Consensus Analysis

Stephen P. Borgatti and Daniel S. Halgin
University of Kentucky

Introduction

As developed by Romney, Weller and Batchelder (1986), consensus analysis is both a theory and a method. As a theory, it specifies the conditions under which agreement among people can be seen as a sign of knowledge or "getting it right". Many folk epistemological systems rely on the connection between agreement and truth. An obvious example is the court jury system, which does not regard a prosecutor's claims as true unless a jury of 12 independent people agrees. Another example is the scientific practice of measuring things multiple times and taking the average as the best estimate. Influential books such as "The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations" (Surowiecki, 2004) are based on this argument. Similarly, Adam Smith argued that markets make better decisions about how to allocate goods than any single body could. We also see it in common phrases like "50,000,000 Elvis fans can't be wrong."[1] And yet, it seems obvious that billions of people *can* be wrong. Agreement does not always imply getting it right. What the theory of consensus analysis does is work out the special circumstances under which agreement really does imply knowledge.

As a method, consensus analysis provides a way of conceptualizing and coping with individual variability. As D'Andrade (1987:194) put it, "For a long time there has been a scandal at the heart of the study of culture. This scandal goes as follows: All human groups have a culture; culture is shared knowledge and belief; but when we study human groups, we find that there is considerable disagreement concerning most items of knowledge and belief." For example, we would like to report that, in western cultures, the color white connotes purity, but a survey asking respondents to associate colors with qualities will never give a 100% association between white and purity. What the method of consensus analysis does is to provide three things. First, it provides a way to determine whether observed variability in beliefs is cultural, in the sense that our informants are drawn from different cultures with systematically different beliefs, or idiosyncratic, reflecting differences in individual familiarity with elements in their own culture (e.g., some people know the names of more dog breeds than others). Second, within a group that has been determined to constitute a single culture, the method provides a way of measuring how much of the culture each individual knows – "cultural competence". Third, for each culture represented in a dataset, the method tries to ascertain the culturally correct answer to every question we have put to our informants.

The method has been found useful in a wide variety of settings across multiple disciplines. For example, consensus analysis has been used to investigate cultural diversity within social movements (Caulkins and Hyatt, 1999), Celtic cultures (Caulkins, 1999) and Welsh-American populations (Caulkins, Offer-Westort, and Trosset, 2005). Scholars have used the method to investigate public health issues such as

[1] Name of an album released by Elvis Presley

perceptions of diseases among Guatemalans (Romney et al., 1986), postpartum hemorrhage among Bengalis (Hruschka, Sibley, Kalim, and Edmonds, 2008), pain among Anglo-American and Chinese (Moore, Brodsgaard, Miller, Mao and Dworkin, 1997) and AIDS, diabetes, the common cold, empacho, and mal de ojo among multiple ethnic groups (Weller and Baer, 2001). Others have used consensus analysis to distinguish experts from novices in domains such as fish (Boster and Johnson, 1989; Miller, Bartram, Marks and Brewer, 2004), pollution and food safety (Johnson and Griffith, 1996), ecological knowledge (Shafto and Coley, 2003), medical beliefs (De Munck et al, 1998) and alphabet systems (Jameson and Romney, 1990).

Consensus theory, unlike most work in the social sciences, is developed quite formally. It is based on an underlying abstract model, which is then used as a basis for deriving implications which constitute the theory itself. This then makes certain methods possible. We use this structure to organize our discussion.

The Abstract Model

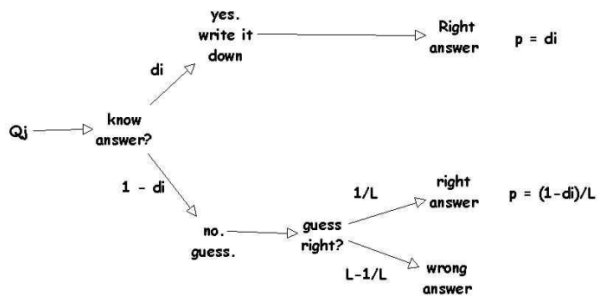
The simplest way to explain the consensus model is to start with a simplified, artificial context. For example, consider a multiple choice exam given in an introductory anthropology class. A possible question might be "The author of the book *Tristes Tropiques* was _____", followed by 5 choices.

