption of multivariate normality [11]:

(1) lambda Y (LY): the factor loadings of the observed Y's on the unobserved dependent variables;

(2) lambda X (LX): the factor loadings of the observed X's on the unobserved independent variables;

(3) theta delta (TD): the covariance matrix of δ, the residuals or error term for the measurement model of the latent independent variables;

(4) theta epsilon (TE): the covariance matrix of ϵ, the residuals or error term for the measurement model of the latent dependent variables;

(5) beta (BE): the causal coefficients among the dependent variables;

(6) gamma (GA): the causal coefficients linking the dependent and independent variables;

(7) phi (PH): the covariance matrix of the latent independent variables; and

(8) psi (PS); the covariance matrix of the residuals in the structural model.

In addition to estimating the measurement and structural-equation models simultaneously, another important aspect of the LISREL model is that it indicates how well the predicted covariance matrix $\boldsymbol{\Sigma}$ reproduces or fits the observed covariance (or correlation) matrix **S**. This indicator is an approximation to χ^2 with appropriate degrees of freedom for the number of free and constrained parameters. The lower χ^2, the better the fit of the model to the data [12]. Contrasted with ordinary regression techniques, which assess only the parameter estimates and their significance, the ability to test the overall fit of a model using LISREL represents a significant advance in model testing and estimation. Furthermore, two models may be compared by subtracting the χ^2 values and degrees of freedom. The difference in the χ^2's is itself a χ^2 statistic whose significance indicates whether the models are, in a statistical sense, significantly different. On the basis of the work of Tucker and Lewis (1973), Bentler and Bonett (1980) proposed a goodness-of-fit measure ρ for maximum-likelihood estimates that is independent of the degrees of freedom. Their approach also yields a similar statistic for comparing the differences in fit of two models. Thus LISREL provides a powerful method of model estimation, comparison and revision.

Finally, not all models that can be stated in terms of LISREL can be estimated using real data. Such models are those having too many unknown parameters or parameters which cannot be determined uniquely (i.e., the model is underidentified) [13]. However, restricted models may be estimated by imposing reasonable constraints on such models.

Category Reliability and Validity Problems

Two of the most important problems in content analysis are the reliability and validity of content-analysis dictionaries.

RELIABILITY

Reliability in its broadest sense refers to the consistency of measurements. Prior to computer-aided content analysis, the principal reliability problems stemmed from the consistency with which human coders applied rules for classifying words in a text. Coder reliability problems were solved by the use of computers, which required as a first step the formalization of coding rules. Once formalized, the rules would be applied consistently by a valid computer program.

Some of the remaining sources of error in computer-aided content analysis concern the formalization of the rules to be applied by the computer. There are two major difficulties: first, category definitions may be ambiguous, so that some words are erroneously assigned to categories; second, words themselves are often ambiguous. We first consider the limitation of category definitions.

The usual procedure when coders disagree is to estimate the extent of their agreement (reliability) and to accept or reject the coding on the basis of some level thereof. This procedure was not followed in the construction of the Lasswell Value Dictionary (Lasswell and Namenwirth, 1968; Namenwirth and Weber, 1974) or the latest Harvard dictionary (Dunphy et al., 1974). In both instances, it was felt that there should be substantial agreement once the rules of classification were known. Consequently, whenever disagreement occurred the coders were asked to resolve their dispute in terms of commonly accepted standards. Although disputes were resolved in all cases, the decision rules frequently remained implicit or ad hoc. Consequently, the rules of classification are contained explicitly only in the actual assignments of words to categories, and are represented only partly in the definitions of these groupings.

The ambiguity in the meaning of words needs only little illustration. Take the word "kind". What kind of word is that? In one sense it refers to a class of objects, while in another it describes or evokes a benevolent disposition. Another example is the word "just". Often it denotes a concern with ethics, but more frequently it means just something else, if it did not just happen. How are these riddles resolved?

As part of the General Inquirer (Kelly and Stone, 1975), the latest versions of the Harvard dictionary and the Lasswell Value Dictionary incorporate rules for distinguishing among the various senses of homo-

graphs, i.e., words with more than one sense or meaning. These "disambiguation routines" were validated using broad samples of text (Kelly and Stone, 1975), and they can be easily modified as necessary. In addition to the Kelly–Stone procedures, the Dutch linguist Boot (1974, 1977a–c, 1978a,b) has developed computer methods for the disambiguation of Dutch, German and English texts.

Another source of error stems from the common practice of single classification; that is, assigning each word sense to only one basic category and then weighting equally all words assigned to the same category. But do all words in a category reflect that category to the same extent? For example, in their economic senses, do the words "bank", "interest", "buy" and "investment" equally indicate concern with wealth or economic matters? And, if not, what are the consequences for reliability and validity?

Even though there has been substantial progress in error reduction, the amount of error remaining and its sources have yet to be investigated empirically. The use of general-purpose dictionaries would rest on firmer ground if empirical evidence were available regarding the measurement properties of content categories.

If category construction is viewed as a problem in scale or test construction, then according to classical test theory (e.g., Lord and Novick, 1968; Jöreskog, 1974) there are three possible measurement models for test or category construction: (1) parallel measures; (2) τ-equivalent measures; and (3) congeneric measures. Figure 1 represents one scale with two measures. Each of the classical test measurement models imposes a different set of restrictions. Specifically, the parallel test model is given in equation form by

$$y_1 = \lambda_{11}\eta_1 + \epsilon_1$$

$$y_2 = \lambda_{21}\eta_1 + \epsilon_2$$

$$\lambda_{11} = \lambda_{21}$$

$$\sigma^2(\epsilon) = 1$$

$$\sigma^2(\epsilon_1) = \sigma^2(\epsilon_2)$$

where the y's are observed variables, the λ's are factor loadings, the η's are latent variables, and the ϵ's are errors.

The parallel test model states that each item measures the scale to the same extent ($\lambda_{11} = \lambda_{21}$), that the variance of the scale is unity, and that both items are fallible to the same extent ($\sigma^2(\epsilon_1) = \sigma^2(\epsilon_2)$).

