A. KIMBALL ROMNEY
University of California, Irvine
SUSAN C. WELLER
University of Pennsylvania
WILLIAM H. BATCHELDER
University of California, Irvine

Culture as Consensus: A Theory of Culture and Informant Accuracy

This paper presents and tests a formal mathematical model for the analysis of informant responses to systematic interview questions. We assume a situation in which the ethnographer does not know how much each informant knows about the cultural domain under consideration nor the answers to the questions. The model simultaneously provides an estimate of the cultural competence or knowledge of each informant and an estimate of the correct answer to each question asked of the informant. The model currently handles true-false, multiple-choice, and fill-in-the-blank type question formats. In familiar cultural domains the model produces good results from as few as four informants. The paper includes a table showing the number of informants needed to provide stated levels of confidence given the mean level of knowledge among the informants. Implications are discussed.

THE CONCEPT OF CULTURE has long been the central focus of study in anthropology. Writing in the early 1950s Kroeber (1952:139) observed that "The most significant accomplishment of anthropology in the first half of the twentieth century has been the extension and clarification of the concept of culture." More recently Goodenough (1964:36) has asserted that "the anthropologist's basic task, on which all the rest of his endeavor depends, is to describe specific cultures adequately. . . . Culture, being what people have to learn as distinct from their biological heritage, must consist of the end product of learning: knowledge, in a most general, if relative, sense of the term." In this paper we present a way of describing and measuring the amount and distribution of cultural knowledge among a group of informants in an objective way.

We are reminded of the need for an objective approach to culture and ethnography by the recent controversy generated by the publication of Derek Freeman's book, *Margaret Mead and Samoa* (1983). In the introduction to a special section of the *American Anthropologist,* Ivan Brady (1983:908) commented that "The book has been reviewed all over the world and has raised questions of authenticity and viability in ethnographic research. . . . One broad but undetermined topic of enduring value that emerges from these essays, it seems to me, is the problem of how anthropologists get to know what they know and write in the first place." The model provided in this paper is an attempt to make objective the criteria by which we might measure our confidence in inferring correct answers to cultural questions, i.e., to help answer the epistemological question of "How do we know when we know?"

A. KIMBALL ROMNEY is Professor, School of Social Sciences, University of California, Irvine, CA 92717. SUSAN C. WELLER is Assistant Professor, Department of Medicine, Clinical Epidemiology Unit, University of Pennsylvania, 2L NEB/S2, 420 Service Drive, Philadelphia, PA 19104. WILLIAM H. BATCHELDER is Professor, School of Social Sciences, University of California, Irvine, CA 92717.

Despite examples of differing views of ethnographers such as Mead and Freeman in Samoa and Redfield and Lewis in Tepoztlan, Mexico, anthropologists may tend to underestimate the probable effects of the ethnographer in selecting and shaping the data and in forming impressions contained in the final ethnographic report. This points to our need to find more objective ways to investigate culture.

The assumption in fieldwork has been that the investigator is a valid and reliable instrument and that the informant provides valid and reliable information. We suggest that informants' statements should be treated as probabilistic in character. When, for example, an informant states that the name of an object is "X," we should assume that there is some probability (that we can estimate) that the statement is correct. This probability may be close to 1 in the case of a very knowledgeable informant and close to 0 in the case of an uninformed informant. The more informants there are who agree (when questioned independently) on an answer the more likely it is to be the correct cultural response.

Informant interviews are a main source of data for anthropology. Evaluation and analysis of such data, including theory construction and testing, constitute a vital part of the research activity of the profession. Frequent disagreement among informants confronts the investigator with two major problems: first, how can the "cultural knowledge" of different informants be estimated, and, second, how can the "correct" answers to specific questions be inferred and with what degree of confidence? This paper supplies a formal approach that answers these two questions for a variety of cultural information domains that lend themselves to systematic question formats, e.g., true-false, multiple-choice, fill-in-the-blank.

The aspect of culture that our theory attempts to account for is the part that is stored in the minds of its members. Roberts (1964:438–439) has said that "It is possible to regard all culture as information and to view any single culture as an 'information economy' in which information is received or created, stored, retrieved, transmitted, utilized, and even lost. . . . In any culture information is stored in the minds of its members and, to a greater or lesser extent, in artifacts." In a similar vein, D'Andrade has developed the notion of culture as a shared and learned "information pool."

> It is not just physical objects which are products of culture. . . . Behavior environments, consisting of complex messages and signals, rights and duties, and roles and institutions, are a culturally constituted reality which is a product of our socially transmitted information pool. . . . In saying that an object—either a physical object like a desk, or a more abstract object like a talk or a theorem—is a product of culture, I mean that the cultural pool contains the information which defines what the object is, tells how to construct the object, and prescribes how the object is to be used. Without culture, we could not have or use such things. [1981:180]

The size of the cultural information pool virtually dictates that knowledge be distributed and shared. Roberts points out that there is a limit to what an individual or combination of individuals can learn and that "it is safe to assert that no tribal culture is sufficiently small in inventory to be stored in one brain" (1964:439). D'Andrade, in a closely reasoned discussion, places some loose bounds on the possible size of the information pool that an individual may control. "Upper limits can be obtained by considering time constraints; e.g., to learn ten million chunks would require that one learn more than a chunk a minute during every waking hour from birth to the age of twenty" (1981:180). The large size of the information pool is also related to the division of labor in society. "One of the characteristics of human society is that there is a major division of labor in who knows what" (D'Andrade 1981:180). Clearly we cannot study all of culture but rather we must have a strategy for sampling smaller, coherent segments of the total information pool constituting culture. We also need to make a provision for the possible unequal distribution of knowledge among "experts" or specialists and nonspecialists in a society.

One segment of culture that provides a reasonable focus derives from Kroeber's (1948) classic discussion of "systemic culture pattern." Systemic culture patterns are characterized as coherent subsystems of knowledge that tend to cohere and persist as a unit limited

primarily to one aspect of culture. A systemic culture pattern has sufficient internal organization that it may diffuse as a unit. As examples, Kroeber (1948) discusses plow agriculture and the alphabet. Roberts and his colleagues have studied such diverse systemic culture patterns as eight-ball pool (Roberts and Chick 1979), pilot error (Roberts, Golder, and Chick 1980), women's trapshooting (Roberts and Nuttrass 1980), and tennis (Roberts et al. 1981). They have demonstrated that the relevant behavior events for each of these domains are coded into what Roberts has called "high-concordance codes" that cultural participants understand and use with ease.

Each systemic culture pattern may be thought of as having an associated semantic domain that provides a way of classifying and talking about the elements in the culture pattern. A semantic domain may be defined as an organized set of words (or unitary lexemes), all on the same level of contrast, that jointly refer to a single conceptual sphere, e.g., a systemic culture pattern. The words in a semantic domain derive their meaning, in part, from their position in a mutually interdependent system reflecting the way in which a given language classifies the relevant conceptual sphere. This definition corresponds to Frake's discussion on levels of contrast (1961). Examples of semantic domains include kinship terms, linguistic "tags" for the behavior events in games like tennis (Roberts et al. 1981), manioc names in Aguaruna (Boster 1986), disease terms and characteristics (Weller 1983, 1984a, 1984b), Buang clan membership (Sankoff 1971), and color terminology.

A recent example of ethnographic data that could be analyzed by our theory was collected by Boster (1986) on Aguaruna manioc classification. He asked informants to identify growing manioc plants in a garden. "Data were collected by guiding informants through the gardens, stopping at each plant and asking, *waji mama aita,* 'What kind of manioc is this?' " Boster concluded that the more an informant agreed with others the more knowledge that informant had about manioc. Since he was able to assess differences among informants as to cultural knowledge, he was able to establish that women knew more than men and that women in the same kin and residential groups were more similar to each other in knowledge than nonrelated women.

In a test-retest analysis he found that "within informant agreement is strongly correlated with between informant agreement" (Boster 1986). Since the informants who agree with the group the most on the first test are those who agree most with themselves on the retest, Boster argues that agreement among informants indicates knowledge. The results, he says, "helped confirm the inference of cultural knowledge from consensus." We agree with Boster that knowledge can be inferred from consensus.

The aim of this paper is to derive and test a formal mathematical model for the analysis of informant consensus on questionnaire data that will stimultaneously provide the following information: first, an estimate of the cultural competence or knowledge of each informant, and second, an estimate of the correct answer to each question asked of the informants.

The plan for the remainder of the paper is as follows: first, after a brief, informal verbal description of the theory we will present the formal mathematical model for the analysis of true-false, multiple-choice, and fill-in-the-blank profile data. Derivations are kept as simple as possible. Further technical details on the model and related models can be found in Batchelder and Romney (1986). Applications of earlier informal versions of the theory can be found in Romney and Weller (1984), Weller (1984b), and Weller, Romney, and Orr (1986). Second, we apply the model to quasi-experimental data consisting of a general information test where answers are known a priori to illustrate how the model works. We also analyze a small subset of informants and discuss sample size requirements of the model. Third, we apply the model to some field data on disease classification collected in Guatemala. This illustrates the application of the model in a naturally occurring situation where the answers are not known a priori and the results may have important theoretical implications. Finally, we will discuss the implications of the model and relate it to some of the current research concerns in anthropology.

Description and Development of the Formal Model

The central idea in our theory is the use of the pattern of agreement or consensus among informants to make inferences about their differential competence in knowledge of the shared information pool constituting culture. We assume that the correspondence between the answers of any two informants is a function of the extent to which each is correlated with the truth (Nunnally 1978:chap. 6). Suppose, for example, that we had a "perfect set" of interview questions (cultural information test) concerning the game of tennis. Suppose further that we had two sets of informants: tennis players and non-tennis players. We would expect that the tennis players would agree more among themselves as to the answers to questions than would the non-tennis players. Players with complete knowledge about the game would answer questions correctly with identical answers or maximal consensus, while players with little knowledge of the game would not. An insight like this one allowed Boster to identify those informants with the most cultural knowledge in his study of manioc plants.