Each student's set of responses consists of a vector of numbers ranging between 1 and L, where L is the number of alternatives in each multiple choice question (5 in our example). If we call a given student's vector **z**, then her response to question J can be denoted z_j. We can also arrange the whole collection of student response vectors as a response matrix, in which the rows are students and the columns are questions. If we call this matrix Z, then cell z_{ij} gives the response of student I to question J. If N gives the number of students taking the test, and Q gives the number of questions on the test, the matrix Z has N rows and Q columns.

The instructor's answers can also be arranged as a vector **t** that looks much the same as each student's vector. When we grade the test, we effectively compare each student's row in the response matrix Z with the instructor's vector t. If the two vectors are the same across all questions, the student gets a perfect score. If the vectors are quite different, the student has missed many questions and gets a low score. Thus, a student's exam score or *accuracy* is actually a measure of similarity between the student's and instructor's vectors. The measure of similarity is the simple match coefficient -- the number of times that $z_j = t_j$, where j ranges across all questions, divided by the number of questions, q.

Now, to work out the relationship between knowledge, responses on the test, and agreement among students, we build a simple model of how a student goes about responding to a given question on the test. Based on classical test theory, the model is as follows (see Figure 1). When a student I confronts a given question, he either knows it or he doesn't. We denote the (unknown) probability of knowing it as d_i and the probability of not knowing it is then $1-d_i$. The quantity d_i is the student's competence in the domain – it is the proportion of all possible questions that could be asked in a given topic that he knows the answer to. In short, d_i is a measure of the amount of knowledge possessed by student I.

Figure 1



Returning to the model, if the student knows the answer, he writes it down and gets it right (this can be made more complicated later to allow for people getting questions wrong that they in fact knew the answer to). If she doesn't know the answer, which occurs with probability 1- d_i, she guesses. For expository purposes, we will assume she guesses randomly among the L alternatives. However, we can also complicate this model by incorporating certain biases (e.g., adherence to the rule 'if you don't know the answer, choose "c"') or person-varying talent for knocking out some obviously wrong alternatives.

We can see in the model that the overall probability m_i that a given student I gets a randomly chosen question right is equal to $d_i + (1-d_i)/L$, as shown in Equation 1. In other words, the probability of getting a question right is a function of the probability of knowing the answer plus the probability of not knowing but guessing correctly. By rearranging terms, we can also write the probability of knowing the answer (i.e., competence) in terms of the proportion of correct answers. As shown in Equation 2, a person's competence in a given domain can be modeled as the proportion of correct answers they give, but adjusted for chance guessing. This is a well-known model in classical test theory (Novick, 1966).

$$m_i = d_i + (1 - d_i)/L \qquad \text{Equation 1}$$

$$d_i = (Lm_i - 1)/(L-1) \qquad \text{Equation 2}$$

Using this simple model, we can now begin to work out the relationship between agreement and knowledge. We begin by formulating the probability that two students I and J (with knowledge levels d_i and d_j respectively) give the same answer to a given question. As shown in Table 1, there are just four ways this can happen. Case 1 is where both students happen to know the answer and therefore provide the same (correct) answer. This happens with probability $d_i d_j$, which is simply the probability that I knows the answer times the probability that J knows the answer. Case 2 occurs when student I knows the answer (with probability d_i) and student J does not know the answer but happens to guess right (which happens with probability $(1-d_j)/L$). The probability of both of these events occurring jointly is their product, $d_i (1-d_j) /L$. Case 3 is the mirror image of Case 2. In this case, student J knows the answer and student I does not, but guesses right. This occurs with probability $d_j (1-d_i) /L$. Finally Case 4 is where neither student knows the answer and they guess the same thing. Note that it is not necessary that they guess correctly in order to agree. Given that both students are guessing, the probability of agreeing with each other is L times $1/L^2$, which is just 1/L. To see this, consider the probability that both students guess answer "a". Since each is choosing option "a" with probability 1/L, the probability of both choosing "a" at the same time is 1/L times 1/L or $1/L^2$. Since there are L possible choices, the overall probability of guessing the same thing is L times $1/L^2$, which is simply 1/L. Thus, Case 4 occurs with probability (1-di)(1-dj)/L, which is the joint probability that neither student knows the answer and whatever they guess happens to be the same.