The τ-equivalent measurement model is similar, but without the requirement of equal error variances; that is, $\sigma^2(\epsilon_1)$ does not have to equal $\sigma^2(\epsilon_2)$. The congeneric measurement model places no restrictions on either the factor loadings (the λ's), the variance of the scale (η_1), or the error structure.

For the model in Fig. 1, the congeneric measurement model is

$$y_1 = \lambda_{11}\eta_1 + \epsilon_1$$

$$y_2 = \lambda_{21}\eta_1 + \epsilon_2$$

To the best of my knowledge, these measurement models have never been estimated using computer-aided content-analysis data. These models can be estimated via LISREL with appropriate parameter constraints. The observed variables would be medium- and high-frequency word senses [14] and the latent variables would be content categories. The measurement model in Fig. 1 is elaborated substantially below in the context of validity.

Using the best-fitting test model, the reliability of each category can be calculated by summing the squared factor loadings λ. However, another procedure is required to obtain an estimate of reliability for the entire set of categories included in the model. Given that the LISREL measurement model is the same as a first-order common factor-analysis model but with added constraints, the Ω (Heise and Bohrnstedt, 1970) reliability coefficient can be calculated from the LISREL results.

Another powerful feature of LISREL is that the same model may be estimated simultaneously for more than one group or set of documents. Furthermore, some or all of the parameters to be estimated may be constrained to be equal across sets of documents, or they may be permitted to vary completely across sets. The equality of coefficients across groups is tested by comparing the goodness-of-fit χ^2 when parameters are constrained to be equal, with the χ^2 when they are permitted to vary across groups. This provides a powerful facility for determining which parameters are most sensitive to the particular set of documents analyzed, thereby giving an indication of reliability across sets of documents.

VALIDITY

In addition to questions concerning reliability, content analysis has always faced difficult validity problems. Validity refers to the extent to which the theoretical concept it is intended to measure is actually measured. It is useful to distinguish three major types of validity: (1) criterion validity; (2) face or content validity; and (3) construct validity.

A measure has criterion validity to the extent that it predicts some behavior. For example, if it was desired to create a test which predicted academic performance in college, then that measure would have criterion validity to the extent that it was in fact correlated with a variable such as grade point average. But content categories reflect meaning or shared understandings of language. And what variables can be used as external criteria for content categories? It would seem that there are no external criterion

variables for content categories, but the issue is not clear-cut. For example, in pilot research designed to cross-validate some of the categories of the Lasswell Value Dictionary for German language text, Klingemann et al. (1982b) found a strong relationship between economic fluctuations and concern with "wealth" categories in the speeches of the Kaiser over the period 1870–1914. This finding is consistent with earlier results based on American political documents and economic performance (Namenwirth, 1969b). Is this criterion validity, and if so, what criterion variables exist for categories such as "love" or "uncertainty"? Even if some categories can be criterion-validated, most cannot. Therefore, content analysis must rest on a different sort of validation.

Until now, content analysis has relied heavily on face or content validity. A measure is content-valid to the extent that the contents of the items or the meanings of the words in a scale or category appear to measure what is intended. Although great care was taken in the construction of the Harvard and Lasswell dictionaries, some remain unconvinced of the validity of these instruments. Part of this scepticism stems from the ambiguities of category definitions and word senses noted above, especially in borderline cases, of which there are many.

A measure of a theoretical variable has construct validity if it "behaves" as the concept it measures should. The most powerful form of construct validity is external construct validity. With respect to content analysis, external validity means that content variables are related to other phenomena in accordance with a theory or model. In a number of studies involving cultural indicators based on long series of documents, the results have been consistent with pertinent interpretations of social, political and economic change in America (Namenwirth, 1969b, 1973; Namenwirth and Lasswell, 1970), Great Britain (Weber, 1981, 1982a), Sweden (Rosengren, 1981) and Germany (Mohler, 1978; Klingemann et al., 1982b).

In addition to external construct validity, measures may have internal construct validity. For example, if there are available various measures of alienation, and the data are consistent with a measurement model that presumes nine subscales, then judged on internal criteria these measures have internal validity. A content-analysis dictionary has internal construct validity if data based on several sets of texts are consistent with a measurement model for that dictionary. Such a measurement model is proposed just below. Indeed, a main objective of the research outlined in this paper is the assessment of the internal validity of content-analysis dictionaries.

In the above discussion of category reliability, the measurement model illustrated in Fig. 1 decomposed variation in relative word frequency into two components: variance in common with other word senses that comprise a content category, and random measurement error. This measurement

model is an oversimplification in two respects. First, it fails to take into account systematic sources of error, that is, variance which can be measured reliably but which is not a result of the construct whose measurement is intended. Second, the model ignores the causes of category variation. Failure to account for these sources of variance constitutes mis-specification of the measurement model.

Krippendorff (1980, p. 121) and others argue that procedures such as the GI are flawed because they ignore the larger semantic context of the words analyzed. This assertion is in error in at least three ways. First, the disambiguation rules for homographs discusssed above are based on the usage of words within sentences. In addition to syntactical information, the GI uses information about other words in a sentence to distinguish the various senses of homographs. Furthermore, a group of words that constitute a semantic unit can be counted as a single occurrence: for example, phrases such as

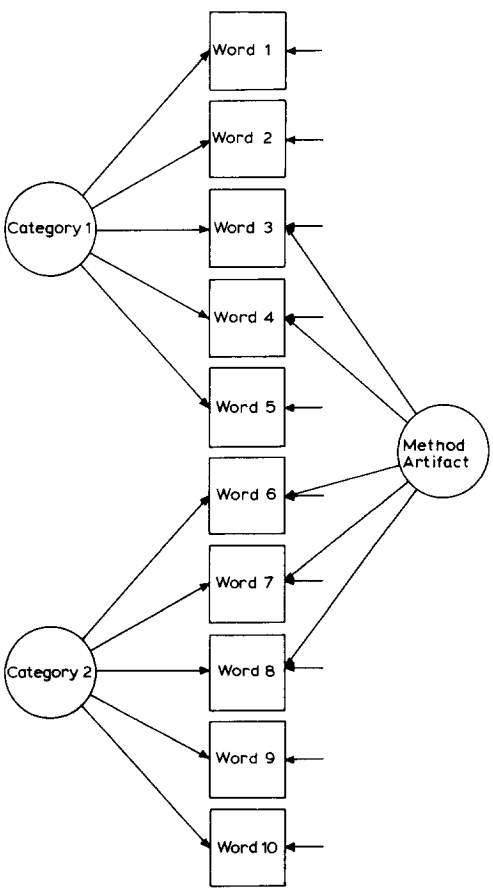

Fig. 3. Measurement Model for Words, Categories and a Method Artifact.

