A CRITICAL EXAMINATION OF THE ANALYSIS OF DICHOTOMOUS DATA*

WILLIAM H. BATCHELDER AND LOUIS NARENS†

University of California, Irvine

This paper takes a critical look at theory-free, statistical methodologies for processing and interpreting data taken from respondents answering a set of dichotomous (yes-no) questions. The basic issue concerns to what extent theoretical conclusions based on such analyses are *invariant* under a class of "informationally equivalent" question transformations. First the notion of Boolean equivalence of two question sets is discussed. Then Lazarsfeld's latent structure analysis is considered in detail. It is discovered that the best fitting latent model depends on which one of the many informationally equivalent question sets is used. This fact raises a number of methodological problems and pitfalls with latent structure analysis. Related problems with other methodologies are briefly discussed.

1. Introduction. Much of the methodology in the social sciences proposes techniques for drawing conclusions from data represented in the form of answers to dichotomous (yes-no) questions (or attributes) about some domain of inquiry. Among the many examples are latent structure analysis (e.g., [5]), multivariate uncertainty analysis (e.g., [2]), multidimensional scaling of respondents, taxonomy construction, and the classification of respondents by linear discriminant analysis ([4]). Such methodologies postulate a theory-free set of mathematical models, and the application of a particular methodology to particular data involves selecting the "best fitting" member of the set of models for that methodology. While a methodology may be applied solely as a data reduction tool in the spirit of curve fitting, more frequently, efforts are made to interpret the selected model theoretically in substantive terms. Related examples of this strategy in the social sciences (for data other than dichotomous) are well known in applications of analysis of variance, factor analysis, linear regression, and multidimensional scaling.[1]

*Received April, 1976; revised August, 1976.

†The authors would like to thank Albert Ahumada and Jerry Kaiwi for helpful discussions during the development of this paper.

[1] Many of the points made in this paper have direct consequences for taxonomic work in the biological sciences, in particular the debate concerning those who prefer phylic to phenic attributes in taxonomy construction (c.f. [7]). However, for a number of reasons, it seemed preferable to confine our analysis to methods in common use in the social sciences.

Philosophy of Science, 44 (1977) pp. 113-135.

In the dichotomous case, a particular research project generally proceeds in four separate stages: (1) questionnaire development, (2) data collection and preliminary analysis into a respondents-by-questions matrix, (3) the application of some standard methodology to such data, and (4) the substantive interpretation of the resulting model. While there is some body of literature on what constitutes a "good questionnaire," we have not been able to find any literature on the relationship of the questionnaire to the conclusions drawn from the data by the methodology employed. Most researchers suspect that one's knowledge about a domain will be heavily dependent on which questions are asked, but the extent to which the substantive conclusions of a data analysis routine will depend on the questions asked is not known. Basically this paper is concerned with the central question of whether or not *the best fitting model (and therefore the substantive conclusions) of a statistical methodology for dichotomous data is invariant under a class of question transformations*. Precise definitions of "question transformations" will be developed in the next section.

There is an old saw in mathematics that most things of mathematical or theoretical interest are invariant under important classes of transformations. For example, topological properties of a rubber sheet are invariant under various distortion operations such as stretching, and the areas and volumes of geometrical objects are invariant under the transformation of rotating coordinate system, etc. Invariants have also played a major role in the development of science, e.g., the conserved quantities of mass, energy, and momentum are invariant under radically different description frameworks. Even social science has its invariants, e.g., the correlation coefficient as well as the t and F statistics are invariant under linear transformations in the dependent variables. In fact one way that statistical texts can be organized is around the scale type of the dependent variables which in turn by the work of Stevens ([8]) is related to measurement scale transformations and invariants. Multidimensional scaling into Euclidean space has the feature that interobject distances are invariant under coordinate axis rotation, etc.

The emphasis of this paper is the question of the invariance of the theoretical structuring of a domain of inquiry under transformations on the questions asked about it. A corollary of this perspective is that when invariance is absent the case must be made for why the questions or attributes used were the "correct ones" for the methodology. The paper first presents a technical discussion of dichotomous questions and introduces the concept of Boolean transformations. Then, after preliminary notation is developed, latent structure analysis

is considered in the context of general Boolean transformations on the question set. Following a brief summary of related results for other methodologies, a number of methodological issues are raised.

2. Boolean Transformations. Dichotomous questions are questions that can be answered *yes* or *no*. There are two types of dichotomous questions that will be considered: *objective questions* and *subjective questions*. Objective questions are formulated and answered by the researcher and are questions about data on the members of a population; subjective questions are questions that are formulated by the researcher and answered by members of the population. If the researcher asks a subject "Do you have a cold today?" the question is considered subjective; if the researcher asks (of himself) about a subject "Did he answer 'yes' when asked if he had a cold today?" the question is considered objective.

Sometimes answers to some questions uniquely determine in some sense answers to other questions. To analyse this phenomena it is convenient to introduce a calculus of questions called *the propositional calculus*. Since for our purposes a "yes" answer to a question Q is equivalent to the assignment of the truth value *true* to the proposition expressed by Q, we will often call questions propositions. Furthermore, the symbol 1 will be used to denote a *yes* response or assignment of *true* to Q and 0 will denote a *no* response or an assignment of *false* to Q. Also in some contexts, questions are called attributes, and the assignment of 1 for subject j for attribute P will mean that subject j has attribute P.

Let $\mathcal{Q} = \{Q_1, Q_2, \ldots, Q_M\}$ be a set of M propositions. Then by a Boolean function P on $\mathcal{Q}$, we mean a logical formula using the propositions in $\mathcal{Q}$ and the standard logical connectives: disjunction ("or," $\vee$), conjunction ("and," $\wedge$), negation ("not," $\sim$), bicondition ("if and only if," $\leftrightarrow$), etc. It is a well-known theorem of propositional logic that all possible logical connectives can be defined in terms of the connectives $\wedge$ and $\sim$. By this definition there are infinitely many Boolean functions on $\mathcal{Q}$. However, if one identifies two Boolean functions P and Q as being logically equivalent if and only if $P \leftrightarrow Q$ is a tautology, then there are exactly 2^{2^M} different types of Boolean functions on $\mathcal{Q}$ such that a Boolean function of one type is not logically equivalent to a Boolean function of a different type.

While a researcher may collect data from each population member only on $\mathcal{Q}$, it is useful for our analysis to imagine that data exist on all 2^{2^M} Boolean function types generated by $\mathcal{Q}$. In the case of

objective[2] questions, it seems reasonable to expect that the usual methods of truth assignments and logic apply since we are supposing that the researcher is a good researcher and is therefore logical. However, in the case of subjective questions, the respondent may well be illogical and his response to a Boolean function on $\mathcal{Q}$ need not be uniquely determined by his response to $\mathcal{Q}$. Nevertheless, a logical respondent is certainly a possibility, and it is our aim to study the impact on standard data analysis routines of logical respondents. If the conclusions of a data analysis routine are very dependent on which set of Boolean functions the respondent is asked, then in some cases it could be argued that the data analysis routine demands illogical respondents to preserve its conclusions.