Table 1. Exhaustive enumeration of cases leading to agreement between student I and student J

Case	**Probability**
1. Both I and J know the right answer	$p_1 = d_i d_j$
2. Student I knows the right answer, and student J guesses right.	$p_2 = d_i (1-d_j) /L$
3. Student J knows the right answer, and student I guess right	$p_3 = d_j (1-d_i) /L$
4. Neither knows the answer, both guess the same answer	$p_4 = (1-d_i)(1-d_j) /L$

Now, since the four cases are independent, to get the overall probability m_{ij} that students I and J give the same answer to any given question, we simply add up the probabilities across all four cases (Equation 3). By rearranging terms, we can simplify this to Equation 4, which says that the agreement m_{ij} between students I and J is a function of the degree to which each is highly knowledgeable plus some agreement due to chance. This shows that, at least in our simplified multiple-choice exam setting, agreement really is an indicator of knowledge.

$$m_{ij} = d_i d_j + d_i(1-d_j)/L + d_j(1-d_i)/L + (1-d_i)(1-d_j)/L \qquad \text{Equation 3}$$

$$m_{ij} = d_i d_j + (1 - d_i d_j) /L \qquad \text{Equation 4}$$

From a practical point of view, what is particularly interesting about Equation4 is that we can use it to estimate the amount of knowledge each person has, and we don't need an answer key to do it. To see this, we first rearrange the equation so that joint knowledge ($d_i d_j$) is on the left side of the equation and (Lm_{ij} - 1)/(L-1) is on the right. Since L is just a constant, the entire right side is just a linear scaling of m_{ij}, which is the amount of agreement between students I and J. For convenience, we call this adjusted agreement m^*_{ij}.

$$d_i d_j = (Lm_{ij} - 1)/(L-1) = m^*_{ij} \qquad \text{Equation 5}$$

The first important thing to realize is that the matrix M* is observable. In our multiple-choice test, we can count up, for each pair of students, how often they agree with each other, and then adjust this for chance agreement to get M*. So it is only the left side that contains unknowns. The second important thing to realize is that the equation $d_i d_j = m^*_{ij}$ is solvable for the values of d. Indeed, standard algorithms for principal factor analysis will take the observed m^*_{ij} values as input and estimate the values of d, which are the factor loadings. In other words, using factor analysis, we can determine each student's amount of knowledge or competence simply by analyzing the pattern of observed agreements among students.

Obviously, there must be a catch. It can't be that we can discover the truth about, say, physical laws, by taking a poll of physicists' opinions. There must be conditions under which this is true. Indeed, there are conditions, and these are entirely contained in the response model shown in Figure 1. All that has been shown is that if the response model is true, then we can divine knowledge without the benefit of an answer key. When the model is not true, we really can't draw any conclusions about knowledge.

This may seem very restrictive, and indeed it is, but the good news is that the response model does not have to be exact in every detail. It can be shown (Batchelder & Romney, 1988) that Equation 4 holds for any model which embodies the following three assumptions: common truth, conditional independence, and item homogeneity. We discuss each of these in turn.

Common Truth. This is the assumption that there is one and only one right answer for every question. It is implicit in the response model which assumes that for each multiple-choice question, one of the answers is correct and the others are wrong. The model shows that there are two branches of the probability that lead to correct answers: knowledge and guessing right.

Conditional Independence. This is the assumption that students' responses are independent from each other, both across questions and other students, conditional on the answer key. In other words, the students do not cheat from each other, and the answers to certain questions do not change the answers to other questions. This assumption is implicit in the second fork of the response model, where, if a student does not know the answer, she guesses randomly among the available choices. The implication of this assumption is that there is nothing attracting students to the same answer other than both knowing the correct answer.