"Quality and Quantity" or "Center for Surveys, Methods and Analysis", and idioms such as "point of view", "broken heart" or "have in common".

Second, the usage of words in a text often indicates concern with a particular issue, theme or message. On other occasions word usage reflects stylistic considerations or syntactical necessity. In a given set or type of document, this stylistic and syntactical usage may be reflected in systematic rather than random variance. To the extent that this variance is systematic, it will adversely affect the estimation of content-analytic measurement and causal models. Word-count techniques do not directly model stylistic or syntactical sources of variance. Hence this systematic source of variance is a method artifact. Figure 3 presents a measurement model for two categories with the addition of a method artifact that causes variation in some but not all word senses. The existence of this method artifact is a hypothesis which remains to be tested. The model in Fig. 3 may be estimated using LISREL if

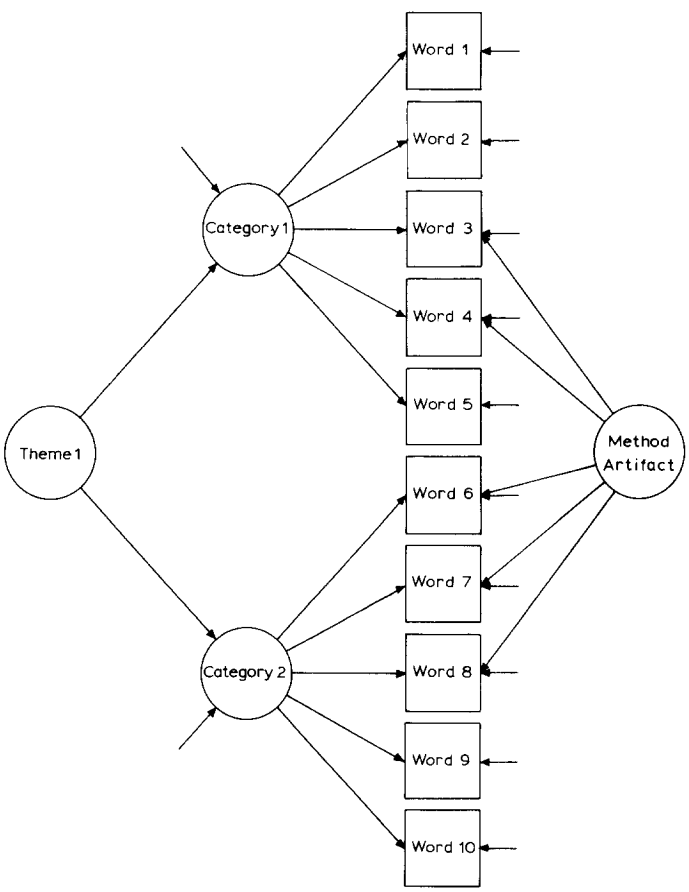

Fig. 4. Measurement Model for Words, Categories, a Method Artifact and a Theme.

appropriate constraints are imposed to permit identification of the model [15].

Finally, Krippendorff is in error in one other respect. It is crucial not to lose sight of the fact that the words analyzed were written or spoken to communicate some message about some issue or set of issues. The models above imply that word usage occurs in the abstract, that is, independent of concern with particular issues or themes. Consider Fig. 4, which represents a measurement–causal model of word-sense variation for a particular set of texts. This model includes the method factor noted just above, but also includes one latent variable representing a theme.

This model is a confirmatory second-order factor-analysis model and incorporates several hypotheses, including: (1) that observed variation in word senses can be decomposed into reliable and valid category variance, random measurement error, and reliable method variance; (2) that variation in the latent category variables is caused by concern with (latent) themes and by random error; (3) that variation in latent themes is measured without error; and (4) that, taking into account measurement error and the effects of themes, the categories are uncorrelated [16]. The causes of concern with themes are not included in this model. To the extent that this measurement model fits the data for several sets of texts, then the Lasswell and Harvard dictionaries (or other dictionaries) have both reliability and internal validity.

Single versus Multiple Word Classification

In classifying a word into a particular dictionary category, one is really answering the question of whether the entry generally has a certain attribute (or set of interrelated attributes). There are two answers to this question. Yes, the entry does, and it is therefore thus classified. Or, the answer is no, and therefore the entry is not classified under this heading. This formulation indicates two complications. In the first place, having one attribute does not logically exclude the possession of another. Secondly, not all entries need have the same attribute to the same extent. The qualities in terms of which words are classified may be continuous rather than dichotomous, thus leading to variation in intensity. Intensity is taken into account in the preceding model, and is indicated by the magnitudes of the factor loadings. Double or multiple classification of entries resolves the first problem, but creates others. For the Lasswell dictionary it was decided that the gain in semantic precision would not outweigh the loss of logical distinctiveness and exclusiveness (Namenwirth and Weber, 1974; Lasswell and Namenwirth, 1968). Above all, logical exclusiveness is a precarious precondition of all classification for subsequent statistical analysis. Therefore, when the Lass-

well dictionary was constructed, if an entry could be classified under more than one category it was classified in the category which seems most appropriate, most of the time, for most texts. As regards intensity, although it is true that not all category entries will have the same pertinence to a category, a dichotomous rather than a weighted classification scheme was chosen nonetheless.