Let $\mathcal{Q} = \{Q_1, \ldots, Q_M\}$ be a set of M propositions and Γ be the set of Boolean functions on $\mathcal{Q}$. If P,R are elements of Γ, then P is said to be *Boolean equivalent* to R (in symbols, P eq R) if and only if $P \leftrightarrow R$ is a tautology. Let $\mathcal{P}$ be a set of M Boolean functions on $\mathcal{Q}$. Then $\mathcal{P}$ is said to be *Boolean equivalent* to $\mathcal{Q}$ (in symbols, $\mathcal{P}$ eq $\mathcal{Q}$) if and only if for each Q in $\mathcal{Q}$ there exists a Boolean function P on $\mathcal{P}$ such that P eq Q.

Example 2.1. Let $\mathcal{Q} = \{Q_1, Q_2\}$ and $\mathcal{P} = \{P_1, P_2\}$, where $P_1 = Q_1 \leftrightarrow Q_2$ and $P_2 = Q_2$. Then P_1 eq $Q_1 \leftrightarrow Q_2$, P_2 eq Q_2, Q_1 eq $P_1 \leftrightarrow P_2$, and Q_2 eq P_2, i.e., $\mathcal{P}$ is Boolean equivalent to $\mathcal{Q}$. Note the truth assignment correspondence given by Equation 2.1.

Q_1	Q_2	P_1	P_2
1	1	1	1
1	0	0	0
0	1	0	1
0	0	1	0

(2.1)

Note also that the truth assignments given to $\mathcal{P}$ are a permutation η of the truth assignments given to $\mathcal{Q}$, where $\eta(1,1) = (1,1)$, $\eta(1,0) = (0,0)$, $\eta(0,1) = (0,1)$, and $\eta(0,0) = (1,0)$. That truth assignments on $\mathcal{P}$ are permutations of those on $\mathcal{Q}$ will be shown in the following to be a characteristic property of Boolean equivalence.

Let $\mathcal{Q} = \{Q_1, \ldots, Q_M\}$ be a set of M propositions and $\mathcal{P} = \{P_1, \ldots, P_M\}$ be a set of M Boolean functions on $\mathcal{Q}$. Then $\mathcal{P}$ is said to be *informationally equivalent* to $\mathcal{Q}$ if and only if the truth assignments on $\mathcal{P}$ are permutations of the truth assignments to $\mathcal{Q}$:

[2]It would be possible to become more technical at this point and introduce the idea of a dichotomous attribute being tied to a measurement. For example, $Q_1 \leftrightarrow Q_2$ might not be obtained by first measuring Q_1 and Q_2 and then performing logic. Such issues are interesting, but their elaboration here would only obscure our main point.

in other words, $\mathscr{P}$ is informationally equivalent to $\mathscr{Q}$ if and only if there exists a 1-1 function η on the set of truth assignments χ (viewed in the usual truth table format as ordered M-vectors of 1's and 0's) on $\mathscr{Q}$ and functions $\eta_1, \ldots, \eta_M$ from χ into $\{0,1\}$ such that for each $\underset{\sim}{x}$ in χ, (*i*) $\eta_i(\underset{\sim}{x}) = 1$ if and only if P_i is assigned the value 1 by $\underset{\sim}{x}$, and (*ii*) $\eta(\underset{\sim}{x}) = (\eta_1(\underset{\sim}{x}), \ldots, \eta_M(\underset{\sim}{x}))$. Thus if $\mathscr{P}$ is informationally equivalent to $\mathscr{Q}$, then each truth assignment to elements of $\mathscr{Q}$ uniquely determines a truth assignment to elements of $\mathscr{P}$ and vice versa, i.e., $\mathscr{P}$ and $\mathscr{Q}$ are informationally indistinguishable.

Theorem 2.1. Let $\mathscr{Q} = \{Q_1, \ldots, Q_M\}$ be a set of M propositions and $\mathscr{P} = \{P_1, \ldots, P_M\}$ be a set of M Boolean functions on $\mathscr{Q}$. Then $\mathscr{Q}$ and $\mathscr{P}$ are Boolean equivalent if and only if they are informationally equivalent.

Proof. Let χ be the set of truth assignments on $\mathscr{Q}$. For each $\underset{\sim}{x} \in \chi$ and each $i \leq M$, let $\eta_i(\underset{\sim}{x}) = 1$ if and only if P_i is true under the assignment $\underset{\sim}{x}$, and let $\eta(\underset{\sim}{x}) = (\eta_1(\underset{\sim}{x}), \ldots, \eta_M(\underset{\sim}{x}))$.

Suppose $\mathscr{P}$ is informationally equivalent to $\mathscr{Q}$, i.e., suppose η is 1-1. For $i \leq M$, let

$$\mathscr{Y}_i = \{\underset{\sim}{x} | \underset{\sim}{x} \in \chi \text{ and } \eta_i^{-1}(\underset{\sim}{x}) = 1\}.$$

Now for each $\underset{\sim}{x} \in \mathscr{Y}_i$, write *the conjunctive normal form* $\mathscr{P}(\underset{\sim}{x})$ *of* $\mathscr{P}$ *corresponding to* $\underset{\sim}{x}$, e.g., if $\underset{\sim}{x} = (1,0,0)$ write

$$\mathscr{P}(\underset{\sim}{x}) = P_1 \wedge \sim P_2 \wedge \sim P_3.$$

Then let

$$R_i = \bigvee_{\underset{\sim}{x} \in \mathscr{Y}_i} \mathscr{P}(\underset{\sim}{x}), \tag{2.2}$$

where Eq. 2.2 represents the sentential formula obtained by "or-ing" all the $\mathscr{P}(\underset{\sim}{x})$ for $\underset{\sim}{x} \in \mathscr{Y}_i$. It is readily established that R_i eq Q_i. To see this note that if Q_i is true, then the set of possible truth assignments on $\mathscr{P}$ is given by $\mathscr{Y}_i$; consequently, $\mathscr{P}(\underset{\sim}{x})$ is true for some $\underset{\sim}{x} \in \mathscr{Y}_i$ and therefore R_i is true. On the other hand, if R_i is true, then $\mathscr{P}(\underset{\sim}{x})$ is true for some $\underset{\sim}{x} \in \mathscr{Y}_i$ from which it follows that $\eta_i^{-1}(\underset{\sim}{x}) = 1$ and thus Q_i is true. Thus $\mathscr{P}$ eq $\mathscr{Q}$.