We are assuming that there exists a "high concordance code" of a socially shared information pool concerning tennis, that informants vary in the extent to which they know this culture, and that each informant answers independently of each other informant. Once we know how "competent" each informant is we can figure out the answers to the questions by weighting each informant's input and aggregating to the most likely answer. That is, we put more weight on the more knowledgeable informants than the less knowledgeable ones. The model we develop is simply an elaboration and formalization of these ideas and their implications.

Although the model and associated data analysis methods are new, we incorporate derivations and concepts from previous well-established theories. The major sources of concepts include the following: the overall structure as well as more general ideas are influenced by signal detection theory (Green and Swets 1966). Approaches used by psychometricians in test construction to study items were adapted to apply to informants rather than items (Lord and Novick 1968; Nunnally 1978). There are structural identities to latent structure analysis (Lazarsfeld and Henry 1968) although again our applications are to informants rather than questions. The relevance of the Condorcet jury trial problem is also important (e.g., Grofman, Feld, and Owen 1983). Techniques common in decision analysis like Bayesian estimation are common in many fields and can be found in any mathematical statistics book (e.g., Hogg and Craig 1978).

At this point we need to introduce some definitions and notation in order to present the formal model. We start with an informant by question Response Profile Data Matrix of the following form:

(1)

$$\mathbf{X} = \begin{array}{c|cccccccc} \text{Informant} & 1 & 2 & \cdot & \cdot & k & \cdot & \cdot & M \\ \hline 1 & X_{11} & X_{12} & \cdot & \cdot & X_{1k} & \cdot & \cdot & X_{1M} \\ 2 & X_{21} & X_{22} & \cdot & \cdot & X_{2k} & & & X_{2M} \\ \cdot & \cdot & & & & & & & \\ \cdot & \cdot & & & & & & & \\ i & X_{i1} & X_{i2} & \cdot & \cdot & X_{ik} & \cdot & \cdot & X_{iM} \\ \cdot & \cdot & & & & & & & \\ \cdot & \cdot & & & & & & & \\ N & X_{N1} & X_{N2} & \cdot & \cdot & X_{Nk} & \cdot & \cdot & X_{NM} \end{array} \quad \text{Question}$$

where X_{ik} is the ith informant's response to the kth question. There are N informants and M questions. The model assumes a questionnaire where each question has L possible response alternatives with only one "correct" answer. In a true-false questionnaire $L = 2$ and the response X_{ik} would be "true" or "false" (coded, perhaps, as 1's and 0's, respec-

tively). In a multiple-choice questionnaire, for example, there might be four alternatives so that $L = 4$ and the possible responses X_{ik} might be coded as "A," "B," "C," or "D." A fill-in-the-blank questionnaire can be thought of as a special case of the model with a very large value of L, and X_{ik} would be the actual, possibly edited, written response (possibly blank) of the ith informant to the kth question.

Our notational conventions are:

A. Response Profile Data. $\mathbf{X} = (X_{ik})_{N \times M}$ where X_{ik} is the subject i's response to question k coded as described above.

B. Answer Key. $\mathbf{Z} = (Z_k)_{1 \times M}$ where Z_k is the code for the correct answer to question k (initially unknown to us).

C. Performance Profile Data. $\mathbf{Y} = (Y_{ik})_{N \times M}$, where

$$Y_{ik} = \begin{cases} 1 \text{ if subject } i \text{ is correct on item } k \\ 0 \text{ if subject } i \text{ is wrong on item } k. \end{cases}$$

D. Response Bias.[1] g_{il} is a bias to respond with an alternative l when informant i does not know answer. The range of g_{il} is between 0 and 1. No bias is $1/L$, e.g., in true-false a bias of $1/2$ means that if the informant does not know the answer to the question that they will choose either alternative with equal probability. In the derivations below we assume no bias.

E. Cultural Competence. D_i is the probability that informant i knows (not guesses) the answer to a question, where $0 \leq D_i \leq 1$, and negative D_i's are not allowed by model. This is a theoretical parameter of the model and cannot be observed directly. We assume each informant has the same D_i for all questions.

Since the notation is crucial to understanding what follows we will review and expand on the above. The response profile data **X** is the original raw data from the interviews, and it simply refers to the whole profile data matrix in Eq. 1 consisting of N rows of informants and M columns of questions. The answer key (**Z**) can be estimated from the model but it is not known a priori, it consists of a single vector or line of data that contains the code for the correct answer. In anthropological work we usually do not know the correct answers a priori and the model provides a method for inferring the answer key from the response profile data. When the response data has been recoded as correct or incorrect based upon an answer key we call it performance profile data denoted by **Y.** In psychological test theory it is usually assumed that the investigator knows the answers to the questions while in anthropological fieldwork we do not normally know the answers a priori and hence the need for the model.

The assumptions of the formal model may be stated as follows:

Assumption 1. Common Truth. *There is a fixed answer key "applicable" to all informants, that is, each item* k *has a correct answer,* Z_k, k = 1,2,. . .,M. This simply states the assumption that the informants all come from a common culture, i.e., that whatever the cultural reality is, it is the same for all informants in the sample.

Assumption 2. Local Independence. *The informant-item response random variables satisfy conditional independence, that is,*

$$\Pr[(X_{ik})_{N \times M} \mid (Z_k)_{1 \times M}] = \prod_{i=1}^{N} \prod_{k=1}^{M} \Pr(X_{ik} \mid Z_k) \tag{2}$$

for all possible response profiles (X_{ik}) and the correct answer key (Z_k). This assumes that each informant's answers are given independently of each other informant. The correlations among informant's answer patterns are an artifact of the extent to which each is correlated with the "answer key," i.e., **Z.** To the extent that the data fit the model correlations among informants will be high if computed on the response profile data but close to 0 if computed on the performance profile data.

Assumption 3. Homogeneity of Items. *Each informant* i *has a fixed "cultural competence,"* D_i, *over all questions.* This is a strong assumption that says that questions are all of the same difficulty level. In some situations one might want to make a weaker assumption: namely, that the informants who do better on one subset of the questions will do better on another subset of questions. This generalization might be called the monotonicity assumption and is related to ensuring that the questions are drawn from a coherent domain. Thus, for example, if the tennis experts do better than the nonexperts on one part of the questions concerning tennis, they should do better on another part concerning tennis. The analysis, however, has proven to be very robust in practice under the more restricted homogeneity assumption.

We might note that these assumptions define the ground rules for the operation of our model. They also make it possible to make formal derivations in mathematical terms. It is important to stress that not all response profile data will conform to these assumptions. They require, for example, that all informants are positively correlated with each other (except for sampling variability). In effect this means that our theory assumes that if two people are members of the same culture that they are responding to the questions in terms of a common understanding, i.e., the culture is similar for both informants. By similar we do not mean they will respond identically since there will be misunderstanding of the questions, random guesses, etc. The model measures the shared knowledge of the culture. True negative correlations among informants would mean that they do not have common knowledge in the domain sampled by the questions. Thus when the empirical data show that any of the assumptions are violated then the model does not apply and we infer one of the following: (1) we are not dealing with a culturally defined domain, (2) the informants do not share common knowledge of the cultural domain, or that (3) something else has gone wrong. We will give an example of a violation of the assumptions later.

The formal derivations of the model will be presented in the following steps: first, we derive a method for estimating D_i, the cultural competence, of each informant from the data in the response profile matrix. This estimation is made on the basis of the pattern of shared knowledge (as indexed by proportion of matched responses among all pairs of informants), using the notion that the more consensus the more knowledge. Second, we show how to make inferences as to the correct answers together with statistical confidence levels based on an application of Bayes' Theorem in probability theory (for example, Mosteller, Rourke, and Thomas 1961:146). The model we present is a special case of a more general family of models (Batchelder and Romney 1986) and is referred to there as the High Threshold Model.

Derivations from the Model

We now turn to the task of deriving the cultural competence of the informants from the proportion of matches among them. The parameter D_i is informant i's cultural competence, namely, the probability that informant i "knows" the correct answer to any item ($0 \leqslant D_i \leqslant 1$). If the informant does not know the correct answer (with probability $[1\text{-}D_i]$), then they guess the answer with probability $1/L$ of a correct answer, where $(1\text{-}D_i)$ is the probability of not knowing the answer and L is the number of alternative answers to the question. For example, assume an informant's competence is .7 ($D_i = .7$) for a five-item multiple-choice questionnaire. In addition to expecting that the informant will get .7 of the questions correct we would also expect the informant to get some of the .3 $(1\text{-}D_i)$ questions correct by guessing. Namely, $1/L$ or 1/5 of the remaining .3 of the questions or $(1\text{-}D_i)/L$, i.e., $.3 \times 1/5 = .06$ would be guessed correctly. We add this to the .7 giving a total expected correct of .76. More generally the probability of any question k being answered correctly by any informant i is given by

$$(3) \qquad \Pr(Y_{ik} = 1) = D_i + (1\text{-}D_i)/L,$$

and the probability of answering incorrectly is given by

$$\Pr(Y_{ik} = 0) = (1\text{-}D_i)\ (L\text{-}1)/L.$$

Note that even if we knew the proportion of questions the informant got "correct" (which we do not usually know) we could not observe the effects of the theoretical parameter D_i directly because the proportion correct includes the proportion the informant got right by guessing. In case the correct answer key is known it is easy to simply count the number of correct responses and divide by M, the number of questions, to obtain the proportion of correct responses T_i for informant i. In order to obtain an estimate of D_i we use the empirically observed T_i in place of $\Pr(Y_{ik} = 1)$ in Eq. 3 and solve for D_i to obtain

$$\hat{D}_i = (LT_i\text{-}1)/(L\text{-}1), \tag{4}$$

where the hat over the D_i is the usual convention to indicate that it is an estimate of the underlying competency D_i. All Eq. 4 does is to adjust the proportion correct for guessing, and this is used routinely in aptitude testing by ETS and other agencies.