The category scheme of the current Harvard dictionary was constructed on a somewhat different strategy. It has a set of elementary or "first-order" categories in which entries are assigned on a mutually exclusive basis. These basic categories are then combined into higher-order categories.

A second major methodological question focuses on multiple classification. A strategy for investigating this problem follows naturally from the second-order factor-analysis model described just above. As depicted in Fig. 4, each word sense loads on only one category. Although the procedure for dictionary construction and the measurement model posit independence of categories, some categories are, however, conceptually close. For example, consider the "wealth" subcategories of the Lasswell dictionary: "wealth-participants", "wealth-transactions" and "wealth-other". The first category contains the names of those persons or positions involved in the creation, maintenance and transfer of wealth, such as "banker". The "transaction" category contains references to exchanges of wealth, such as "buying", "selling" and "borrowing". The "wealth-other" category contains wealth-related words not classified in the other two categories. Perhaps references to "banker" indicate a concern with both "participants" and "transactions". After all, bankers do execute transactions.

Given LISREL estimates for the measurement model in Fig. 4, it is possible to use certain diagnostic information produced by the LISREL program to determine whether a better-fitting model might result from the double or multiple classification of some word senses. The latest version of LISREL (Jöreskog and Sörbom, 1981) provides two important diagnostics. First, inspection of the normalized residual covariances may yield indications of specification errors in the model. Second, it is possible to examine a matrix consisting of the ratio between the squared first- and second-order derivatives of the fitting function for the estimated and fixed parameters. Large entries in this matrix suggest that the estimated value is not the true or correct parameter value. In addition, large derivatives are sometimes observed for parameters that were initially fixed to zero, but which are in fact nonzero.

For each word sense in the measurement model developed above, the factor loadings are constrained to zero for all categories except one. A large first derivative of a loading fixed at zero would suggest that word sense should be doubly or multiply categorized; that is, that the loading on that

category is not zero. Hence, large derivatives for the factor loadings suggest that a double or multiple classification scheme is more appropriate than the single classification approach used thus far in the Harvard and Lasswell dictionaries. Should the data indicate the need for multiple or double classification, then within the limits imposed by identification problems, the consequences of double or multiple classification can be explored in conjunction with the three problems discussed in the remainder of the paper.

A Priori versus Inferred Categories

Compared with hand-coding, one advantage of computer-based content analysis is that one set of texts can be classified by different dictionaries. However, this leads to multiple descriptions of the same textual reality. Consequently, there arose an important debate over whose dictionary should be used. Some (e.g., Stone et al., 1966; Dunphy et al., 1974; Namenwirth and Weber, 1974) held that the category scheme should be justified theoretically, and therefore the investigator's categories should be used. For example, the earliest Harvard psycho–social dictionaries were based in part on Parsonian and Freudian concepts (Stone et al., 1966), whereas the Lasswell Value Dictionary (Laswell and Namenwirth, 1968; Namenwirth and Weber, 1974) was based on Lasswell and Kaplan's (1950) conceptual scheme for political analysis [17].

Others (e.g., Iker, 1969; Cleveland et al., 1974; Krippendorff, 1980, (p. 126)) argued that a priori category schemes impose the reality of the investigator on the text. The better course of action, they argued, is to use the categories of those who produced the text. These categories are inferred from covariation among high-frequency words using factor analysis or similar techniques. As a result, different category schemes were inferred from different sets of texts, which then required a theory of categories in order to explain variation in category schemes (Namenwirth and Weber, 1974).

This dispute stems from both difficult methodological problems and from conceptual confusion. Let the term "category" be reserved for groups of words which have *similar* meanings and/or connotations (Stone et al., 1966; Dunphy et al., 1974). For example, the words "banker", "money" and "mortgage" might be classified in a "wealth" or "economic" category. Now let the term "theme" refer to clusters of words with *different* meanings or connotations that taken together refer to some theme or issue. For instance, the sentence, "New York bankers invest money in many industries both at home and abroad" in part reflects concern with economic issues or themes. This disagreement over categories is largely a dispute between those who define categories as words with different meanings that covary (inferred

categories), and those who define categories as words with similar meanings that covary.

Moreover, each approach entails a different measurement model. Specifically, first-order exploratory factor analysis is the statistical model that corresponds to inferring themes from word covariation. In Jöreskog's (1974) notation, the model for first-order exploratory factor analysis is given by

$$\Sigma = \Lambda\Phi\Lambda' + \Theta$$

where Σ is the covariance matrix among the observed variables, Λ is the matrix of factor loadings, Φ is an identity matrix if the factors are orthogonal, or a matrix of correlations among the factors if the solution is oblique, and Θ is a matrix whose diagonal elements are error variances and whose off-diagonal elements are zero. When used to derive so-called inferred content categories, the variables are words and the cases are documents or some other units of text such as paragraphs or themes. As is well known, without imposing constraints on this model there is no unique solution, although the factor loadings may be substantively more interesting after certain rotations than after others. As discussed extensively above, the measurement model for single-classification a priori dictionaries corresponds to a restricted second-order confirmatory factor analysis.

I am unaware of any attempt to analyze the same texts using both measurement models. Therefore, it is uncertain whether these different approaches yield similar or different substantive findings. To investigate the consequences of using very different measurement models for the substantive results, both models should be applied to various sets of documents and the results compared. In addition, it would be noteworthy to determine whether themes inferred from word covariation are more or less strongly related to noncontent variables, such as type of newspaper (mass/elite) or economic fluctuations. This same strategy of inquiry can be applied to a variety of text data sets and the generality of the findings assessed.

Units of Aggregation

The choice of "document" as the logical unit of analysis is only one of several possibilities. For example, sentences, paragraphs or themes might be used. There is some evidence (Saris-Gallhofer et al., 1978) indicating that the reliability of content categories varies according to the level of aggregation: comparing hand- and computer-coded content analyses of the same texts, it was found that sentences and documents had the highest reliabilities, while the reliability for paragraphs was slightly lower. In addition, the reliability at all levels of aggregation was substantially less than the reliabilities for specific words or phrases.