Suppose $\mathscr{P}$ is Boolean equivalent to $\mathscr{Q}$. Suppose that η is not 1-1. A contradiction will be shown. Let $\underset{\sim}{x}, \underset{\sim}{y}$ be elements of χ such that $\underset{\sim}{x} \neq \underset{\sim}{y}$ and $\eta(\underset{\sim}{x}) = \eta(\underset{\sim}{y})$. Since $\underset{\sim}{x} \neq \underset{\sim}{y}$, let Q be an element of $\mathscr{Q}$ such that $\underset{\sim}{x}$ assigns true to Q and $\underset{\sim}{y}$ assigns false to Q. Since $\mathscr{P}$ eq $\mathscr{Q}$, let R be a Boolean function on $\mathscr{P}$ such that R eq Q. Since the elements

of $\mathscr{P}$ are Boolean functions on $\mathscr{Q}$, R may also be considered as a Boolean function on $\mathscr{Q}$. Since R eq Q and R is a Boolean function on Q, $\underset{\sim}{x}$ assigns true to R and $\underset{\sim}{y}$ assigns false to R. However, R as a Boolean function on $\mathscr{Q}$ is assigned true by $\underset{\sim}{x}$ if and only if R as a Boolean function on $\mathscr{P}$ is assigned true by $\eta(\underset{\sim}{x})$ if and only if (since $\eta(\underset{\sim}{x}) = \eta(\underset{\sim}{y})$) R is assigned true by $\eta(\underset{\sim}{y})$ if and only if R (as a Boolean function on $\mathscr{Q}$) is assigned true by $\underset{\sim}{y}$. Since R eq Q, $\underset{\sim}{x}$ assigns true to Q if and only if $\underset{\sim}{y}$ assigns true to $\mathscr{Q}$, and this is a contradiction. □

By theorem 2.1, every $\mathscr{P}$ which is Boolean equivalent to $\mathscr{Q}$ is equivalent to one of $(2^M)!$ sets of M Boolean functions on $\mathscr{Q}$ which correspond to the possible permutations of χ. The members of this finite set of canonical, Boolean equivalent representations constitute the set of informationally equivalent transformations of $\mathscr{Q}$.

3. Data Processing Routines. *Preliminary notation.* Suppose a researcher collects responses on M questions, $\mathscr{Q} = \{Q_1, \ldots, Q_M\}$, from each of N respondents (either objectively or subjectively). It is possible to represent the data in an $N \times M$ data matrix $\underset{\sim}{D}$ whose ij^{th} term, $\underset{\sim}{D}_{ij}$, is given by

$$\underset{\sim}{D}_{ij} = \begin{cases} 1 \text{ if respondent } i \text{ records a ''yes'' to question } j \\ 0 \text{ otherwise.} \end{cases}$$

Further, it is useful to denote the i^{th} row vector of $\underset{\sim}{D}$ by $\underset{\sim}{D}_{i.}$ and the j^{th} column vector of $\underset{\sim}{D}$ by $\underset{\sim}{D}_{.j}$.

Most of the methods of data analysis that will be considered in this paper proceed by first transforming $\underset{\sim}{D}$ into a M-dimensional binary contingency cube $\underset{\sim}{S}$. The cells in the cube correspond to the 2^M truth value assignments on $\mathscr{Q}$—hereafter known as the *signatures* of $\mathscr{Q}$—and the cell entries are the number out of the N respondents whose answer pattern corresponds to to that cell's signature. The next example illustrates these remarks.

Example 3.1. Suppose that there are 6 respondents to the 2 questions Q_1 and Q_2 yielding the following data matrix:

$$\underset{\sim}{D} = \begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \end{array} \begin{array}{c} \begin{array}{cc} Q_1 & Q_2 \end{array} \\ \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix} \end{array} \tag{3.1}$$

Then the 2-dimensional contingency cube corresponding to $\underset{\sim}{D}$ would be

$$\underset{\sim}{S} = Q_1 \quad \begin{array}{c|cc} & Q_2 = 1 & Q_2 = 0 \\ \hline Q_1 = 1 & 2 & 2 \\ Q_1 = 0 & 1 & 1 \end{array}. \tag{3.2}$$

In general, transformations of a question set $\mathcal{Q} = \{Q_1, \ldots, Q_M\}$ to an informationally equivalent set $\mathcal{P} = \{P_1, \ldots, P_M\}$ not only alter $\underset{\sim}{D}$ but they alter anything based on $\underset{\sim}{D}$ such as $\underset{\sim}{S}$. To illustrate, assume that the $\underset{\sim}{D}$ and $\underset{\sim}{S}$ of Equations 3.1 and 3.2 and that η is the transform of $\mathcal{Q} = \{Q_1, Q_2\}$ as developed in Example 2.1. The derived $\underset{\sim}{D}(\eta)$ and $\underset{\sim}{S}(\eta)$ are as follows:

$$\underset{\sim}{D}(\eta) = \begin{array}{c|cc} & P_1 & P_2 \\ \hline S_1 & 1 & 1 \\ S_2 & 0 & 0 \\ S_3 & 1 & 0 \\ S_4 & 0 & 1 \\ S_5 & 1 & 1 \\ S_6 & 0 & 0 \end{array} \tag{3.3}$$

and

$$\underset{\sim}{S}(\eta) = P_1 \quad \begin{array}{c|cc} & P_2 = 1 & P_2 = 0 \\ \hline P_1 = 1 & 2 & 1 \\ P_1 = 0 & 1 & 2 \end{array}. \tag{3.4}$$

Notice that the effects on $\underset{\sim}{S}$ of the transformation η are to juggle the cell frequencies inside of $\underset{\sim}{S}$. It is always the case that a Boolean equivalent transformation will juggle the frequencies in $\underset{\sim}{S}$; and further, each juggling of the cell frequencies in $\underset{\sim}{S}$ has a corresponding informationally equivalent transformation of the question space that gave rise to $\underset{\sim}{S}$. There are always $(2^M)!$ ways to juggle the cell frequencies in a contingency cube with 2^M cells, and these are in 1-1 correspondence with the class $\mathcal{N}$ of informationally equivalent transformations on $\mathcal{Q}$. Since each $\eta \in \mathcal{N}$ maps the signature space $\mathcal{X}$ onto itself, the cell frequency in cell $\underset{\sim}{x}$ is merely mapped by η into the cell corresponding

to $\eta(\underset{\sim}{x})$. Of course if several cells have identical frequencies, some of the transformations will not change $\underset{\sim}{S}$.

Before turning to the data analysis routines, another observation about the relationship between the logic and the binary contingency cube is in order. Basically for each question set $\mathcal{Q} = \{Q_1, \ldots, Q_M\}$ some of the $\eta \in \mathcal{N}$ merely negate certain questions—i.e., change their truth sense—and others merely reorder the questions. Such transformations preserve the original question set except for reorderings and truth sense changes. Such transformations are called *signed permutations* of $\mathcal{Q}$. Since with M questions there are 2^M possible truth sense changes (including the identity transform $\eta(\underset{\sim}{x}) = \underset{\sim}{x}$ for all $\underset{\sim}{x} \in \mathcal{X}$) and $M!$ reorderings of the questions, there are $2^M \cdot M!$ members of $\mathcal{N}$ that are signed permutations of $\mathcal{Q}$.