The anthropologist, unlike the test-theorist, does not know the correct answers in advance so that we cannot use Eq. 4 to estimate the D_i. Fortunately, and perhaps surprisingly, it is still possible to obtain estimates of the D_i's by examining the proportion of matches among all pairs of informants. The derivation of the procedure follows.

Assume two informants, i and j, whose probabilities for a correct response, from Eq. 3, are:

for informant i,

$$\Pr(\text{correct}) = D_i + (1\text{-}D_i)/L,$$

and for informant j,

$$\Pr(\text{correct}) = D_j + (1\text{-}D_j)/L.$$

Now we want to know the probability of i and j matching responses on any question k in terms of the competence, D_i and D_j, of each. The possible ways of matching are:

1. Both know the answer to the item with probability D_iD_j, that is, the probability that i knows the item times the probability that j knows the item.

2. One informant knows the answer to the item and other guesses the item correctly. This occurs in two ways: i knows the item and j guesses correctly with probability $D_i\ (1\text{-}D_i)/L$ and j knows the item and i guesses correctly with probability $D_j\ (1\text{-}D_i)/L$.

3. Neither knows the item but both guess the same response[2] to the item which occurs with probability

$$(1\text{-}D_i)\ (1\text{-}D_j) \sum_{l=1}^{L} (1/L)^2 = (1\text{-}D_j)\ (1\text{-}D_j)/L.$$

Let us introduce a random variable for matches,

$$M_{ij,k} = \begin{cases} 1 \text{ if } i \text{ and } j \text{ match on question } k \\ 0 \text{ otherwise.} \end{cases}$$

Then adding all the four possibilities we get

$$\Pr(M_{ij,k}=1) = D_iD_j + D_i(1-D_j)/L + D_j(1-D_i)/L + (1-D_i)(1-D_j)/L,$$

which reduces algebraically to

$$\Pr(M_{ij,k}=1) = D_iD_j + [1-D_iD_j]/L. \tag{5}$$

Since Eq. 5 does not have the question subscript k on the right-hand side, we can replace $\Pr(M_{ij,k}=1)$ by the observed proportion of matches, M_{ij}, from the data on all questions and solve for D_iD_j to obtain an estimate of D_iD_j given by

$$\hat{D_iD_j} = (LM_{ij}-1)/(L-1). \tag{6}$$

Note that Eq. 6[3] is a close parallel to Eq. 4. Unlike Eq. 4, however, Eq. 6 cannot be used directly to provide separate estimates of D_i and D_j because there are two unknown competencies and only one equation. The key to the method of estimating competencies lies in the fact that the response profile matrix **X** provides $N(N-1)/2$ independent equations like Eq. 6, one for each distinct pair of informants. Thus, there are $N(N-1)/2$ equations in terms of N unknown competencies, so that as long as $N \geq 3$, there are more knowns than unknowns.

In order to write out the entire set of equations for solution, we define

$$M^*_{ij} = (LM_{ij}-1)/(L-1), \tag{7}$$

which is an empirical point estimate of the proportion of matches between informants i and j *corrected for guessing* (on the assumption of no bias).

The set of equations can be written in matrix notation as follows:

$$\begin{pmatrix} D_1^2 & M^*_{12} & \cdots & M^*_{1j} & \cdots & M^*_{1N} \\ M^*_{21} & D_2^2 & \cdots & M^*_{2j} & \cdots & M^*_{2N} \\ \vdots & \vdots & & \vdots & & \vdots \\ M^*_{i1} & M^*_{i2} & \cdots & M^*_{ij} & \cdots & M^*_{iN} \\ \vdots & \vdots & & \vdots & & \vdots \\ M^*_{N1} & M^*_{N2} & \cdots & M^*_{Nj} & \cdots & D_N^2 \end{pmatrix} = \begin{pmatrix} D_1 \\ D_2 \\ \vdots \\ D_i \\ \vdots \\ D_N \end{pmatrix} (D_1, D_2, \ldots, D_j, \ldots, D_N) \tag{8}$$

where, of course, $M^*_{ij} = M^*_{ji}$ for all pairs of informants i and j.

Equation 8 represents an overspecified set of equations and because of sampling variability it is unlikely that they can be solved exactly. However, it is possible to obtain an approximate solution to Eq. 8 and thereby obtain estimates of the individual competencies D_i. The general approach to such problems is to select some criteria of goodness of fit, say least squares, and then to calculate estimates $\hat{D}_i$ that minimize the sum of the squared discrepancies between observed and predicted values of $\hat{M}^*_{ij}$. A least squares fit of the equation above is directly obtainable through the use of a version of factor analysis called the minimum residual method, first described by Comrey (1962). A version that accomplishes the same end is available on SPSS in the PA2 option (Nie et al. 1975:480). In our application we specify just one factor that gives direct estimates of the D_i for each individual. If the assumptions hold there should only be a single factor so that the first latent root should be large with respect to all other latent roots (see Lord and Novick 1968:381–382). We will discuss the criteria for fitting the model later in the paper. In any event, this procedure gives us the estimates of each informant's cultural competence D_i in terms that can be interpreted as the proportion of the questions they actually "know."

We now turn to the problem of how to infer the correct answers to the question. We give a formal presentation of a Bayes' Theorem approach to the problem in Appendix A. Now we give an example meant to give an intuitive feel for the approach to be taken.

To illustrate our approach, suppose we have only two informants, 1 and 2, and one true-false question. Suppose we know the competencies to be $D_1 = .8$ and $D_2 = .2$. If we know nothing at all, our a priori probabilities that the question is correctly answered true or false are .5 and .5, respectively. However, when we know the informants' responses, we are in a position to compute a posteriori estimates of the probabilities of the correct answer being true or false. The basic information is given in Table 1. The four logically possible response patterns involving two informants and one question are presented, where 1 codes a "true" response and 0 a "false" response. The probability of each pattern is computed on the assumption the correct answer is "true" and on the assumption the correct answer is "false." For example, assume the correct answer is "true," then the probability of both informants answering true (response pattern [1,1]) is the probability of the first informant being correct times the probability of the second informant being correct. This is computed from the competence using Eq. 3 with $L = 2$. Thus we have $(.8 + [1-.8]/2) \times (.2 + [1-.2]/2) = .54$ as shown in Table 1 for a response pattern of (1,1) where the correct answer is "true."

Bayes' Theorem in elementary probability theory provides the machinery for computing the a posteriori probabilities of true and false, respectively, given the a priori probabilities and the "evidence" of the informants' responses. Let X_1 and X_2 be the responses of the two informants, Pr(T) and Pr(F) be the a priori probabilities, and $\Pr(T \mid <X_1, X_2>)$, $\Pr(F \mid <X_1, X_2>)$ the desired a posteriori probabilities. Then Bayes' Theorem, adapted to our case, requires

$$(9) \qquad \Pr(T \mid <X_1, X_2>) = \frac{\Pr(<X_1, X_2> \mid T)\Pr(T)}{\Pr(<X_1, X_2> \mid T)\Pr(T) + \Pr(<X_1, X_2> \mid F)\Pr(F)},$$

where, for example, $\Pr(<X_1, X_2> \mid T)$ is the conditional probability of the evidence if the correct answer is true. Columns 2 and 3 in Table 1 give the conditional probabilities of the evidence given the correct answer is T or F, respectively, and columns 4 and 5 give the a posteriori probabilities from Eq. 9. For example, suppose $X_1 = 1$ and $X_2 = 1$, then

$$\Pr(T \mid <1, 1>) = \frac{.54 \times 1/2}{.54 \times 1/2 + .04 \times 1/2} = .931$$

which is the first entry in column 4. The rest of the values in columns 4 and 5 are obtained similarly from Eq. 9 and from the fact that $\Pr(F \mid <X_1, X_2>) = 1 - \Pr(T \mid <X_1, X_2>)$.

In Appendix A, the approach illustrated by this sample example is extended to cover the general case. This requires extension in the following ways: (1) it must handle the case of an arbitrary number of possible answers; (2) it must allow the evidence to come

Table 1

Illustrative data for computing a posteriori probabilities for a single question given two informants with competencies of .8 and .2, respectively.

Response pattern	Probability if correct answer		A posteriori probability	
	is true	is false	for true	for false
1 1	.54	.04	.931	.069
1 0	.36	.06	.857	.143
0 1	.06	.36	.143	.857
0 0	.04	.54	.069	.931

from an arbitrary number of informants; (3) it must handle an arbitrary number of questions; and (4) it must provide a way of using estimated competencies (from solving Eq. 8) in place of true competencies, which are not known. Despite these extensions, the logic behind the Bayes' Theorem applied to the preceding example is the key to the method of inferring correct answers.

Example 1. General Information Test

In this section of the paper we illustrate our procedures with true-false type data from a general information test. We have chosen an illustrative example where we know the answers a priori so that we can compare the results produced by our procedures to known benchmarks. Such a "validation" procedure provides insight into what we could expect in cases where we really do not know the answers ahead of time. In addition we feel that it is important to provide an example with real data so that the reader can better follow the ideas and procedures.

We constructed a 40-item test of general information from questions developed by Nelson and Narens (1980) to study long-term memory phenomena. Their "300 general-information questions were developed from fact books, with the aid of almanacs, atlases, trivia books, friends, and colleagues. All questions pertained to information that was at least 10 years old" (1980:339). We selected 40 of their questions from the median difficulty range and converted them to a true-false format. Roughly half of the correct answers were true and half false (19 true, 21 false). We collected responses to the 40-item questionnaire from 41 randomly selected students in the UCI student union during their leisure time. A copy of the questionnaire appears in Appendix B to facilitate replication.

We do not mean to imply that the General Information Test pertains to a given culture pattern. It probably does not generalize to other areas of culture, and it is meant only as an illustration. In fact we believe that there is great variability in what informants know about various domains of their culture. Informants who may know a great deal about sailing, for example, do not necessarily know much about tennis. We would also expect that there is greater consensus among informants in some areas than others.