These findings call into question long-standing practices regarding the aggregation of words into larger units in both hand-coded and computer-aided content analysis. Future research should examine the consequences of different levels of aggregation. To evaluate systematically the consequences of aggregation, the second-order factor model developed above could be estimated using data aggregated at both the paragraph and document level. The null hypothesis would be that aggregation makes no difference; that is, that the coefficients do not differ according to level of aggregation. This is unlikely to be the case, if only because the variances of words, categories and themes are likely to be a function of aggregation.

One of the reasons why reliabilities are likely to vary with the level of aggregation stems from the fact that if a word (sense) is used, it is unlikely to be used again immediately [18]. For a given document, the longer the length of text considered, the more likely it is that word usage will fall into a stable pattern. Perhaps there is a threshold number of words below which word usage is unstable and reliability is lower, and above which word usage is stable and reliability is higher. It is unknown whether this threshold, if it exists, is close to the length of the average paragraph. Thus, for each set of documents the investigator could simulate units of text with varying lengths by aggregating over every n sentences. The parameters of the second-order factor model could be estimated several times, each time based on a constructed unit of text (pseudo-paragraph?) aggregated over a different number of sentences. In this way, the relationship between reliability and units of aggregation could be assessed systematically over a wide range of text lengths.

The results of this analysis should indicate whether the reliability and internal validity of dictionaries based on a priori category schemes vary with the level of aggregation. If not, then future investigations can proceed using the level of aggregation most relevant to the substantive problem at hand. In addition, greater confidence could be placed in previous research irrespective of the level of aggregation employed. If reliability and internal validity do vary by level of aggregation, then investigators must be more cautious in selecting units of analysis. Moreover, if, as Saris-Gallhofer et al. (1978) found, greater reliability is associated with smaller units, then it will be necessary to re-evaluate past results based on documents as the unit of analysis.

Impact of Different Dictionaries on Substantive Results

The choice of dictionary is predicated in part on theoretical considerations. For example, if it was wished to study extensively a particular

construct, such as McClelland's "need achievement" (Nach), then a dictionary might be constructed which scores only that variable (e.g., Stone et al., 1966, p. 191). General dictionaries such as the Harvard IV follow a different strategy. The category schemes of general-purpose dictionaries consist of many "common-sense" categories of meaning. These categories are chosen to reflect the wide range of human experience and understanding encoded in language.

Berelson's (1952) assertion that "content analysis stands or falls by its categories" is often quoted. This is certainly true in the operationalization of specific, relatively narrow concepts. However, it is my contention that for general-purpose dictionaries the content classification scheme has little or no effect on the substantive results. That is, if the same text is classified using different general dictionaries and analogous measurement models, then the same substantive conclusions will be reached.

There is already some empirical evidence on this score (Namenwirth and Bibbee, 1973, n. 12). In their analysis of newspaper editorials, Namenwirth and Bibbee classified the text using two different dictionaries and then factor-analyzed separately the two sets of scores. Comparing the results across the dictionaries, Namenwirth and Bibbee found that the factors had similar interpretations. Furthermore, irrespective of which dictionary was used, Namenwirth and Bibbee arrived at similar substantive conclusions [19].

Holding the general measurement model constant, future research should investigate the relationship between the dictionary used to classify a text and the substantive conclusions. Texts can be classified using multiple dictionaries and the results compared. If the substantive conclusions do not depend on the particular category scheme, then this would suggest that Berelson's assertion regarding the importance of categories is a limited truth. A practical benefit of this finding would be that where general dictionaries exist, researchers who have been reluctant to use one or another existing dictionary that did not operationalize their conceptual scheme might now be persuaded to do so. In addition, those who sought to create dictionaries in languages other than English might be persuaded to utilize existing category schemes to maintain cross-language comparability of results.

In the event that the results replicate only partly across dictionaries, additional research should ascertain the circumstances under which the results are similar or variant. Lastly, if the results indicate that Namenwirth and Bibbee's findings are unique, then this would provide empirical evidence that Berelson was right and that investigators must pay close attention to category schemes regardless of whether specific or general dictionaries are used.

Concluding Remarks

Over the last decade there has been some cultural-indicator research using both hand-coded and computer-aided content analysis. These studies have analyzed American (e.g., Namenwirth, 1969b, 1973), British (e.g., Weber, 1978, 1981, 1982a), Dutch (Gallhofer, 1978; Gallhofer and Saris, 1979a,b, 1980; Saris and Gallhofer, 1981), Swedish (Rosengren, 1981) and German (Mohler, 1978; Klingemann et al., 1982b) texts. Taken as a whole, these studies indicate that content analysis may be the preferred, indeed, in some cases, the only way to generate valid and reliable quantitative indicators spanning long periods of time. Substantively, the results of these studies complement and extend other findings regarding long-term social, political and economic change. For example, analyzing German college entrance examinations over the period 1917–1971, Mohler (1978) found that the values of German students remained stable through the 1920s and 1930s, but that there was a profound change in 1945 with the loss of the War and the Allied occupation.

These cultural-indicator studies have not received wide attention in part because content-analysis methodology is suspect. The research suggested above should put content analysis on a sound methodological base. Consequently, content-analytic results may more easily find their way into the mainstream of European and American social science.

Although the immediate focus is methodological, the research proposed above may eventually have its greatest impact on theory. Social scientists will be able to utilize computer-aided content analysis with greater confidence to address a wide variety of theoretical problems involving the relationships among cultural, social, economic and political change. Indeed, given the virtual revolution over the last ten years or so in the statistical analysis of time series (e.g., Box and Jenkins, 1976; McCleary and Hay, 1980; Hibbs, 1974, 1977; Bloomfield, 1976; Jenkins and Watts, 1968), this may be an especially good time to address empirically the relationship between cultural and other indicators.