Signed permutations of $\mathcal{Q}$ preserve a certain structure on the marginal frequencies of "yes" and "no" to each component question. That is, for any set of M questions one can construct a $2 \times M$ matrix $\underset{\sim}{M}$ of the frequencies of respondents (out of N) who answer "yes" or "no" to each question

$$\underset{\sim}{M} = \begin{array}{c|cccc} & Q_1 & Q_2 & \cdots & Q_M \\ \text{yes} & f_{11} & f_{12} & f_{1j} & f_{1M} \\ \text{no} & f_{01} & f_{02} & f_{0j} & f_{0M} \\ & N & N & N & N \end{array},$$

where f_{1j} is the frequency of respondents responding "yes" to question j and f_{0j} is the frequency responding "no" to the question j for $1 \leq j \leq M$. If η is a signed permutation of $\mathcal{Q}$, then $\underset{\sim}{M}(\eta)$ will consist of a permutation of the columns of $\underset{\sim}{M}$ with some number between 0 and M interchanges of the f_{1j}'s and f_{0j}'s. Moreover if η is not a signed permutation of $\mathcal{Q}$, then in general, the resulting $\underset{\sim}{M}(\eta)$ may not conform to these restrictions.

It is easy to define a binary relation $R_{\mathcal{Q}}$ on $\mathcal{N}$ such that for all $\eta_1, \eta_2 \in \mathcal{N}$

$$\eta_1 \, R_{\mathcal{Q}} \, \eta_2$$

if and only if there is a $\theta \in \mathcal{N}$ such that θ is a signed permutation and $\eta_2(\mathcal{Q}) = \theta[\eta_1(\mathcal{Q})]$. That is, $\eta_1(\mathcal{Q})$ and $\eta_2(\mathcal{Q})$ are identical question sets up to possible reorderings of the questions and changes in the truth sense of certain questions. Further, it is easy to show that $R_{\mathcal{Q}}$ is an equivalence relation on $\mathcal{N}$. Therefore we can partition the $(2^M)!$ members of $\mathcal{N}$ into

$$(2^M)!/(2^M \cdot M!) = (2^M - 1)!/M!$$

equivalence classes each with $2^M \cdot M!$ members which are signed permutations of each other. Example 3.2 illustrates some of these points.

Example 3.2. Suppose two questions $\mathcal{Q} = \{Q_1, Q_2\}$ are asked of N respondents giving rise to the 2×2 contingency table (cube)

$$\underset{\sim}{S}_1 = Q_1$$

	Q_2: 1	Q_2: 0
Q_1: 1	a	b
Q_1: 0	c	d

where, of course, $a + b + c + d = N$. For $M = 2$ there are $(2^2)! = 4! = 24$ members of $\mathcal{N}$ and by the preceding remarks there should be 3 equivalence classes each with 8 members. Such is clearly the case and prototypical members of the other two classes are given by

$$\underset{\sim}{S}_2 =$$

a	c
d	b

and

$$\underset{\sim}{S}_3 =$$

a	b
d	c

4. Latent Structure Analysis. Latent structure analysis ([5]) is one of several statistical methods in Social Science designed to resolve structure from data. The data consist of the $N \times M$ matrix $\underset{\sim}{D}$ described in the previous section. Each structural model has the nature of a set of L "latent classes," c_ℓ, to each of which is associated an M-dimensional row vector, $\underset{\sim}{p}_\ell$, whose j^{th} term is the probability of a "yes" response to question j for any respondent falling into class c_ℓ, where $1 \leq \ell \leq L$ and $1 \leq j \leq M$. The analysis proceeds under the assumption that the population of actual or potential respondents to questions about a concept domain can be partitioned into groups each of which presumably represents some intrinsic or latent viewpoint about the concept domain under investigation (cf. [5], Chapter 1). All members of a particular group (latent class) of respondents are thought to look at the concept domain in the same way in the sense

that they all have the same probability of a "yes" response to *any particular question* that might be asked about the concept domain. In addition, responses to a set of questions for members of a latent class are postulated to be independent (the axiom of local independence). The output of the analysis for a particular set of M questions is a "best approximating" number of classes, L, the associated set of L response probability vectors, and the *a posteriori* probabilities that each of the N respondents falls into the various classes (sometimes called recruitment probabilities).

The method of extracting latent classes from $\underset{\sim}{D}$ proceeds by forming the M-dimensional, binary, contingency cube $\underset{\sim}{S}$ and attempting to decompose $\underset{\sim}{S}$ into a sum of L contingency cubes,

$$\underset{\sim}{S} = \underset{\sim}{H}_1 + \ldots + \underset{\sim}{H}_L, \tag{4.1}$$

where each of the $\underset{\sim}{H}$'s manifests approximate statistical independence. Each of the component cubes corresponds to one of the latent classes, c_ℓ, and the response vector for that class, p_ℓ, is given by the M-dimensional row vector of marginal probabilities of "yes" responses to the M questions for the corresponding cube, H_ℓ.

For purposes of our analysis, it is useful to formalize the notion of statistical independence in an M-dimensional, binary, contingency cube $\underset{\sim}{H}$ with N observations. If $F(\underset{\sim}{x})$ is the observed frequency in the cell $\underset{\sim}{H}(\underset{\sim}{x})$, then $\underset{\sim}{H}$ is said to show statistical independence in case for all $\underset{\sim}{x} \in \chi$,

$$F(\underset{\sim}{x}) = N \cdot \Pi_{j=1}^{M} p_j^{x_j} (1 - p_j)^{(1-x_j)} \tag{4.2}$$

where p_j and x_j are the j^{th} members of the row vectors $\underset{\sim}{p}$ and $\underset{\sim}{x}$, respectively. Equation 4.2 requires that each cell frequency be computable from an appropriate product of marginal probabilities times the sample size.

The tasks of selecting L and selecting a model that gives an adequate approximation to Equation 4.1 give rise to complex and interesting problems in applied mathematics. For purposes of this section we shall assume that Equation 4.1 can be solved adequately. At its best a solution to Equation 4.1 gives rise to latent classes that truly reflect underlying pockets of opinion about some cognitive domain in the sense that any other "complete" set of questions about the same domain would give rise to essentially the same latent classes and class memberships despite different response vectors. At its worst, the analysis is merely a data processing routine for rerepresenting $\underset{\sim}{D}$ as an approximation in terms of simpler structures. The first of

these possibilities is important because researchers generally employ the analysis with the intention of interpreting the resulting structure substantively.

One way to assess the theoretical usefulness of latent structure analysis is to see if its output is left invariant under informationally equivalent transformations of $\mathcal{Q}$. Since any of the $(2^M)!$ "complete" question formats could have been used in a questionnaire study, one would hope that the latent classes revealed by the analysis would be the same for each question format. In the event that the latent class structure that emerges from the analysis depends on the question set, a number of methodological issues, previously unraised, emerge.