Our main interest in looking at all 41 students is to illustrate how closely our estimates of each student's competence on the test $\hat{D}_i$, computed from the pattern of matches among students by solving Eq. 8, correspond to the estimates of competence computed from the actual proportion of correct answers using Eq. 4. After presenting these results we will provide a detailed numerical example of all our procedures on a small subset of the larger sample.

To obtain estimates of each student's competence on the General Information Test without using our knowledge of the answers, we first construct a 41-by-41-student matrix of the proportion of matches among all pairs of students. This is done by taking each of the $N(N\text{-}1)/2$ (820 in our example) pairs of students, and for each pair, counting the number of questions to which the pair had identical responses and divide by the number of questions. We then corrected for guessing by using Eq. 7. The resulting matrix constitutes the system of equations from Eq. 8. We then applied the minimum residual method of factor analysis using the PA2 option in SPSS (Nie et al. 1975:480). This provided us with the estimates of each informant's competence on this set of questions.

These estimates of competence may be interpreted as estimates of the proportion of questions that each student "knew" the answer. Since, in this example, we know the answer key for the questions, we can calculate each student's actual score and correct this score for guessing using Eq. 4 to obtain a traditional estimate of competence. Note that both of these methods only provide us with estimates of competence since guessing and other factors add sampling variability to each student's performance so that their "true" underlying competence is never directly observable. The mean and standard deviation of the estimates based on matches is .54 and .17, respectively, while the corresponding figures estimated from the key are .49 and .19. The correlation coefficient between the

two estimates is .93. The two estimates are, for all practical purposes, interchangeable.[4] Thus the estimate obtained without knowledge of the answers is comparable to the estimate based on the knowledge of the answers.

Even though these methods seem robust it is important to note that the original data *must* conform to the assumptions stated earlier or the method may give false or nonsensical answers. The model does not allow for negative true competence, for example. Also, it assumes that the questions are part of a shared belief system, not idiosyncratic preferences, and that the informants are all from a single coherent culture.

How can the researcher know whether the data conform to the assumptions, that the questions are in fact tapping a coherent cultural domain, that the informants are in fact from a single culture, etc.? Fortunately there is a major criterion that normally suffices to ensure that our assumptions are not violated in the structure of the data. The assumptions imply that the matrix of corrected matches in Eq. 8 has a single factor structure. This simply means that a single underlying all-positive factor, in our case competence, accounts for all of the structure in the matrix other than sampling variability. In statistical terms it means that the first factor has all positive values and accounts for several times as much variance as the next factor and that all other factors are relatively small and diminish slowly in size. This is a sufficiently stringent criterion that it normally guarantees that we can assume that the first factor is an estimate of informant competence.

The eigenvalues for each of the first five factors in the General Information Test are 13.95, 2.71, 2.28, 2.06, and 1.96. Note that the first factor, our estimate of competence, is not only all positive but is also over five times as large as the second factor and that the remaining factors are all small and trail off slowly. One can understand why this criterion is crucial if we look at Eq. 8. There we can see only one set of D_i's. If more than one factor were present the system of equations pictured in Eq. 8 would fail to accommodate the surplus data, and to fit the data would require parameters not in the theory. Since the violation of any of the three assumptions affects the factor solution, obtaining an appropriate factor solution is the main requirement in judging the fit of data to the model.

The next task is to infer the correct answers to the General Information Test. To accomplish this we applied the Bayes' Theorem method of estimating the answer key (see Appendix A). This method correctly classified all questions but one as "true" or "false." In fact we found that the a posteriori probabilities were so close to 1 and 0 that they were uninteresting in some sense. This is due in part to the fact that the number of informants is so large. We might note that the one question the method misclassified in the general information test was an item worded as follows: "Burton is the last name of the star of Spartacus." This question is in fact false but the majority of students responded with "true." We can never guard against such errors completely.

Example 2. Subset of General Information Test

We turn now to a detailed numerical example of a small sample of four of the students. In practice anthropologists must often deal with very small samples of informants. We want to illustrate that the method will work with small samples, although there will be a larger error variance than with larger samples. However, based on work reported later (see Table 5), we expect that on high concordance culture patterns samples of six to ten informants will work very well as a base to estimate the answer key. In any event we want to aggregate knowledge across a very small sample of informants to illustrate the possible power of the method.

We picked two high-competence and two low-competence students from the sample of 41 presented above. The response profile data for these four students is given in Table 2.

In order to prepare our system of equations as given in Eq. 8, we go through three simple steps: (1) count the number of matches between each pair of students; (2) divide this count by the number of questions (40) to get proportion of matches; (3) apply Eq. 7 with $L = 2$ to obtain an empirical estimate of the proportion of matches *corrected for guessing*. The results of the three steps are shown in Table 3.

Table 2
Response profile data for four students and 40 questions.

1	1	1	0	0	1	0	0	0	0	1	1	1	0	0	0	0	1	1	0	0	1	0	1	1	0	1	1	1	1	1	1	0	1	0	1	1	0	1	0	1
0	1	1	0	0	1	0	0	1	1	1	0	1	1	0	0	1	1	1	0	1	1	1	0	0	1	1	1	1	1	1	1	0	0	0	1	0	0	1	0	1
0	1	0	0	0	1	0	0	1	1	1	0	1	1	0	0	1	1	1	0	0	0	1	0	0	1	1	1	1	0	1	0	1	0	1	0	1	0	0	1	0	0
0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	1	1	0	1	1	0	1	1	0	0	1	0	0

Table 3
Matches, proportion matches, and corrected matches for four students.

Matches				Proportion				Corrected matches			
...	27	25	22	....	.675	.625	.550	...	.35	.25	.10
27	...	34	21	.675	....	.850	.525	.35	...	.70	.05
25	34	...	23	.625	.850	....	.575	.25	.70	...	.15
22	21	23	...	.550	.525	.575	....	.10	.05	.15	...

Table 4
Three different estimates of competence for the four student example.

Student	Competence from 4	Competence from all 41	Competence from score
2	.36	.33	.35
5	.90	.78	.80
9	.76	.74	.80
29	.12	.18	.25

The corrected matches are then analyzed by the minimum residuals method as implemented in SPSS option PA2. The estimates of competence are compared to the solution for all 41 students and to the known answers in Table 4. A comparison of all three columns indicates that the data based on just four informants yielded reasonably accurate competence estimates.

To infer the correct answers, we used the methods presented in Appendix A. In particular, we used Eq. 16 since $L = 2$. Table 5 summarizes the results.

Table 5 presents the 16 possible response patterns to a given question by the four informants, i.e., the 16 possible patterns of evidence on any questions. The a posteriori probabilities (or confidence levels) for a "true" answer are computed from Eq. 16. The confidence levels for a "false" answer are 1.0 minus the probability of a "true" answer. Since our a priori probabilities are equal for "true" and "false," it is natural to conclude that evidence patterns with values greater than .5 should lead to a reconstructed answer of "true" while patterns with values less than .5 lead to an answer of "false." This is indicated by the break in the table as well as the inference listed in the last column. The number of questions that yielded each response pattern is shown along with the number of questions (3) that were incorrectly classified by the method. It seems remarkable that just four informants, including two rather bad ones, provide enough evidence to correctly classify 37 of the 40 questions.

Note that Table 5 reveals an interesting symmetry, namely, that the values of the a posteriori probability of "true" and "false" are interchanged when the responses of each of the four informants are interchanged (for example, rows 3 and 14). This symmetry is a consequence of the model assumptions of no response bias and the fact that D_i is the

Table 5
Bayesian calculations for the 16 logical response patterns for four informants with D_i's of .37, .91, .77, and .13, respectively.

Response pattern	Likelihood		A posteriori probability of		No.of questions[a]	Inferred answer
	Ratio	Log	True	False		
1 1 1 1	440.362	6.088	.998	.002	8	True
1 1 1 0	264.235	5.577	.996	.004	4	True
0 1 1 1	94.861	4.552	.990	.010	0	True
0 1 1 0	56.920	4.042	.983	.017	6*	True
1 1 0 1	7.731	2.045	.885	.115	1	True
1 1 0 0	4.639	1.534	.823	.177	3**	True
0 1 0 1	1.665	.510	.625	.375	0	True
1 0 1 1	1.001	.001	.5001	.4999	0	True
0 1 0 0	.999	−.001	.4999	.5001	1	False
1 0 1 0	.600	−.510	.375	.625	1	False
0 0 1 1	.216	−1.534	.177	.823	0	False
0 0 1 0	.129	−2.045	.115	.885	0	False
1 0 0 1	.018	−4.042	.017	.983	0	False
1 0 0 0	.011	−4.552	.010	.990	5	False
0 0 0 1	.004	−5.577	.004	.996	5	False
0 0 0 0	.002	−6.088	.002	.998	6	False

[a]No. of questions with this pattern. Each * indicates one error in classification occurred for that pattern.

probability that informant i knows the correct answer to a question irrespective of whether it is a true or false item. Notice that the first and last three response patterns in the table all give a posteriori probabilities of .99 or better, and 28 questions yielded one of these patterns. The inferred answer to these questions can be accepted with a very high degree of confidence, and in fact, all 28 questions are correctly assigned the correct answer by the method. The three questions misclassified by the method were all assigned lower probability levels. Of the ten questions in which two students answered "true" and two students answered "false," seven were correctly classified by the method and all three misclassified questions came from such a pattern.

Sample Size Requirements

For the General Information example, 41 informants decisively and correctly classified all questions as "true" or "false" (except the Spartacus question discussed earlier). In addition, a selected sample of only four informants correctly classified all but three of the classifiable questions, 28 of which were decisively classified with a confidence level exceeding .99. These extremes ($N = 41$ and $N = 4$) suggest an interesting general question, namely, what is the minimal number of informants needed to describe a cultural domain by our methods? By describe, we mean to be able to confidently infer the answer to most questions.