Finally, rather than an endpoint, the reconceptualization and research proposed here are first steps towards the eventual reconciliation of word-count content analysis with approaches based on artificial intelligence. The next stage will be to explore both modes of content analysis as complementary strategies for resolving problems of interpretation and the representation of meaning.

Acknowledgments

The author thanks J.Z. Namenwirth, who suggested that the problems of content analysis might be investigated with modern statistical tools. Although he is not responsible for the way in which I have developed that idea, this paper is the outgrowth of a long dialogue with him and others, including P.J. Stone, P.P. Mohler, and H.-D. Klingemann. C.Z. Lawton and B. Norman provided editorial assistance.

Notes

1 For other definitions see Krippendorff (1980), Holsti (1969) and Gerbner et al. (1969).

2 Other computer-based approaches to content analysis include artificial-intelligence models of human cognition: see for example, Abelson (1963, 1973, 1975), Boden (1977), Shank and Colby (1973), Shank and Abelson (1977), Shank et al. (1980), Weizenbaum (1976) and Winograd (1972).

3 Proportions or percentages are often used to standardize for the length of document. Because the mean and variance of proportions are related, these should be transformed using the arcsin square-root transformation (e.g., Schuessler, 1970, pp. 411–416; Freeman and Tukey, 1950).

4 Kurzweil Computer Products (33 Cambridge Parkway, Cambridge, MA 02142) makes and markets an omnifont optical character reader. In pilot tests conducted at Kurzweil, this machine read the platforms of the Democratic and Republican parties for the period 1968–1972 quite easily and with few errors.

5 In addition to the reasons cited in the text, another reason why content analysis has not become a widely used methodology is that no institutional apparatus has evolved to support its development. The contrast with survey research is especially revealing, because at one time there were a number of "big names", such as Harold Lasswell, who were active in content-analytic research. For reasons that are not at all clear, the money went to support survey research. Janowitz's (1969) appraisal of content analysis is typical of the negative views current in the 1960s. Other reasons for the lack of interest in computer-aided content analysis are the difficulties in using existing computer software; the relatively high costs of computing during the 1960s and early 1970s; and the great expense of encoding text in machine-readable format. The last problem has been largely resolved by omnifont optical character readers and the ability to capture text from other electronic media, such as newswire services or newspaper editing and composition systems.

6 See the references in Note 2.

7 On the other hand, extensive work by Gottschalk (1979) and others demonstrates that content-analytic variables representing the emotional states of individuals are related to a wide range of physiological measures.

8 The Dutch content-analysis group at the Free University of Amsterdam, led by Irmtraud Gallhofer and Willem Saris (see for example, Gallhofer and Saris, 1980, 1979a,b), have made extensive use of contemporary statistical methods to analyze data generated by hand-coded content analysis (see especially Saris-Gallhofer et al., 1978; Saris and Gallhofer, 1981).

9 In addition to the LISREL program, these models can be estimated using MILS, an advanced form of LISREL written by Ronald L. Schoenberg at NIMH.

10 As discussed below, some content-analysis systems can distinguish among words with more than one sense. In this case, the unit of analysis is the word sense.
11 If it is wished not to make distributional and other strong assumptions, most of the models here can be estimated using Wold's (e.g., 1975, 1981) partial least-squares (PLS) procedures instead of LISREL. However, PLS estimation minimizes the errors in predicting the data. PLS parameter estimates are not maximum-likelihood, as in LISREL.
12 It is possible to overfit models to data. The current unwritten rule-of-thumb is that the best fit is obtained when the χ^2 value is about equal to the degrees of freedom.
13 Based on the information matrix, LISREL provides a convenient but not infallible check on the identifiability of the model.
14 High- and medium-frequency word senses are those appearing in a text at a rate of 10 or more per thousand words.
15 Weeks (1980) estimated a similar model by actually modifying the LISREL program. However, as Judd and Krosnick (1981) illustrate, this model can easily be estimated using the regular LISREL program.
16 This hypothesis argues that categories which are distinct conceptually will be unrelated empirically. With suitable constraints, models with some empirical correlation among the categories can be dealt with while maintaining the identification of the model.
17 I would like to call attention to what I immodestly refer to as "Weber's paradox": results using the Lasswell dictionary have not been interpreted or explained using Lasswell's theory, and results using the Harvard dictionary have not been interpreted or explained using Freudian or Parsonian theory.
18 Philip Stone pointed this out to me in a personal communication.
19 It should be noted that both the Harvard and Lasswell dictionaries emphasize institutional aspects of social life. In addition, Zvi Namenwirth played a large role in the creation of the Lasswell and early Harvard dictionaries. Therefore, some will not be surprised if his results do replicate across dictionaries.

References

Abelson, R.P. (1963). "Computer simulation of 'hot' cognition", in S.S. Tomkings and S. Messick, eds., *The Computer Simulation of Personality: Frontier of Psychological Research.* New York: Wiley.

Abelson, R.P. (1973). "The structure of belief systems", in R.C. Shank and K.M. Colby, eds., *Computer Models of Thought and Language.* San Francisco: W.H. Freeman.

Abelson, R.P. (1975). "Concepts for representing mundane reality in plans", in D.B. Bobrow and A. Collins, eds., *Representation and Understanding: Studies in Cognitive Science.* New York: Academic Press.

Bauer, R.A. (1966). *Social Indicators.* Cambridge, MA: MIT Press.

Bentler, P.M. and Bonett, D.G. (1980). "Significance tests and goodness of fit in the analysis of covariance structures", *Psychological Bulletin* 88: 588–606.

Berelson, B. (1952). *Content Analysis in Communications Research.* New York: Free Press.

Bloomfield, P. (1976). *Fourier Analysis of Time Series: An Introduction.* New York: Wiley.

Boden, M. (1977). *Artificial Intelligence and Natural Man.* New York: Basic Books.