It is easily seen that the results of latent structure analysis are invariant under signed permutation transformations. To see this one notes that the right side of Equation 4.2 is invariant under question transformations that reorder the M questions or change particular x_i's from 0 to 1 or from 1 to 0. Thus if Equation 4.1 represents an adequate approximation to $\underset{\sim}{S}$ and η is a signed permutation transformation on $\mathcal{Q}$ then

$$\underset{\sim}{S}(\eta) = \sum_{i=1}^{L} \underset{\sim}{H}_i(\eta)$$

gives an identically adequate approximation to $\underset{\sim}{S}(\eta)$.

Unfortunately, when transformations stray outside the class of signed permutation transformations of $\mathcal{Q}$, the results of latent structure analysis can be radically different. To illustrate this, consider the two question case discussed in Example 3.1, Equations 3.1, 3.2, 3.3, and 3.4. It is readily verified that Equation 4.2 is satisfied for Equation 3.2 but not for Equation 3.4, *i.e.*, Equation 3.2 shows statistical independence but not Equation 3.4. Therefore, a latent structure analysis of $\underset{\sim}{D}$ will reveal one latent class; whereas, the analysis of $\underset{\sim}{D}(\eta)$ will not.[3]

A second and more substantive example of the problems with using latent structure analysis to discover "true" latent classes is the following:

Example 4.1. Suppose that a large data bank has among other things, information from 210 respondents to the following three questions:

[3]The fact that this example and the next involve only two questions ($M = 2$) should not disturb the reader. It is easy but undesirably messy to create examples of $M > 2$ in the spirit of the examples provided. The examples illustrate the fact that the output of a latent structure analysis is necessarily *not invariant* under equivalent question sets.

(1) Q_1, registered Republican, (2) Q_2, voted for Nixon in 1972, and (3) Q_3, believed Nixon innocent in 1974. Note that for simplicity we will assume that the respondent pool has been sifted so that "no" answers on Q_1, Q_2, and Q_3 mean registered Democrat, voted for McGovern in 1972, and believed Nixon guilty in 1974, respectively. Equation 4.3 gives the respondent frequencies corresponding to each labeled signature frequency (data hypothetical).

Label	Q_1	Q_2	Q_3	Frequency
A	1	1	1	60
B	1	1	0	35
C	1	0	1	5
D	1	0	0	40
E	0	1	1	10
F	0	1	0	15
G	0	0	1	5
H	0	0	0	40
				210

(4.3)

Now let us suppose that two astute political scientists—Professor X and Professor Y—decide to apply latent structure analysis to data drawn from the bank in an effort to uncover fundamental American political viewpoints prevalent in the early 1970's. Professor X has a theory which suggests that P_1, "party loyalty," and P_2, "identity with the party image," are critical questions. He decides to use the data bank to define these questions as follows:

$$P_1 = Q_1 \leftrightarrow Q_2$$

and

$$P_2 = Q_1 \leftrightarrow Q_3 .$$

Put less technically, "party loyalty" is defined as "voted for the party you registered for" and "identity with the party image" is approximated by an identical logic used on the assumption that the party image of the Republican party was that Nixon was innocent and that the opposite viewpoint was the stand of the Democratic party.

Let us assume that quite independently Professor Y decides to employ the same methodology but to two different questions. According to Professor Y's theory, critical questions would be R_1, "party loyalty" (same as before), and R_2, "satisfied with 1972 vote" (measured in 1974). He formally defines these by

$$R_1 = P_1$$

and

$$R_2 = Q_2 \leftrightarrow Q_3 .$$

Thus a "yes" is registered to R_2 for Nixon voters who thought Nixon was innocent in 1974 and McGovern voters who thought Nixon was guilty in 1974.

It is of interest that Professor X's and Professor Y's questions are informationally equivalent, *i.e.*, $\{P_1, P_2\}$ is equivalent to $\{R_1, R_2\}$ in the sense of section 2. This fact can be seen by noting the following implications:

$$R_2 = [P_1 \leftrightarrow P_2] = [(Q_1 \leftrightarrow Q_2) \leftrightarrow (Q_1 \leftrightarrow Q_3)] \tag{4.4}$$

and

$$P_2 = [R_1 \leftrightarrow R_2] = [(Q_1 \leftrightarrow Q_2) \leftrightarrow (Q_2 \leftrightarrow Q_3)] . \tag{4.5}$$

It is particularly interesting to note that from Professor X's viewpoint, Professor Y's questions seem unduly complicated since they involve the biconditional in Equation 4.4; however, from Professor Y's viewpoint, Professor X's questions suffer from exactly (!) the same flaw in Equation 4.5.

The actual latent structure analysis is facilitated by Equations 4.6 and 4.7 which convert data bank frequencies into signature frequencies for the new questions:

P_1	P_2	Data Bank	Frequencies
1	1	$A + H$	100
1	0	$B + G$	40
0	1	$C + F$	20
0	0	$D + E$	50
			210

(4.6)

and

R_1	R_2	Data Bank	Frequencies
1	1	$A + H$	100
1	0	$B + G$	40
0	1	$D + E$	50
0	0	$C + F$	20
			210 .

(4.7)

The reader should note that a comparison of Equation 4.6 and 4.7 directly reveals the equivalence of $\{P_1, P_2\}$ and $\{R_1, R_2\}$, and, in terms of the definition $\eta(1,1) = (1,1)$, $\eta(1,0) = (1,0)$, $\eta(0,1) = (0,0)$, and $\eta(0,0) = (0,1)$.

The appropriate Chi-squared test of independence ([3], Chapter 17) for the data in Equation 4.6 yields

		P_2 = 1	P_2 = 0
P_1	1	100	40
	0	20	50

with $\chi^2 = 35$, which is a poor fit of Equation 4.2. On the other hand, for the data in Equation 4.7 one gets

		R_2 = 1	R_2 = 0
R_1	1	100	40
	0	50	20

with $\chi^2 = 0$, which is a perfect fit of Equation 4.2. Thus Professor X would probably find more than one latent class in his analysis; whereas, Professor Y, working with logically equivalent questions, would find only one latent class.

The first point to be made about the noninvariance result illustrated by the previous examples is that it does not hinge on *requiring* the the respondents to answer questions in one set in a manner logically consistent with their answers to the other set. In both examples, it was the analyst and not the respondents who provided the data for analysis from original respondent data. Nevertheless, the example seems very damaging to the efficacy of latent structure analysis because the respondents might have been asked either set of questions. From the examples we can conclude that the subjects would have had to respond *illogically* to both sets to reveal identical latent class structures. It seems to us that a technique for revealing latent structure in a cognitive domain which requires that respondents make responses to questions that are logically inconsistent is foundationally untenable. Of course an alternative method of analysis which *required* that subjects be logically consistent would also be untenable.