It is possible to use our model to derive the minimal number of informants, N, needed as a function of a few crucial factors. The factors that determine the number of informants needed are as follows: first, minimum sample size will depend on the cultural competence of the pool of informants used. The higher the average competence of the sample the smaller the sample needed. Second, the investigator must set an appropriate confidence level, that is, the minimal value of an a posteriori probability that will be acceptable to decisively determine the answer to a question. The higher this level is set, the more informants will be needed. Third, the proportion of questions that one wants to decisively

and correctly classify, given an average cultural competence and a specified confidence level, affects the number of informants needed. The larger the proportion of questions one wants classified the greater the number of informants needed. Questions not decisively classified will either remain unclassified for the given confidence level or be misclassified. In none of the following calculations do the misclassified items constitute as high as 1%.

Although we do not present the derivations here, we have used the model to derive the minimum number of informants needed to achieve the desired accuracy as a function of these three factors. The derivations assume a true-false ($L = 2$) format and a pool of informants that are homogeneous in competence.[5] The results are presented in Table 6.

Table 6 lists competence levels from .5 to .9 in steps of a .1 along the columns. The major row headings list selected confidence levels from .90 through .999. They refer to the lowest acceptable value of the a posteriori probability chosen by the investigator to classify a question as "true" or "false." The minor row heading gives the lowest acceptable proportion of questions that will be decisively classified given various row and column choices. In the body of the table, the integers report the minimal number of informants needed for each of the cells. For example, when average competence is .7, confidence level is .99, and the proportion classified is .95, the minimum number of informants needed is shown to be 9. This means that 9 informants, with mean competence of .7, in

Table 6
Minimal number of informants needed to classify a desired proportion of questions with a specified confidence level when average cultural competence is known (confidence levels of .9, .95, .99, and .999 are included).

Proportion of questions	Average level of cultural competence				
	.5	.6	.7	.8	.9
.90 Confidence level					
.80	9	4	4	4	4
.85	11	6	4	4	4
.90	13	6	6	4	4
.95	17	10	6	6	4
.99	25	16	10	8	4
.95 Confidence level					
.80	9	7	4	4	4
.85	11	7	4	4	4
.90	13	9	6	4	4
.95	17	11	6	6	4
.99	29	19	10	8	4
.99 Confidence level					
.80	15	10	5	4	4
.85	15	10	7	5	4
.90	21	12	7	5	4
.95	23	14	9	7	4
.99	*	20	13	8	6
.999 Confidence level					
.80	19	11	7	6	4
.85	21	13	8	6	4
.90	23	13	10	8	5
.95	29	17	10	8	5
.99	*	23	16	12	7

Note: *Well over 30 informants needed.

response to a true-false questionnaire, have at least a .95 probability of correctly classifying each question with an a posteriori probability or confidence level of at least .99.

The most stringent levels in our table are the .999 confidence level and the .99 value of proportion of questions. When these levels are reached, virtually every question is correctly and decisively classified with near-perfect confidence. It is interesting that only 16 informants are needed to achieve this goal if the average competence is .7 and only 12 are needed if it is .8.

What of the questions that are not classified correctly by the method? Our analysis shows that in no case listed in Table 6 is there as much as a 1% chance of misclassification. Thus when the responses of the informants lead to a decisive classification at the selected confidence level, it is essentially always a correct one.

The use of the method with small samples of subjects and items is in rather striking contrast to related psychometric methods. For example, Nunnally (1978:262), among others, recommends sample sizes of 300 to 1,000 and the use of a large number of items with "at least five times as many persons as items." Lord and Novick (1968) present figures based on a sample of 107,234 cases. Lazarsfeld and Henry (1968) use a small number of questions but say we should have samples of subjects of at least 1,000. Are we really justified in using as few as a half-dozen subjects with only a few dozen items? We feel that the answer is yes for the following reasons: (1) we have a very tight theory whose assumptions are very stringent; (2) we are working with very high concordance codes where consensus is high; and (3) we are only trying to find one "correct" answer for a question rather than, say, differentiating questions on a continuous scale of tendency to be "true" or "false."

Example 3. Disease Classification in Guatemala

In a series of recent articles, one of the authors has contributed to the understanding of intracultural variability and the validation of cultural beliefs utilizing data on diseases collected in Guatemala and Mexico (Weller 1983, 1984a, 1984b). In her research she measured "agreement among informants to assess the relative cultural salience of each illness concept. It is assumed that illness concepts with the highest agreement are culturally more salient than those with lower agreement" (1984a:341). In an urban Guatemala setting (population 21,000), she had informants rank order 27 diseases on "degree of contagion" and "those needing the hottest remedy or medicine to those needing the coldest remedy or medicine" (1984a:342). Using a variety of quantitative methods, she demonstrated that informants agreed more among themselves on the concept of contagion than on the concept of whether hot or cold medicines were needed. In a more detailed study using additional data, she added evidence to the lack of informant agreement on the hot-cold concept (1983). Since the results are documented in detail, we feel that the urban Guatemala data represent a good natural situation to analyze with our methods.

To simplify the analysis we used a dichotomous form of the rank order data that was provided by the fact that each informant was asked to divide the diseases into contagious and noncontagious diseases before doing the complete ranking. The same procedure was followed for the hot-cold data. We used this dichotomized data in the following analysis.

We prepared the matrix of adjusted matches for both sets of data as specified in Eq. 8. The minimum residual method of factoring the matrix gave the proportion of variance accounted for in each of the first four factors for contagion as follows: .69, .08, .05, and .03. Since the first factor is all positive and the contagion data is fit with a one-factor solution, it fulfills the assumptions of the model. Thus, we used the Bayes' Theorem method (Appendix A) to classify the diseases as contagious and noncontagious. Results indicated a high level of competence. The average level as estimated by the first principal component was .82 $\pm$ .11. Each disease was classified at a very high level of confidence (beyond the .9999 level).

The hot-cold data, on the other hand, produced two factors of about the same size (values for first four are .23, .20, .10, and .10), the first of which had 11 negative values. Thus, the data do not satisfy the dominant all positive single factor expected from the model. To check the notion that there might be two separate cultures on hot-cold, we did an analysis of the 12 positive cases. Even this subset of informants did not give a single factor result. The first factor had negative values and the first four factors accounted for the following proportion of variance: .24, .22, .14, and .13. We conclude, therefore, that the informants do not share a coherent set of cultural beliefs concerning what diseases require hot or cold medicines.

One way of getting a graphic idea as to why the method is not appropriate in the case of the hot-cold data is to study the agreement among informants on the raw data. Table 7 lists the 27 diseases studied together with the number of informants classifying each disease as "not contagious" and "not needing cold" medicines. Figure 1 plots the data in Table 7 as frequency distributions. Note that in the case of contagion that the data are very bimodal indicating that informants agree on how they classify the disease. All 24 informants agree that six diseases (arthritis, colic, diabetes, kidney pain, gastritis, and rheumatism) were considered noncontagious and three (measles, whooping cough, and smallpox) were contagious.

The hot-cold data are, by contrast, fairly unimodally distributed. None of the diseases is unanimously classified by all informants (there were 23 rather than 24 informants on

Table 7

Guatemalan disease terms and the number of informants classifying diseases as not-contagious and as not-cold.

Disease	Informants classifying disease as	
	Not-contagious	Not-cold
1. Allergies	20	20
2. Amoebas	8	15
3. Tonsilitis	10	11
4. Appendicitis	23	15
5. Arthritis	24	6
6. Cancer	22	12
7. Colic	24	4
8. Diabetes	24	12
9. Diarrhea	20	10
10. Diphtheria	5	7
11. Kidney pain	24	19
12. Gastritis	24	16
13. Flu	1	1
14. Hepatitis	3	14
15. Intestinal	22	12
16. Malaria	16	5
17. Mumps	1	13
18. Polio	17	11
19. Rheumatism	24	3
20. Rubella	2	13
21. Measles	0	13
22. Tetanus	21	12
23. Typhoid fever	3	10
24. Whooping cough	0	6
25. Tuberculosis	1	6
26. Chicken pox	2	12
27. Smallpox	0	13

the hot-cold task). Thus the evidence of consensus analysis reinforces the previous analysis by Weller, and it provides an example of how the violation of the model assumptions by the data is signaled in the analysis.

Summary and Discussion

We have described a model for the analysis of culture knowledge based on the consensus among informants. We illustrated the analysis on data from a general information test on a sample of 41 informants and then focused on a sample of only four informants for a more detailed empirical examination. Finally we reanalyzed some data on the classification of diseases from Guatemala. There remain, however, several important topics that need further clarification and discussion. These topics will be discussed in the following order: (1) The problem of the validity of the method, does it measure what we think it measures? (2) The problem of the reliability of the method, how well do we measure what we measure? (3) How much consensus among informants can be expected and how much is necessary to define culture patterns? (4) How robust is the method and what are some of the dangers and pitfalls? (5) How many informants do we need? (6) Possible directions and prospects for further methodological developments.

1. Validity

The validity of the theory is of prime importance. When we factor the matrix of adjusted matches and obtain estimates of cultural competence represented in the theory by the D_i's (assuming a legitimate single factor solution) do they "really" indicate that informants with high cultural competence "know" more about the domain tested than do less-competent informants? Validity is a complicated concept and difficult to define precisely: however, we take it to mean that our measures relate in known and precise ways to other variables that we accept as measuring substantially the "same" thing as we think we are measuring.

For example, when our competence estimates in the General Information Test correspond closely to the "known" scores on the test we assume that this fact provides evidence for validity. In other words, we found that after going through some rather detailed computations on a matrix of informant-by-informant proportion of matches data, we came out with essentially the same results as simply adjusting the proportion correct. Since we accept at "face value" the idea that competence is related to proportion correct, we then feel that our procedure measures the "same thing." Some might say that our procedure worked in this one case because of chance or because of the special nature of the data which is objective fact (not, for example, a shared cultural belief). Are there other examples that might be interpreted as indicating validity?

In two class examinations of a multiple-choice format,[6] we applied the model in order to test the correspondence between estimates of competence based on "known" answers and those obtained by consensus methods. In the first test, a final in a History of Psychology class with a sample of 60 students, using a five alternative multiple choice format, the correlation between the two estimates was .978. In the second test, a midterm in a Psychology of Humor class with 35 students, using a four alternative multiple choice format, the correlation between the two estimates was .89. These results fully support the analysis used in the paper, although they are limited to objective test situations in a classroom. Would the results hold up in other areas?