Item Homogeneity. This assumption specifies that questions are drawn randomly from a universe of possible questions about a specific topic, so that the probability d_i that student I knows the answer to a question is the same for any randomly chosen question. In other words, all questions are on the same topic, about which a given student has a certain level of knowledge. This is implied in the response model by the use of a single parameter d_i to characterize a respondent's probability of knowing the answer. Thus, item homogeneity implies that questions about basketball are not mixed with questions about wine, since people can be expected to have different competencies in the two domains.

Together, these three conditions are sufficient to guarantee that we can recover the amount of knowledge possessed by any student, even without the benefit of an answer key. In addition, we can use this method to determine the culturally correct answer key.

Methodology and Empirical Illustration

The consensus method starts with an n-by-m person-by-question response matrix X, in which cell x_{ij} gives the response of person i to question j. We shall assume to start with that the responses are categorical choices obtained from a multiple-choice test (including a true-false test that just has two choices for each question). We presume every question has exactly L possible responses which are numbered 1 through L. The analysis begins by constructing a person-by-person agreement matrix M in which m_{ij} equals the number of questions for which persons I and J give the exact same answers. We then adjust this for chance guessing to obtain the matrix M*, using the formula in Equation 6, where L is the number of choices in the multiple-choice test.

$$m^*_{ij} = (Lm_{ij} - 1)/(L-1) \qquad \text{Equation 6}$$

This matrix is then subjected to a principal factor analysis using a method such as Comrey's (1962) method which ignores the diagonals of the input matrix. This results in a set of eigenvectors and associated eigenvalues. We sort the eigenvectors in descending order by the eigenvalues. The principal eigenvector (with the largest eigenvalue) contains the factor loadings which, when all three conditions hold, indicate the amount of knowledge each person has. The eigenvalue can be used to assess the extent to which agreements among persons are explained by a single factor, corresponding to the existence of a single answer key. If the largest eigenvalue is quite a bit larger than the next-largest eigenvalue (say, 3 times as large), we consider this as evidence that the common truth and conditional

independence conditions hold, which means we can interpret the factor loadings as competence. In addition, to be consistent with the common truth, all factor loadings should be non-negative or, if negative, negligibly small.

Once each person's competence has been estimated, it is then possible to infer the culturally correct answer to each question by examining the distribution of responses by the competence of each responder. The greater the competence, the more likely their response is the correct one. Effectively, we choose the answer that would maximize the probability of obtaining the pattern of answers we do, assuming the model is perfectly true. The details of this Bayesian inference process are given by Romney, Weller and Batchelder (1986).

We now turn to an empirical illustration using an actual multiple-choice exam administered in a Sociology 101 class at the University of South Carolina. The exam consists of 58 multiple-choice questions with 5 possible responses for each question (L=5). A total of 101 students took the test, yielding a 101x58 response matrix X. The analysis procedure in UCINET (Borgatti, Everett & Freeman, 1992) was used to analyze the data. After forming the chance-corrected agreement matrix, the program runs Comrey's (1962) minimum residual factor analysis algorithm to obtain the competence estimates **d**. As shown in Table 2, the pattern of eigenvalues is highly consistent with the assumption of a single set of right answers: this is seen in the largest eigenvalue being many times larger than the next largest. This is what we expect in a classroom setting with a single teacher. If the exam had been given to two different classes of Sociology 101 with different professors, we might have seen a result in which the two largest eigenvalues were of similar size, reflecting two different right answer keys in operation and corresponding to the perspectives of different teachers.

Table 2: Eigenvalues for Sociology 101 exam.

Largest eigenvalue:	47.903
2nd largest eigenvalue:	2.839
Ratio of largest to next:	16.871

Since the first eigenvector turns out to be sufficiently dominant, we can go ahead and interpret the factor loadings (the values of the eigenvector) as estimates of each student's amount of knowledge. Given that in this case we know the correct answers (by definition, they are the teacher's answers) we can compare the model's ranking of students with the instructor's grades for each student. The comparison is shown in Table 3. The Pearson correlation between the model's competence scores and the instructor's letter grades is 0.937.[2] Thus, if we regard the instructor's answer key as the gold standard, the model has recovered those scores extremely well, especially considering the letter grade system groups a range of scores together into a single value.