A natural response to the preceding argument would be that the data for a latent structure analysis represent a sample from an underlying probability distribution over the signature space. In such cases, "illogical" responding would seem to be a necessary consequence of small sample theory. It should be observed, however, that our discussion could just as easily be pitched at the population level

of analysis. Viewed in this way, corresponding to each informationally equivalent transformation on $\mathcal{Q}$ would be a new probability distribution over the signature space. Our results then can be viewed as showing that if the output of a latent structure analysis is to be invariant under informationally equivalent transformations, then the corresponding transformations on the probability distributions on the signature space is not mirrorable by the logic. For example, if P_1 eq $[Q_1 \leftrightarrow Q_2]$, then it is not consistent that

$$Pr(P_1 = 1) = Pr(Q_1 = 1 \ \& \ Q_2 = 1) + Pr(Q_1 = 0 \ \& \ Q_2 = 0).$$

That such natural transformations of probability distributions are incompatible with latent structure analysis is viewed by us as a serious fundamental problem on a par with the practical problem illustrated by the two examples.

Another implication of the examples is that latent structure analysis is rather limited in situations involving unobtrusive data gathering procedures. In such cases the experimenter (perhaps a botanist) gathers a body of yes-no data on each of N subjects (perhaps fossilized plants) without awareness on the part of the subjects. An experimenter could pose and answer M questions from the data for each subject, from a $\underset{\sim}{D}$ matrix, and perform a latent structure analysis. However, the examples show us that the result of the analysis would depend heavily on the questions selected by the experimenter.

A natural response to the preceding argument might be that while the results of the analysis might vary with question format, each result might be interesting in its own right. It is true that each informationally equivalent question set gives rise to its own set of latent classes; however, initial in their determination is the axiom of local independence alluded to earlier. Lazarsfeld and Henry ([5], p. 22) write:

> The defining characteristic of the latent structure models is the axiom of local independence, stated here for the case of discrete classes. . . .
>
> AXIOM OF LOCAL INDEPENDENCE. Within a latent class, α, responses to different items are independent. The within class probability of any pattern of response to any set of items is the product of the appropriate marginal probabilities. . . .
>
> Notice that our definition applies to any set of items, no matter how large. . . .

We of course agree that the methodology produces a best fitting model for each informationally equivalent question set; however, because local independence is so dependent on the format used, we

doubt its usefulness as a primitive assumption in theoretical interpretations of data.

If the analysis is incompatible among various informationally equivalent transformations, it might be argued that it is intended to work only for sets of M "elementary" ("simple," "atomic") questions. Such a restriction would seem to resolve our problem by simply basing the analysis on the set $\mathscr{Q}$. However, from a semantic standpoint, the simple questions may be as complicated logically as the sententially more involved questions in $\mathscr{P}$. For example, suppose a subject's semantics of disease included the notion that "to have a cold" (Q) means "to have a runny nose (Q_1) and a postnasal drip (Q_2) but not a sinus headache (Q_3)." Then we have $Q = Q_1 \wedge Q_2 \wedge \sim Q_3$. In other words, any cognitive domain worthy of interest will have a semantic structure that reflects opinion. The originally unknown logical character of this semantic structure would preclude an *a priori* assessment of which questions actually represent "elementary" propositions about the cognitive domain. While some might be tempted to observe that correct question selection might constitute an art, it should be pointed out that no such observations are evident in the considerable literature on the subject.

Even more damaging to the argument that latent structure analysis is suited to deal with only "simple" questions is the obvious fact that simplicity is judged with respect to a base. Thus in Example 4.1, the $\{R_1, R_2\}$ set is just as complex logically when viewed in terms of $\{P_1, P_2\}$ as a base as is $\{P_1, P_2\}$ viewed in terms of $\{R_1, R_2\}$. In fact, Equations 4.4 and 4.5 show that the R's can be written logically in terms of the P's in exactly the same way that the P's can be written in terms of the R's. Put loosely, each equivalent question format regards the alternative, equivalent formats as occupying various points on a complexity scale with itself as the least complex. Nothing in the logical form of the transformations bears on complexity of a question format unless we know *on other considerations* which is the most basic format. A reasonable "other consideration" would be that the questions used were based on some theory of the cognitive domain under investigation. However, the existence of such a theory would likely render unnecessary the use of latent structure analysis in the first place. In any event, no such connections are established in the literature on the methodology.

5. Other Related Methodologies. A number of methodologies, other than latent structure analysis, for dealing with dichotomous data are flawed by similar lack of invariance results. A brief summary of these is presented in this section.

Several methodologies start with $\underset{\sim}{D}$ and produce an $N \times N$ matrix of similarity or resemblance coefficients based on a simple count of respondent agreements. The similarity matrix, $\underset{\sim}{T}$, based on $\underset{\sim}{D}$ is an $N \times N$ matrix whose ij^{th} term is given by

$$\begin{aligned} t_{ij} &= \Sigma_{k=1}^{M} D_{ik} \cdot D_{jk} + (1 - D_{ik})(1 - D_{jk}) \\ &= \underset{\sim}{D}_{i.} \cdot \underset{\sim}{D}_{j.} + [(\underset{\sim}{I} - \underset{\sim}{D}_{i.}) \cdot (\underset{\sim}{I} - \underset{\sim}{D}_{j.})] \\ &= \underset{\sim}{D}_{i.} \oplus \underset{\sim}{D}_{j.}, \end{aligned} \tag{5.1}$$

where "$\cdot$" is the vector product and "$\underset{\sim}{I}$" is the M-dimensional row vector of 1's. Thus t_{ij} is a simple count of the number of agreements between respondent i and j and is thought to reflect their similarity.

Among the methodologies based on $\underset{\sim}{T}$ are the various nonmetric scaling programs[4] (see [6] for a taxonomy) and the taxonomy construction methods based on phenic characters prevalent in biology (see [7]). The aim of such methodologies is to rerepresent the data in $\underset{\sim}{D}$ in a format designed to make similar respondents "close together." In the case of nonmetric scaling, respondents are represented as vectors and closeness is defined in terms of the Euclidean or Minkowskian metrics, and in the case of taxonomy the close respondents fall into higher cells of a taxonomic hierarchy.