Boot, M. (1974). "PASP: some views on automated syntactical parsing of large language corpuses", *Review of Applied Linguistics* (Institute Voor Toegepaste Linguistiek) 23: 23–38.

Boot, M. (1977a). "Key words in natural languages", *Annals of System Research* 6: 111–121.
Boot, M. (1977b). "An experimental design for automatic syntactic encoding of natural language texts", *ALLC Bulletin* 5: 237–241.
Boot, M. (1977c). "Linguistic data structure, reducing encoding by hand and programming language", in A. Jones and R.F. Churchhouse, eds., *The Computer in Literary and Linguistic Studies.* Cardiff: University of Wales Press.
Boot, M. (1978a). "Homographie: ein Beitrag zur automatischen Wortlassenzuweisung in der Computerlinguistik", unpublished dissertation, Utrecht.
Boot, M. (1978b). "Ambiguity and automated content analysis", *Methoden en Data Nieuwsbrief* 3: 117–137.
Box, G.E.P. and Jenkins, G.M. (1976). *Time Series Analysis: Forecasting and Control.* 2nd edn. San Francisco: Holden–Day.
Carley, M. (1981). *Social Measurement and Social Indicators: Issues of Policy and Theory.* London: Allen and Unwin.
Cleveland, C., McTavish, D. and Pirro, E. (1974). "Quester-contextual content analysis methodology", paper presented at the 1974 Pisa Conference on Content Analysis.
De Weese, L.C., III (1977). "Computer content analysis of 'day-old' newspapers: a feasibility study", *Public Opinion Quarterly* 41: 91–94.
Dunphy, D.C., Bullard, C.G. and Crossing, E.E.M. (1974). "Validation of the General Inquirer Harvard IV Dictionary", paper presented at the 1974 Pisa Conference on Content Analysis.
Freeman, M.F. and Tukey, J.W. (1950). "Transformations related to the angular and the square root", *Annals of Mathematical Statistics* 21: 607–611.
Gallhofer, I.N. (1978). "Coder's reliability in the study of decision making concepts, replications in time and across topics", *Methoden en Data Nieuwsbrief* 3: 58–74.
Gallhofer, I.N. and Saris, W. (1979a). "The decision of the Dutch Council of Ministers and the Military Commander-in-Chief relating to the reduction of armed forces in autumn 1916", *Acta Politica* 14: 95–105.
Gallhofer, I.N. and Saris, W. (1979b). "An analysis of the argumentation of decision makers using decision trees", *Quality and Quantity* 13: 411–430.
Gallhofer, I.N. and Saris, W. (1980). "Decision theory applied to foreign policy decisions", paper presented to International Society of Political Psychology, Boston, 1980.
Gerbner, G., Holsti, O.R., Krippendorff, K., Paisley, W. and Stone, P.J., eds. (1969). *The Analysis of Communication Content.* New York: Wiley.
Gottschalk, L.A. (1979). *The Content Analysis of Verbal Behavior: Further Studies.* New York: SP Medical and Scientific Books.
Heise, D.R. and Bohrnstedt, G.W. (1970). "Validity, invalidity, and reliability", in E.F. Borgatta and G.W. Borhnstedt, eds., *Sociological Methodology 1970.* San Francisco: Jossey–Bass.
Hibbs, D.A., Jr. (1974). "Problems of statistical estimation and causal inference in time series regression models", in H.L. Costner, ed., *Sociological Methodology 1973–1974.* San Francisco: Jossey–Bass.
Hibbs, D.A., Jr. (1977). "On analyzing the effects of policy interventions: Box–Tiao versus structural equations models", in H.L. Costner, ed., *Sociological Methodology 1977.* San Francisco: Jossey–Bass.
Hosti, O.R. (1969). *Content Analysis for the Social Sciences and Humanities.* Reading, MA: Addison–Wesley.
Iker, H.P. (1969). "A computer systems approach toward the recognition and analysis of content", in G. Gerbner et al., eds., *The Analysis of Communicaton Content.* New York: Wiley.

Janowitz, M. (1969). "Content analysis and the study of the 'symbolic environment'", in A.A. Rogow, ed., *Politics, Personality, and Social Science in the Twentieth Century: Essays in Honor of Harold D. Lasswell.* Chicago: University of Chicago Press.

Jenkins, G.M. and Watts, D.G. (1968). *Spectral Analysis and its Applications.* San Francisco: Holden–Day.

Jöreskog, K.G. (1974). "Analyzing psychological data by structural analysis of covariance matrices", in D.H. Krantz, R.C. Atkinson, R.D. Luce and P. Suppes, eds., *Contemporary Developments in Mathematical Psychology. Vol. 2.* San Francisco: W.H. Freeman.

Jöreskog, K.G. and Sörbom, D. (1979). *Advances in Factor Analysis and Structural Equation Models.* Cambridge: Abt Books.

Jöreskog, K.G. and Sörbom, D. (1980). *LISREL IV (Analysis of Linear Structural Relationships by the Method of Maximum Likelihood) Users' Guide.* Chicago: National Educational Resources, Inc.

Jöreskog, K.G. and Sörbom, D. (1981). *LISREL V (Analysis of Linear Structural Relationships by Maximum Likelihood and Least Squares Methods) Users' Guide.* Chicago: National Educational Resources, Inc.

Judd, C.M. and Krosnick, J.A. (1981). "Attitude centrality, organization and measurement", unpublished paper, Harvard University.

Kelly, E.F. and Stone, P.J. (1975). *Computer Recognition of English Word Senses.* Amsterdam: North-Holland.

Klingemann, H.D., Mohler, P.P. and Weber, R.P. (1982a). "Cultural indicators based on content analysis", *Quality and Quantity* 16: 1–18.

Klingemann, H.D., Mohler, P.P. and Weber, R.P. (1982b). "Das Reichtumsthema in den Thronreden des Kaisers und die ökonomische Entwicklung in Deutschland, 1871–1914", in H.-D. Klingemann, ed., *Computerunterstutzte Inhaltsanalyse in der empirischen Sozialforschung.* Kronberg: Athenäum.