Other research that presents data relevant to the validity of related methods include Romney and Weller's (1984) attempt to predict the accuracy of recall, Boster's (1986) study on consensus and cultural knowledge, and Garro's (1983) finding that older women and curers have more consensus concerning beliefs about diseases than do younger women and noncurers. A final example can be found in Weller, Romney, and Orr (1986) where individual correspondence to consensual values regarding the appropriateness of discipline techniques was used to predict familial use of corporal punishment.

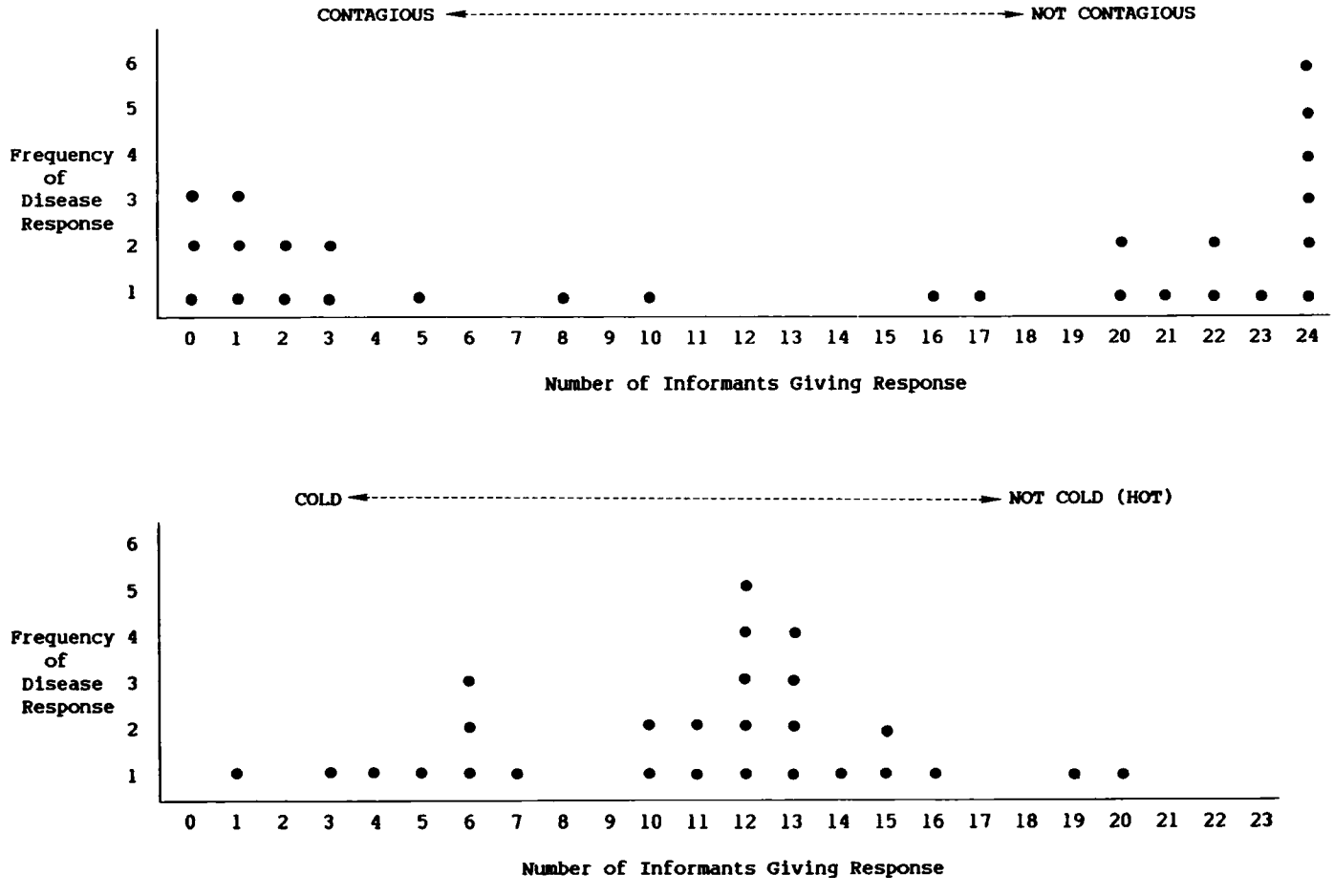

Figure 1
Distribution of informants saying that a disease does not have a given characteristic (contagion at top and cold at bottom).

We feel that the kinds of evidence referred to here demonstrate that the model does have a certain amount of validity in the sense that it does measure what we think it is measuring in those areas so far tested. Of course it remains to conduct more sophisticated tests and to establish how far we can generalize the application of the theory and still obtain valid results.

2. Reliability

The second traditional concern with new methods centers upon the reliability of the results of the analysis. We interpret reliability to be related to the stability and accuracy of the measurement based on criteria such as retesting or internal consistency. In a variety of simulation studies, which we do not have space here to discuss at length, we have some relevant findings that are worth mentioning. We have found that, other things being equal, fill-in-the-blank type questions are more reliable than multiple-choice type, which are in turn more reliable than true-false type. In replicating the General Information Test in a fill-in-the-blank type format, for example, we found that it took about five true-false type questions to produce the same test-retest reliability as one fill-in-the-blank question. Thus the format of questions chosen by Boster to study manioc (1986) was optimal for reliability.

We also found in other contexts that there are two kinds of reliability that need to be distinguished based upon whether one is dealing with *response* profile data or *performance* profile data (see Batchelder and Romney 1986 for formal definitions). In this paper we have dealt solely with *response* profile data, which enables us to measure the reliability of the informants to sort questions into response categories, e.g., "true" or "false." In traditional psychometric work the data are coded "correct" or "incorrect" and so we label it *performance* profile data. Here reliability refers to how well the test items measure differences among the subjects (Nunnally 1978).

In traditional usage, like Cronbach's (1961) alpha, for example, the measure of item reliability would increase as the variability of the subjects increases. For example, it is easier to reliably differentiate between very smart and very dumb people than it is to reliably measure the difference between people almost the same in intelligence. In anthropological fieldwork we frequently deal in high concordance cultural codes so that all of the informants have high cultural competence. The lack of variability makes traditional measures of item reliability based on performance data inapplicable.

A final comment on reliability. There seems to be a limit on the extent to which you can simultaneously have, for a given set of data, both high informant reliability (on the *response* data) and high item reliability (on the *performance* data). The highest possible informant reliability (ability of informant to reliably sort questions) would arise with all informants being totally competent and the same number of questions in each response category. If the informant variability is low, as in this case, then by definition the items cannot reliably distinguish among informants, that is, low item reliability. This phenomenon needs further thought. In particular, anthropologists, unlike psychologists, need not seek high item reliabilities for tests so long as informant reliabilities on response data are high.

3. Consensus

How much consensus does it take to define a culture pattern? In our theory we have built in the assumption that all the informants are reporting on the same culture and that all informants have non-negative competence. One of the problems that we face is how to decide how much consensus is necessary in order for us to infer the existence of a clearly defined culture pattern. In order to get an idea of how much consensus may be expected, we can compare the average correlation among informants in the data sets studied so far. There are $N(N-1)/2$ pairs of informants in any study and by taking the mean of the correlations across all these pairs we can get a good idea of how much consensus exists for that group on that domain. Mean competence could also serve as an indicator. In

fact, for a true-false format, it can be shown that the two measures are closely related to each other, the mean informant-by-informant correlation is approximately the square of the mean competence.

In the data reported in this paper the mean of the informant-by-informant correlations was .30 among the 41 informants on the General Information Test, .69 among the 24 informants on the contagion data, and .13 among the 23 informants on the hot-cold data.

The data for the 24 Guatemalan informants on the contagion of diseases clearly stand out. The average informant-by-informant correlation is more than twice as large as any of the other samples. It is the only sample we have that clearly reports on a cultural belief pattern. We feel that Roberts and Chick's (1979) notion of high concordance codes is fully applicable in this situation. In general, we believe that cultural patterns are high consensus codes, and we have shown in Table 7 that a relatively small number of informants is sufficient to correctly classify most questions. Notice that the hot-cold data not only do not satisfy the assumption of a single factor structure but that the informant-by-informant correlations are low.

Even though the General Information Test data do not deal with cultural patterns, they provide crude indicators of what to expect in vaguely coherent domains. The requirement that the first factor is several times as large as the second and all positive will ensure that one is dealing with a single culture. Clearly the lower the level of consensus the more informants one needs to reach the same degree of confidence in the results.

4. Robustness

Just how far can the methods be generalized and how robust can we expect the assumptions to be? One of the important assumptions that we make that is probably frequently violated is that the questions are of equal difficulty. In simulation studies we have found that the method is not very sensitive to even fairly large violations of this assumption. However, there is always a serious danger that a few questions are so different from the others in which they are embedded that the inferences we make about them are wrong. Probably it is unwise to put too much faith in the inferences concerning any particular item since at the moment we do not know how to prove that any particular item belongs in the domain and is of a comparable difficulty level.

In the case of the disease study, for example, we feel very confident that the informants could classify diseases on the basis of contagion. There is a certain parallelism in asking the same question about each disease. In other domains we may not have as good a reason for feeling that the questions all concern the same domain and are of comparable difficulty levels.

The method seems very sensitive, and lacks robustness, with respect to "negative" cultural competence, that is, informants who "know" incorrect answers. This is probably a good feature since it does not make theoretical sense to posit a coherent culture in which, nevertheless, each informant may have an idiosyncratic view. In such a world we could not agree on the meanings of words, symbols, etc. and the cumulation of culture as we know it would not be possible. Other kinds of models need to be constructed to describe such "preference" data.