It is clear that informationally equivalent transformations on $\mathscr{Q}$ will change the pattern of similarities between respondents. To see this, consider the data in Example 3.1, Equation 3.1. The 6×6 matrix of similarities obtained from Equation 3.1 by using Equation 5.1 is

$$\underset{\sim}{T} = \begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \end{array} \begin{array}{c} \begin{array}{cccccc} S_1 & S_2 & S_3 & S_4 & S_5 & S_6 \end{array} \\ \begin{bmatrix} 2 & 1 & 0 & 1 & 2 & 1 \\ 1 & 2 & 1 & 0 & 1 & 2 \\ 0 & 1 & 2 & 1 & 0 & 1 \\ 1 & 0 & 1 & 2 & 1 & 0 \\ 2 & 1 & 0 & 1 & 2 & 1 \\ 1 & 2 & 1 & 0 & 1 & 2 \end{bmatrix} \end{array}. \tag{5.2}$$

When $\mathscr{Q}$ is transformed by η to yield the $\underset{\sim}{D}$ given by Equation 3.3 the similarity matrix becomes

[4] Nonmetric scaling programs require as input a matrix of similarities between pairs of entities. The method of getting $\underset{\sim}{T}$ described in Section 5 is only one of many utilized. Our concerns do not extend in a natural way to other methods of getting $\underset{\sim}{T}$.

$$\underset{\sim}{T}(\eta) = \begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \end{array} \begin{array}{c} \begin{array}{cccccc} S_1 & S_2 & S_3 & S_4 & S_5 & S_6 \end{array} \\ \begin{bmatrix} 2 & 0 & 1 & 1 & 2 & 0 \\ 0 & 2 & 1 & 1 & 0 & 2 \\ 1 & 1 & 2 & 0 & 1 & 1 \\ 1 & 1 & 0 & 2 & 1 & 1 \\ 2 & 0 & 1 & 1 & 2 & 0 \\ 0 & 2 & 1 & 1 & 0 & 2 \end{bmatrix} \end{array}. \tag{5.3}$$

Examination of Equation 5.2 indicates that for the $\mathcal{Q}$ set $1 = t_{12} > t_{13} = 0$; however, for the transformed data $0 = t_{12}(\eta) < t_{13}(\eta) = 1$. There are other reversals of the distances. However, it should be noted that $t_{ij} = 2$ if and only if $t_{ij}(\eta) = 2$. This clearly follows because η is a 1-1 function on χ so that if two subjects have identical answer patterns to $\mathcal{Q}$ then they will have identical patterns in $\eta(\mathcal{Q})$, for all $\eta \in \mathcal{N}$. It is also easy to show that the subject similarity matrix is invariant under signed permutation of $\mathcal{Q}$.

The preceding example clearly shows that the results of nonmetric scaling of subject similarities are not invariant under informationally equivalent transformations. A plausible idea for getting a matrix of similarities that are reflective of different question sets might be to compute the average t_{ij} over all $\eta \in \mathcal{N}$,

$$\bar{t}_{ij} = E_{\mathcal{N}}(t_{ij}) = (1/2^M!)\,\Sigma_{\eta \in \mathcal{N}}\, t_{ij}(\eta).$$

The next theorem shows that this plan is frustrated by a surprising result:

Theorem 5.1 Let $\underset{\sim}{D}$ be a $N \times M$ data matrix with questions $\mathcal{Q} = \{Q_1, \ldots, Q_M\}$ and let $\mathcal{N}$ be the set of Boolean equivalent transformations on $\mathcal{Q}$. Then for all $1 \leq i,j \leq N$

$$\bar{t}_{ij} = \begin{cases} M & \text{if } \underset{\sim}{D}_{i.} = \underset{\sim}{D}_{j.} \\ M(2^{M-1} - 1)/(2^M - 1) & \text{if } \underset{\sim}{D}_{i.} \neq \underset{\sim}{D}_{j.} \end{cases}$$

Proof: Clearly if $\underset{\sim}{D}_{i.} = \underset{\sim}{D}_{j.}$, $t_{ij}(\eta) = M$ for all $\eta \in \mathcal{N}$ since η is 1-1; hence $\bar{t}_{ij} = M$.

If $\underset{\sim}{D}_{i.} \neq \underset{\sim}{D}_{j.}$ then every signature pair $\underset{\sim}{x}$ and $\underset{\sim}{x}' \in \chi\,(\underset{\sim}{x} \neq \underset{\sim}{x}')$ is assumed by subjects i and j equally often as η ranges over $\mathcal{N}$. In fact there are $2^{M-2}!$ members of $\mathcal{N}$ that result in $\underset{\sim}{D}_i$ and $\underset{\sim}{D}_j$ taking any fixed pair of distinct signatures. Consequently,

$$\begin{aligned} \bar{t}_{ij} &= (1/2^M!)\,\Sigma_{\eta \in \mathcal{N}}\, t_{ij}(\eta) \\ &= \Sigma_{\underset{\sim}{x}, \underset{\sim}{x}' \in \chi}\,(\underset{\sim}{x} \oplus \underset{\sim}{x}')\, 2^{M-2}!/2^M! \end{aligned}$$

$$= \Sigma_{k=0}^{M-1} k\binom{M}{k}/(2^M - 1)$$

$$= M(2^{M-1} - 1)/(2^M - 1). \qquad \square$$

Theorem 5.1 implies that the average similarity between any pair of subjects who differ at all in their answers to $\mathcal{Q}$ is the same regardless of how many different answers they had to $\mathcal{Q}$.[5] Such a result would seem to place in jeopardy the practice of substantively interpretating similarity matrices based on subject similarity scores from an arbitrary question set. However, as with latent structure analysis, our results do not invalidate these methods if viewed merely as data reduction tools.

The methodological implication of Theorem 5.1 is similar to the results of the preceding section. Put simply, efforts to group respondents by their answer patterns to a set of questions are heavily dependent on which set of questions is used. In particular, naturally related question sets produce incompatible groupings. The considerable literature in the social sciences on these and related methodologies have ignored the syntactic issue altogether, preferring instead, to emphasize the complex applied mathematical issues that arise in such analyses.

Related observations are possible for other methodologies for analyzing dichotomous data. In particular, substantive conclusions based on the information theoretic approach[6] ([2]) and on linear threshold logic[7] ([1]; [4]) are severely restricted to the question format utilized by the methodology. While either methodology may be useful

[5] After we had arrived at the theorem, a related result by Watanabe ([9]) came to our attention. The theorem of Watanabe is called the "Theorem of the ugly duckling." Basically Watanabe shows that an ugly duckling and a swan are just as similar as two swans if one compares entities on the full Boolean lattice of 2^{2^M} predicates based on any starting set $\mathcal{Q}$ and M predicates. We had similar experiences to Watanabe in explaining this result to various scientists: Watanabe states:

> . . . It is curious that when I talked about the statement and proof of the theorem on different occasions since 1961, some people have manifested their surprise and delight, while others grumbled that they knew something like this must be true. But when I asked the latter group of people where they had read or written it, I could get no clear answer. . . . ([9], p. 376)

Our Theorem 5.1 is not the same as Watanabe's theorem, but it amounts to the same thing.

[6] While total uncertainty in a contingency S is invariant under informationally equivalent transformations on $\mathcal{Q}$, the contingent uncertainty as well as the maximum uncertainty depend heavily the question format used.