Krippendorff, K. (1980). *Content Analysis: An Introduction to its Methodology.* Beverley Hills: Sage.

Land, K.C. and Spillerman, S. (1975). *Social Indicator Models.* New York: Russell Sage.

Lasswell, H.D. and Kaplan, A. (1950). *Power and Society: A Framework for Political Inquiry.* New Haven: Yale University Press.

Lasswell, H.D. and Namenwirth, J.Z. (1968). *The Lasswell Value Dictionary.* Vols. *1–3.* New Haven: Yale University (mimeo).

Lord, F.M. and Novick, M.R. (1968). *Statistical Theories of Mental Test Scores.* Reading, MA: Addison–Wesley.

McCleary, R. and Hay, R.A., Jr. (1980). *Applied Time Series Analysis for the Social Sciences.* Beverley Hills: Sage.

Markoff, J., Shapiro, G. and Weitman, S. (1974). "Toward the integration of content analysis and general methodology", in D.R. Heise, ed., *Sociological Methodology 1975.* San Francisco: Jossey–Bass.

Mohler, P.P. (1978). *Abitur 1917–1971: Reflecktionen des Verhältnisses zwischen Individuum und kollektiver Macht in Abituraufsätzen.* Frankfurt: Peter Lang.

Namenwirth, J.Z. (1969a). "Marks of distinction: a content analysis of British mass and prestige newspaper editorials", *American Journal of Sociology* 74: 343–360.

Namenwirth, J.Z. (1969b). "Some long and short term trends in one American political value", *Computer Studies* 3: 126–133 (reprinted in G. Gerbner et al., eds., *The Analysis of Communication Content.* New York: Wiley).

Namenwirth, J.Z. (1970). "Prestige newspapers and the assessment of elite opinions", *Journalism Quarterly* 47: 318–323.

Namenwirth, J.Z. (1973). "The wheels of time and the interdependence of value change", *Journal of Interdisciplinary History* 3: 649–683.

Namenwirth, J.Z. and Bibbee, R. (1973). "Speech codes in the press", *Journal of Communication* 25: 50–63.

Namenwirth, J.Z. and Bibbee, R. (1976). "Change within or of the system: an example from the history of American values", *Quality and Quantity* 10: 145–164.

Namenwirth, J.Z. and Lasswell, H.D. (1970). *The Changing Language of American Values: A Computer Study of Selected Party Platforms.* Beverley Hills: Sage.

Namenwirth, J.Z. and Weber, R.P. (1974). "The Lasswell Value Dictionary", paper presented at the 1974 Pisa Conference on Content Analysis.

Namenwirth, J.Z. and Weber, R.P. (1980). "Directed and contingent value changes in American and British political documents", *Methoden en Data Nieuwsbrief* 5: 3–44.

Ogilvie, D.M. (1966). "Procedures for improving the interpretation of tag scores: the case of windle", in P. Stone et al., eds., *The General Inquirer: A Computer Approach to Content Analysis.* Cambridge, MA: MIT Press.

Sheldon, E.B. and Moore, W.E. (1968). *Indicators of Social Change.* New York: Russell Sage.

Stone, P.J., Dunphy, D.C., Smith, M.S. and Ogilvie, D.M. (1966). *The General Inquirer: A Computer Approach to Content Analysis.* Cambridge, MA: MIT Press.

Tucker, L.R. and Lewis, C. (1973). "A reliability coefficient for maximum likelihood factor analysis", *Psychometrika* 38: 1–10.

Weber, R.P. (1978). "The dynamics of value change, transformations and cycles: British Speeches from the Throne, 1689–1972", unpublished Ph.D. dissertation, University of Connecticut.

Weber, R.P. (1981). "Society and economy in the Western world system", *Social Forces* 59: 1130–1148.

Weber, R.P. (1982a). "The long-term problem-solving dynamics of social systems", *European Journal of Political Research* 00: 00–00.

Weber, R.P. (1982b). "Content analytic cultural indicators", in G. Melischek, K.E. Rosengren and J. Stappers, eds., *Cultural Indicators.* Vienna: Austrian Academy of Sciences.

Weeks, D.G. (1980). "A second-order longitudinal model of ability structure", *Multivariate Behavioral Research* 15: 353–365.

Weizenbaum, J. (1976). *Computer Power and Human Reason.* San Francisco: W.H. Freeman.

Wilcox, L.D. (1972). *Social Indicators and Societal Monitoring: An Annotated Bibliography.* Amsterdam: Elsevier.

Winograd, T. (1972). *Understanding Natural Language.* New York: Academic Press.

Wold, H. (1975). "Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach", in J. Gani, ed., *Perspectives in Probability and Statistics, Papers in Honour of M.S. Bartlett.* London: Academic Press.

Wold, H. (1981). "Model construction and evaluation when theoretical knowledge is scarce: on the theory and application of partial least squares", in J. Kmenta and J. Ramsey, eds., *Evaluation of Econometric Models.* New York: Academic Press.

Rosengren, K.E. (1981). *Advances in Content Analysis.* Beverley Hills: Sage.

Saris, W.E. and Gallofer, I.N. (1981). "A coding instrument for empirical research of political decision making", unpublished manuscript, Free University of Amsterdam.

Saris-Gallhofer, I.N., Saris, W.E. and Morton, E.L. (1978). "A validation study of Holsti's content analysis procedure", *Quality and Quantity* 12: 131–145.

Schuessler, K. (1970). *Analyzing Social Data: A Statistical Orientation.* Boston: Houghton–Mifflin.

Shank, R.C. and Abelson, R.P. (1977). *Scripts, Plans, Goals and Understanding.* Hillsdale, NJ: Erlbaum Associates.

Shank, R.C. and Colby, K. (1973). *Computer Models of Thought and Language.* San Francisco: W.H. Freeman.

Shank, R.C., Lebowitz, M. and Birnbaum, L. (1980). "An integrated understander", *American Journal of Computational Linguistics* 6: 13–30.