5. Number of Informants

One of the important findings that needs comment is the ability to estimate the number of informants necessary to classify a given proportion of questions at a specified level of confidence. It should be noted that the numbers given in Table 6 are minimal numbers. Due to the noncontinuous nature of the binomial distribution it is sometimes possible for a larger number of informants to actually classify a smaller proportion of questions at a given level than might be obtained with a smaller number. For this reason it is best not to take Table 6 too literally. The numbers are meant as a rough guide and to illustrate that it is possible to get stable results with fairly small samples.

It may also be noted that Table 7 does not list average competence below $D = .5$. The reason for this is that the model applies to high concordance cultural codes. If the average informant-by-informant correlations fall much below .3 the assumptions of the model probably are not met.

6. Prospects

We have presented the barest outline of a possible theory of cultural consensus. The first and most obvious need is to generalize to a wider variety of question formats. We are currently working to construct models for the analysis of judged similarity data collected using triads, for rank order data (a more difficult problem), judgment of quantitative variables, etc.

We also have more flexible ways of analyzing dichotomous data. Measuring agreement with covariance, for example, completely avoids the problem of response bias as covariance is invariant under different levels of response bias (Batchelder and Romney 1986). Thus in work where one is worried that informants may have differential response biases, one can compare the results from the matching approach (which is sensitive to response bias) to the results from the covariance approach (which is sensitive to proportion "true").

Methods need to be developed that provide better criteria for judging the coherence of the questions to ensure that they apply to a single well-defined cultural domain. Related to this we need to develop measures for the variability in difficulty in the questions and to find if adjustments in the confidence estimates are needed. Further work is also needed on the problem of reliability of both the estimates concerning the informants and the estimates concerning the questions.

Implications

We feel that the consensus model opens up the possibility of measuring the cultural competence of informants in a variety of domains. In addition it allows one to reconstruct the "culturally relevant" answers to the questions posed along with confidence limits on the reconstruction. These abilities of the model have potentially far-reaching implications for the study of anthropology.

Its use enables us to pick out our best (most competent) informants in each area of culture. Clearly no informant masters all the cultural knowledge, so the best set of informants in one area may not be the best set for another area or domain of culture. One of the most important implications is that we can rely on a small number of good informants as shown in Table 6. We do not have to have large samples to objectively ensure that we are confident of the answers. The model is sufficiently well defined and has stringent enough assumptions that we can expect stable results with a half-dozen or so informants in areas of high concordance. This is the first time, to our knowledge, that we can defend at the formal mathematical level the use of such small samples for the aggregation of cultural knowledge.

A closely related implication is that we can objectively distinguish between informants that have specialized knowledge of exotic or specialized cultural domains from those who do not. Garro's (1983) study of curers' versus noncurers' knowledge of disease characteristics in Pichataro, Mexico, is an example. The finding by Boster (1986) that women have more knowledge of manioc names than men is another example. The ability to objectively test the cultural knowledge of different subgroups of informants (without knowing the answers to the questions ourselves) should greatly expand the possibilities of studying intracultural variability.

The model should also contribute to the solution of some questions about what the cultural beliefs actually are in some cases. For example, many writers have assumed that the definition of the hot-cold concepts of medicines and diseases in Latin America are cultural beliefs analogous to severity or contagions (see review in Weller 1983). The

model allows an objective comparison between the hot-cold beliefs and other beliefs. Our theory makes it possible to compare the cohesiveness and strength of cultural beliefs from one domain to another.

The theory also facilitates comparisons across cultures. For example, if one believes that the failure of the hot-cold concept to emerge in the Guatemala studies was an accident of the particular cultural group studied, one can replicate the study in any number of societies and make comparisons. One of the valuable features of the theory is that the estimates of the parameters (e.g., cultural competence or D_i's) are always stated in the same metric (proportion of items known) so that one can compare from one culture to the next.

The comparability of results from one study to another raises the possibility that anthropologists could more easily compare results both within and among cultures. A variety of results could be accumulated and greatly facilitate our understanding of the distribution of cultural knowledge both within and among cultures.

Notes

Acknowledgments. The authors would like to thank Tarow Indow who provided the insight in solving Eq. 8 with factor-analytic procedures and to Katie Faust, Kathy Maher, and Ryozo Yoshino who served as research assistants on the project. John Boyd wrote the program to solve Eq. 8. Portions of the work were funded by NSF Grant Number SES-8320173 to Romney and Batchelder.

[1]The model becomes more complicated when different informants can have different response biases. Batchelder and Romney (1986) provide methods of handling response bias in case $L = 2$.

[2]In this case it is possible for two informants to have matching errors to a question. On the other hand, in the other two cases a match can only occur on a correct response. For a fill-in-the-blank format, it is essential that mutual blanks (no response) not be counted as a match.

[3]A practical problem with Eq. 6 is that it can yield a negative value of $\hat{D_i D_j}$ Of course the same problem can occur in Eq. 4, which is the usual formula for correcting for guessing when the answer key is known. The computer solution yielding separate competence estimates described next can tolerate a few negative values of $\hat{D_i D_j}$ and still yield individual non-negative competence estimates.

[4]Actually the estimates of competence based on the key average slightly lower than those based on matches. This is probably due in part to a bad question about the star of Spartacus discussed later. Another factor accounting for the difference in these two estimates is undoubtably sampling variability since the estimates are based on different aspects of the data.

[5]It is possible to show that if heterogeneity in competence is allowed, the N will never get higher than provided in Table 7 provided the mean competence in the heterogeneous sample is compared with a homogeneous group of the same competence.

[6]The analysis of these two examples was performed by Kathy Maher, a graduate student at UCI, and will be reported in detail in a later manuscript.

Appendix A: Bayes' Theorem Method

Given the response profile data $\mathbf{X}_{N \times M}$ in Eq. 1, we want to reconstruct the correct answers to the M questions with specified confidence levels. We generalize methods based on Bayes' Theorem provided in Batchelder and Romney (1986) and Nitzan and Paroush (1982) for the case of an arbitrary number of possible correct answers, namely, L. The method has several steps as follows: first, we consider a single question and assess our a priori probabilities of each of the L possible correct answers. Second, we assume known competencies of each informant, and given the evidence of the informants' responses, we compute a posteriori probabilities of each of the L possible correct answers using Bayes' Theorem. Third, we replace the informants' competencies in the Bayes' Theorem solution by the estimated competencies from solving Eq. 8. Finally, we note that each question can be classified independently.

First, without knowing the informants' responses to a particular question k, it is reasonable to set our a priori probabilities for each possible response equal, that is, the a priori probability of response l is set to $P_l = 1/L$ for each $l = 1,2, \ldots, L$.

Second, the evidence relevant to our a posteriori classification consists of the kth column of the response profile matrix in Eq. 1 which we denote for convenience by

$$E_k = <X_{ik}>_{i=1}^{N}$$

We want to compute a posteriori probabilities denoted by

$$(10) \qquad P_l(E_k) = \Pr(Z_k = l \mid <X_{ik}>_{i=1}^{N})$$

for each possible $l = 1,2,\ldots,L$. To computer Eq. 10, Bayes' Theorem (see any elementary probability book, e.g., Mosteller, Rourke, and Thomas 1961:146) can be applied. The result is

$$(11) \qquad P_l(E_k) = \frac{\Pr(<X_{ik}>_{i=1}^{N} \mid Z_k = l)\, P_l}{\sum_{z=1}^{L} \Pr(<X_{ik}>_{i=1}^{N} \mid Z_k = z)\, P_z}.$$

Since $P_z = 1/L$ for each response z, all that is needed to solve Eq. 11 are the conditional evidence probabilities of the form

$$\Pr(<X_{ik}>_{i=1}^{N} \mid Z_k = l).$$

To compute these conditional probabilities, first note that the model implies

$$(12) \qquad \Pr(X_{ik} = l \mid Z_k = l) = D_i + (1-D_i)/L$$

and

$$\Pr(X_{ik} \neq l \mid Z_k = l) = (1-D_i)(L-1)/L$$

for each response l. Next by the local independence assumption, Eq. 2,

$$(13) \qquad \Pr(<X_{ik}>_{i=1}^{N} \mid Z_k = l) = \prod_{i=1}^{N} \Pr(X_{ik} \mid Z_k = l).$$

To put these facts together we need to introduce the random variables

$$(14) \qquad X_{ik,l} = \begin{cases} 1 \text{ if } X_{ik} = l \\ 0 \text{ otherwise.} \end{cases}$$

and also replace the unknown competencies D_i in Eq. 12 by their estimated values $\hat{D}_i$ from solving Eq. 8. The result is

$$(15) \qquad \Pr(<X_{ik}>_{i=1}^{N} \mid Z_k = l) = \prod_{i=1}^{N} [\hat{D}_i + (1-D_i)/L]^{X_{ik,l}}[(1-\hat{D}_i)(L-1)/L]^{1-X_{ik,l}}$$

$$= \prod_{i=1}^{N} \left[\frac{\hat{D}_i(L-1) + 1}{(L-1)(1-\hat{D}_i)}\right]^{X_{ik,l}} \frac{(1-\hat{D}_i)(L-1)}{L}$$

Note that the quantities $X_{ik,l}$ and $\hat{D}_i$ both can be obtained from the response profile matrix $\mathbf{X}$, namely, Eq. 14 and Eq. 8, respectively. The numerical values of Eq. 15 for each response l can be inserted into Eq. 11 to yield a posteriori probabilities or confidence levels for each possible response l. Finally, we note that from the local independence assumption in Eq. 2, each question k can be decided independently if we are willing to accept the estimated $\hat{D}_i$'s. This step in replacing D_i with $\hat{D}_i$ in the computation is analogous to Lazarsfeld's method of computing recruitment probabilities (see, for example, Lazarsfeld and Henry 1968:69).