Table 3: Comparison of letter grades and competence scores for 101 students

Letter Grade	Comp.	Letter Grade	Comp.	Letter Grade	Comp.	Letter Grade	Comp.	Letter Grade	Comp.
A	0.89	B	0.79	C+	0.79	D+	0.70	D	0.59

[2] The original numeric scores used to assign letter grades are no longer available. The correlation was obtained by converting letter scores to numeric as follows: A = 4.00, B+ = 3.50, B = 3.00, C+ = 2.5, C = 2.0, D+ = 1.5, D = 1.0, F+ = 0.5, F = 0. Minuses were not used at this university.

A	0.89	B	0.79	C+	0.77	D+	0.69	D	0.59
A	0.87	B	0.78	C+	0.77	D+	0.69	D	0.57
A	0.86	B	0.78	C+	0.77	D+	0.68	D	0.56
A	0.85	B	0.77	C+	0.73	D+	0.66	D	0.52
A	0.85	B	0.77	C+	0.73	D+	0.65	D	0.52
B+	0.86	B	0.77	C+	0.73	D+	0.65	D	0.50
B+	0.86	B	0.77	C+	0.73	D+	0.65	D	0.47
B+	0.85	B	0.76	C+	0.72	D+	0.65	D	0.46
B+	0.84	B	0.74	C+	0.72	D+	0.64	D	0.46
B+	0.82	B	0.73	C+	0.70	D+	0.64	F	0.41
B+	0.81	B	0.72	C+	0.64	D+	0.63	F	0.38
B+	0.80	B	0.82	C	0.75	D+	0.63	F	0.38
B+	0.80	B	0.80	C	0.73	D+	0.63	F	0.35
B+	0.79			C	0.71	D+	0.63		
B+	0.77			C	0.70	D+	0.63		
B+	0.77			C	0.68	D+	0.63		
				C	0.68	D+	0.63		
				C	0.68	D+	0.62		
				C	0.66	D+	0.62		
				C	0.64	D+	0.62		
				C	0.62	D+	0.62		
				C	0.60	D+	0.61		
						D+	0.60		
						D+	0.60		
						D+	0.60		
						D+	0.59		
						D+	0.58		
						D+	0.57		
						D+	0.57		
						D+	0.56		
						D+	0.56		
						D+	0.54		

The Model in Anthropological Context

In this section we translate the abstract consensus model into a more general anthropological context. We begin by considering the meaning of the three assumptions.

The first assumption, common truth, is fundamentally a statement about the kinds of informant variability that may exist. From the model's point of view, there are basically two sources of variability in informant responses: culture and competence. Cultural variability refers to variability in responses due to belonging to different cultures, which have systematically different ways of looking at the world – in effect, having different answer keys. Competence variability refers to differences in responses within a

given culture, which is to say among people for whom the same answer key applies. In any culture, some people simply know more of the cultural truth than others do. For example, some people know the names of many different kinds of trees, while others know only a few, even when they belong to the same culture.

The common truth assumption essentially states that, if you want to be able to measure the competence or cultural literacy of informants, they must be drawn from the same culture. Otherwise, the differences in responses you observe could be due to cultural differences.

The conditional independence assumption specifies that the only systematic force leading people to give the same answers is the cultural truth. In other words, when people are mistaken about some cultural "fact", their mistakes are not correlated with each other. They might agree on the same wrong answer, but it is just by chance. If this assumption were not true, it would lead to the model overestimating the knowledge of informants, because they would be agreeing at high rates, which would be interpreted as having high levels of knowledge. A key thing to consider here in applying the consensus model is whether the questions are of a factual, recall sort of nature, or whether there are heuristics for figuring out the answer. For example, suppose informants don't actually know whether a given leaf is a maple or not. They may have a rule available to them that says "maple leaves have 5 lobes", which would lead them to answer "maple" for many different kinds of leaves, sometimes correctly, sometimes not. Heuristics of this type provide an extra degree of association between responses that are not consistent with the conditional independence assumption.