[7] The main result here is that for each classifactory partition of the signature space for $\mathcal{Q}$ there exists an informationally equivalent transformation of $\mathcal{Q}$ that permits a description of that partition by a linear threshold function (the dichotomous case of a linear discriminant function) on the derived question set. In other words, the validity of a set of questions can not be based on the existence or nonexistence of an efficient classification rule that is based on a linear threshold function of the question set.

as a data reduction and/or description method, theoretical conclusions based on such analyses are radically altered by transformations among informationally equivalent question sets. While each question format produces a best fitting model, the primitive assumptions of the method seem too particular to us to suggest that each might be useful in its own right.

6. Discussion. In this paper we have studied methodologies for dealing with dichotomous data in the context of transformations on the data space that preserve information of a logic-theoretic sense. In the case that the data space consists of all M-dimensional one-zero vectors corresponding to M dichotomous questions or attributes, the transformations (see section 2) amount to the class of permutations of the respondent signature frequencies over the set of 2^M possible signature patterns. Since these transformations may strike the reader as rather radical and of little interest, it is reasonable to renew and restate the case for their importance.

When a scientist approaches a domain of inquiry with the intention of collecting data from members of a population, he must select questions or attributes for study. There are always other questions or attributes that might have been selected, and, in particular, every set of M dichotomous questions gives rise to a number of alternative sets of M questions each of which is an equivalent syntactic representation of the information obtained from the members of the population. Lacking any reasonable theoretical basis of choosing among such question sets, it seems desirable that theoretical conclusions from data analysis routines be invariant under transformations among informationally equivalent question sets. A number of methodological analyses in current practice, in particular latent structure analysis and methodologies based on respondent similarities, do not yield results that are invariant under this class of transformations. While such methodologies may be convenient as data reduction tools, theoretical conclusions based on these methodologies must be held with serious suspicion until criteria are developed for selection and evaluation of sets of questions for data processing.

The preceding paragraph summarizes the position taken in this paper. A number of objections to aspects of the argument have been communicated to us from readers of preliminary drafts of the paper, some of which have been dealt with in earlier sections. In the remainder of this section several of these are considered in numbered paragraphs.

I. It might be objected that respondents should not be expected to reply to a Boolean function of questions in a manner that is consistent with their answers to the original questions. Thus it would appear

inappropriate to permute signature frequencies when changing question sets.

There are two replies to this objection. First, the analysis of dichotomous data often takes the form of a researcher both formulating dichotomous attributes and assessing members of a population himself. In this case one would expect the researcher to be bound to the dictates of Boolean logic. Second, we are not proposing a response model based on formal logic; however, if a respondent *did behave* according to Boolean logic, then a methodology should not yield differing conclusions depending on which set of informationally equivalent, Boolean functions is selected for analysis. Our point is that response patterns *must violate* Boolean logic if the conclusions from several methodologies are to be invariant under our class of question transformation, and thus we see reason to doubt such methodologies on foundational criteria. As remarked in Section 4, our concerns extend to the population as well as the sample level of analysis.

II. One might argue that the original set of questions is simpler than any other set in that its members are "elementary propositions," whereas, any other set of equivalent questions has as its members Boolean functions of the "elementary propositions." Thus by requiring that methodologies only be applied to these "elementary propositions," one removes the need for invariance.

This is a tricky point, and it is true that if "squatter's rights" are granted to the utilized set of questions, all other equivalent sets appear to be more complex. However, such simplicity is always judged relative to a base. What is simple from the point of view of one question set is complex from another's point of view. Example 4.2 nicely illustrates this point by showing that if either of two equivalent question sets is chosen as the base, the other appears *identically* complex. Put differently, the idea of propositions corresponding to atomic predicate symbols in logic and the idea of an elementary notion from a substantive view point are not the same. For example, "setting up the chess men correctly" is a basic notion for chess players; however, its description is a very complex Boolean function of certain propositional systems such as the chess rules or action sequences. While a serious analysis of these semantic considerations would require a richer logic than the propositional calculus, the direction of that analysis is clearly indicated by our examples.

III. An objection related to the previous one is that the original questions were formulated by the researcher carefully using his intuition for what constitutes basic notions in the domain of inquiry. While researchers may err in their ability to construct basic questions,

methodologies are intended for good question sets and therefore ought to be expected to yield conclusions invariant under equivalent formulations.

While this point appears to us to have some validity, there are two important points to be made. First, nothing in the formalisms of the methodologies we have analyzed states the conditions under which a set of questions for a domain of inquiry is appropriate. Perhaps new work could be directed toward formulating such criteria; however, there is no current justification for asserting that these methodologies are intended for use on "basic" questions only. In the area of medical diagnosis, for example, considerable latitude over popular lay symptomologies is exercised in developing diagnostic methodologies. Even if adequate criteria for a basic question set could be developed, they would depend on already acquired knowledge of the area of inquiry. Thus the use of many methodologies in, for example, anthropological analyses offers no escape from the well-known pitfalls of culture bias, *i.e.*, it appears to us that the type of knowledge required to determine if a particular set of questions is fundamental for studying the belief system of an exotic culture may turn out to be the type of knowledge that some anthropologists have hoped to achieve by employing the supposedly "culture free" data processing methodologies.

IV. Finally, it must be stated that many of the results in this paper are of a negative character, *i.e.*, they argue against a current practice without clearly pointing the way to a substitute.

We feel that the results in the paper are positive from a foundational standpoint, since they provide a criterion for the sorts of measurements which can be taken seriously in drawing conclusions from dichotomous data. Once the requirements on a measure are clearly stated, the task of developing more sophisticated tools of measurement is greatly facilitated. (The authors are currently completing a sequel to this paper that proposes new measurement methodologies, as well as providing useful criteria for the fundamentalness of question sets.)

REFERENCES

[1] Chow, C. K. "Statistical independence and threshold functions." *IEEE Transactions on Electronic Computers* EC-14 (1965): 66-68.

[2] Garner, W. R. *Uncertainty and Structure as Psychological Concepts.* New York: Wiley, 1962.

[3] Hays, W. L. *Statistics for Psychologists.* New York: Holt, Rinehart, and Winston, 1963.

[4] Krishman, T. "On Linear Combinations of Binary Item Scores." *Psychometrika* 38 (1973): 291-304.

[5] Lazarsfeld, P. F. and N. W. Henry. *Latent Structure Analyses.* Boston: Houghton Mifflin Co., 1968.

[6] Shepard, R. "A Taxonomy of some Principal Types of Data and of Multidimensional Methods for their Analysis." In Shepard, R. N., Romney, A. K. and Nerlove, S. B. (eds.). *Multidimensional Scaling*, Vol. I, New York: Seminar Press, 1972. Pages 21-47.
[7] Sokal, R. R. and Sneath, P. H. A. *Principles of Numerical Taxonomy*. San Francisco: W. H. Freeman, 1963.
[8] Stevens, S. S. "Measurement and Man." *Science* 127 (1958): 383-389.
[9] Watanabe, S. *Knowing and Guessing*. New York: John Wiley and Sons, 1969.