In the case of dichotomous response data, it is possible to simplify Eq. 11. If the response data are coded

$$X_{ik} = \begin{cases} 1 \text{ if response of informant } i \text{ to question } k \text{ is "true"} \\ 0 \text{ otherwise,} \end{cases}$$

then by making substitutions in Eq. 15, Eq. 11 becomes

$$(16) \quad P_1(E_k) = \frac{\prod_{i=1}^{N} [(1 + \hat{D}_i)/(1-\hat{D}_i)]^{X_{ik}}(1-\hat{D}_i)}{\prod_{i=1}^{N} [(1 + \hat{D}_i)/(1-\hat{D}_i)]^{X_{ik}}(1-\hat{D}_i) + \prod_{i=1}^{N} [(1-\hat{D}_i)/(1 + \hat{D}_i)]^{X_{ik}}(1 + \hat{D}_i)}$$

$$= 1/\left[1 + \prod_{i=1}^{N} [(1 + \hat{D}_i)/(1-\hat{D}_i)]^{1-2X_{ik}}\right],$$

and, of course, $P_0(E_k) = 1\text{-}P_1(E_k)$ is the a posteriori probability or confidence level for a classification of question k as "false." Naturally we would construct the correct answer for question k as $Z_k = 1$ if and only if $P_1(E_k) > .5$ and this is easily seen to occur in case

$$(17) \quad \prod_{i=1}^{N} [(1 + \hat{D}_i)/(1-\hat{D}_i)]^{1-2X_{ik}} < 1.$$

Taking natural logarithms of both sides of Eq. 17 yields the criterion that $P_1(E_k) > .5$ if and only if

$$(18) \quad \sum_{i=1}^{N} (2X_{ik}\text{-}1) \ln[(1 + \hat{D}_i)/(1-\hat{D}_i)] > 0.$$

Equation 18 is interesting because it shows that each informant's response coded as $2X_{ik}$-1 can be weighted by $\ln[(1 + \hat{D}_i)/(1-\hat{D}_i)]$ and summed to determine the reconstructed answer key. Further the weighting factor can be interpreted from Eq. 12 as the natural logarithm of the ratio of the probability of a correct response to the probability of a wrong response in case $L = 2$.

Appendix B: General Information Test with Answer Key

1. F Gyropilot is the name of the navigation instrument used at sea to plot position by the stars.
2. T Sputnik is the name of the first artificial satellite put in orbit by Russia in 1957.
3. T Nightingale is the last name of the woman who began the profession of nursing.
4. F Backgammon is the game in which the standard pieces are of staunton design.
5. F Brunn is the last name of the man who first studied genetic inheritance in plants.
6. T Bismarck is the name of Germany's largest battleship that was sunk in World War II.
7. F Clurmont is the name of the mansion in Virginia that was Thomas Jefferson's home.
8. F Dodgson is the last name of the author who wrote under the pseudonym of Mark Twain.
9. T Albany is the capital of New York.
10. T Shoemaker is the last name of the jockey with the most lifetime winners in horse racing.
11. T The last name of the man who invented the telegraph is Morse.
12. F Lock is the last name of the man who wrote the "Star Spangled Banner."
13. T Frank Lloyd Wright's profession was an architect.
14. T Hancock is the last name of the first signer of the "Declaration of Independence."
15. F Burton is the last name of the movie actor who portrayed Spartacus.
16. F The first country to use gunpowder was Nepal.
17. F The Italian city that was destroyed when Mount Vesuvius erupted in 79 A.D. was called Herculaneum.
18. T Nero is the name of the Roman emperor who fiddled while Rome burned.

19. T The capital of Hungary is Budapest.
20. F The spleen is the organ that produces insulin.
21. F Pitchblende is the name of the scientist who discovered radium.
22. T The Rhine is the river on which Bonn is located.
23. T Young is the last name of the actor who portrayed the father on the television show "Father Knows Best."
24. F The unsuccessful auto that was manufactured by the Ford Motor Company from 1957–1959 was the "Model N."
25. F Sabin is the last name of the doctor who first developed a vaccine for polio.
26. T Backus is the last name of the man who was the voice of Mr. Magoo.
27. T The mountain range in which Mount Everest is located is the Himalayas.
28. T Burr is the last name of the actor in the role of Perry Mason on television.
29. T The European city in which the Parthenon is located is Athens.
30. F The U.S. Naval Academy is located in the city of Arlington.
31. T The "Nautilus" is the name of the submarine in Jules Verne's "20,000 Leagues Beneath the Sea."
32. F Carrol is the last name of the author who wrote "Oliver Twist."
33. F Kane was the last name of the man who was the radio broadcaster for the "War of the Worlds."
34. F Iceland is the largest island excluding Australia.
35. T Chaucer is the last name of the man who wrote "Canterbury Tales."
36. F Ptolemy is the last name of the astronomer who published in 1543 his theory that the earth revolves around the sun.
37. F Buenos Aires is the capital of Brazil.
38. T The collar bone is called the clavicle.
39. F Jefferson is the last name of the man who was president directly after James Madison.
40. F The name of Alexander Graham Bell's assistant was Sanders.

References Cited

Batchelder, W. H., and A. K. Romney
1986 The Statistical Analysis of a General Condorcet Model for Dichotomous Choice Situations. *In* Information Pooling and Group Decision Making. B. Grofman and G. Owen, eds. Connecticut: JAI Press. (In press.)

Boster, J. S.
1986 Requiem for the Omniscient Informant: There's Life in the Old Girl Yet. *In* Directions in Cognitive Anthropology. J. Dougherty, ed. Urbana: University of Illinois Press. (In press.)

Brady, I.
1983 Speaking in the Name of the Real: Freeman and Mead on Samoa. American Anthropologist 85(4):908–947.

Comrey, A. L.
1962 The Minimum Residual Method of Factor Analysis. Psychological Reports 11:15–18.

Cronbach, L. J.
1961 Coefficient Alpha and the Internal Structure of Tests. Psychometrika 16:297–334.

D'Andrade, R. G.
1981 The Cultural Part of Cognition. Cognitive Science 5:179–195.

Frake, C. O.
1961 The Diagnosis of Disease Among the Subanun of Mindanao. American Anthropologist 63(1):113–132.

Freeman, D.
1983 Margaret Mead and Samoa: The Making and Unmaking of an Anthropological Myth. Cambridge, MA: Harvard University Press.

Garro, L.
1983 Individual Variation in a Mexican Folk Medical Belief System: A Curer-Lay Comparison. Ph.D. dissertation, University of California, Irvine.

Goodenough, W. H.
1964 Cultural Anthropology and Linguistics. *In* Language in Culture and Society. D. Hymes, ed. New York: Harper & Row.

Green, D. M., and J. A. Swets
1966 Signal Detection Theory and Psychophysics. New York: Wiley.

Grofman, B., S. L. Feld, and G. Owen
1983 Thirteen Theorems in Search of the Truth. Theory and Decisions 15:261–278.
Hogg, R. V., and A. T. Craig
1978 Introduction to Mathematical Statistics. New York: Macmillan.
Kroeber, A. L.
1948 Anthropology. New York: Harcourt, Brace.
1952 A Half-Century of Anthropology. *In* The Nature of Culture. A. L. Kroeber, ed. Pp. 139–143. Chicago: University of Chicago Press.
Lazarsfeld, P. F., and N. W. Henry
1968 Latent Structure Analysis. Boston: Houghton Mifflin.
Lord, F. M., and M. R. Novick
1968 Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley.
Mosteller, F., R. E. K. Rourke, and G. B. Thomas, Jr.
1961 Probability with Statistical Applications. Reading, MA: Addison-Wesley.
Nelson, T. O., and L. Narens
1980 Norms of 300 General-Information Questions: Accuracy of Recall, Latency of Recall, and Feeling-of-Knowing Ratings. Journal of Verbal Learning and Verbal Behavior 19:338–368.
Nie, N. H., C. H. Hull, J. G. Jenkins, K. Steinbrenner, and D. H. Brent
1975 SPSS: Statistical Package for the Social Sciences. 2nd edition. New York: McGraw-Hill.
Nitzan, S., and J. Paroush
1982 Optimal Decision Rules in Uncertain Dichotomous Choice Situations. International Economics Review 23:289–297.
Nunnally, J. C.
1978 Psychometric Theory. New York: McGraw-Hill.
Roberts, J. M.
1964 The Self-Management of Cultures. *In* Explorations in Cultural Anthropology. W. H. Goodenough, ed. Pp. 433–454. New York: McGraw-Hill.
Roberts, J. M., and G. E. Chick
1979 Butler County Eight Ball: A Behavioral Space Analysis. *In* Sports, Games, and Play. J. H. Goldstein, ed. Pp. 65–99. Hillsdale, NJ: Erlbaum Associates.
Roberts, J. M., G. E. Chick, M. Stephanson, and L. L. Hyde
1981 Inferred Categories for Tennis Play: A Limited Semantic Analysis. *In* Play as Context. A. B. Cheska, ed. Pp. 181–195. West Point, NY: Leisure Press.
Roberts, J. M., T. V. Golder, and G. E. Chick
1980 Judgement, Oversight, and Skill: A Cultural Analysis of P-3 Pilot Error. Human Organization 39(1):5–21.
Roberts, J. M., and S. Nuttrass
1980 Women and Trapshooting. *In* Play and Culture. H. B. Schwartzman, ed. Pp. 262–291. West Point, NY: Leisure Press.
Romney, A. K., and S. C. Weller
1984 Predicting Informant Accuracy from Patterns of Recall Among Individuals. Social Networks 4:59–77.
Sankoff, G.
1971 Quantitative Analysis of Sharing and Variability in a Cognitive Model. Ethnology 10:389–408.
Weller, S. C.
1983 New Data on Intracultural Variability: The Hot-Cold Concept of Medicine and Illness. Human Organization 42(3):249–257.
1984a Cross-Cultural Concept of Illness: Variation and Validation. American Anthropologist 86(2):341–351.
1984b Consistency and Consensus Among Informants: Disease Concept in a Rural Mexican Town. American Anthropologist 86(4):966–975.
Weller, S. C., A. K. Romney, and D. P. Orr
1986 The Myth of a Sub-Culture of Corporal Punishment. Human Organization. (In press.)