It must be emphasized that the function of the three assumptions is to guarantee that the factor loadings of the agreement matrix can be interpreted as estimates of informant competence in their culture. When the assumptions don't hold, the factor loadings cannot be seen as competence because other factors may be accounting for inter-informant variability. However, estimating competence is not the only important output of the model. In many cases, the key research question is whether the informants belong to a single culture or not, and this can be diagnosed by examining the pattern of eigenvalues. The assumptions of the model do not need to hold in order to make this diagnosis. We now turn to powerful examples of consensus analysis in practice made possible by the loosening of these assumptions.

Consensus Analysis in Practice

Consensus analysis can be used to analyze multiple types of data including true-false, yes-no, multiple choice, and even open-ended, fill-in-the-blank questions (assuming that the assumptions of the model hold). In addition, ways have been proposed to work with ordinal and interval scale data (e.g., Chavez, 1995; Romney, Batchelder, and Weller, 1987) and social network analysis data (Batchelder, Kumbasar, and Boyd, 1997; Kumbasar, 1996).

Consider the archival data from NCAA American football shown in Table 4[3]. Before the start of each season, sports journalists from competing magazines assess the quality of all football teams and subjectively rank the strongest 25 teams (1 is considered the best). The decisions are made before the teams have played any games and are often influenced by factors such as number of returning players, quality of incoming recruits, strength of schedule, coaching ability, etc. To use consensus analysis with

[3] These data are publicly available at http://football.stassen.com/

these data, we recoded the rankings so that the teams were sorted into six tiers[4] in descending quality. Table 4 presents the team quality ratings from 10 different judges:

Table 4: NCAA Football Ratings

	Athlon	Street & Smith	Sporting News	Phil Steele	College Football News	Lindy	ATS	Game Plan	CPA	CNN/SI
Alabama	5	5	6	6	5	5	6	3	6	6
Arizona State	6	3	6	6	6	5	6	5	6	5
Arkansas	6	6	6	5	6	6	4	6	5	6
Auburn	1	2	1	3	1	2	2	1	3	2
Colorado	6	6	6	6	6	5	5	6	6	6
CSU	6	5	4	6	6	4	6	6	6	6
Florida	4	6	6	6	5	6	4	6	4	6
Florida State	3	4	3	3	5	4	6	2	2	3
Fresno State	6	6	5	6	6	6	6	6	6	6
Georgia	2	3	3	4	3	3	5	2	2	1
Kansas State	1	2	2	3	2	1	3	1	1	2
LSU	4	4	3	3	3	3	5	4	2	4
Maryland	4	3	4	4	2	2	4	4	4	4
Miami-Florida	1	1	1	2	1	1	1	2	1	1
Michigan	2	1	1	1	4	1	1	2	2	3
Minnesota	6	5	6	6	6	6	6	6	6	6
Mississippi	6	6	6	5	6	6	6	6	6	6
Missouri	5	6	6	6	5	6	6	6	6	6
N. C. State	3	4	5	4	3	4	5	6	4	2
Nebraska	6	6	6	3	6	6	6	3	5	6
Notre Dame	5	5	4	4	4	3	3	4	4	4
Ohio State	1	1	1	1	1	1	1	1	1	1
Oklahoma	1	1	1	1	1	1	1	1	1	1
Oklahoma State	6	5	5	6	6	6	4	4	6	5
Oregon	6	6	6	6	6	6	6	6	6	5
Oregon State	4	6	6	6	4	6	6	6	6	6
Penn State	6	6	6	5	6	6	4	6	5	5
Pittsburgh	2	2	4	1	2	3	2	4	3	3
Purdue	4	6	3	4	5	4	2	3	5	6
Southern Cal	2	3	2	2	2	2	6	3	2	2
Tennessee	5	3	4	2	3	2	2	5	3	4
Texas	2	1	2	1	1	2	1	1	1	1
Texas A&M	6	6	6	6	6	6	6	5	6	5
TCU	6	6	5	5	6	5	6	5	6	6
UCLA	6	6	6	6	4	6	6	6	6	6
Virginia	5	4	2	6	4	5	3	6	4	3
Virginia Tech	3	2	3	2	2	3	2	2	3	2
Washington	3	2	2	2	6	4	3	3	3	3
West Virginia	6	6	6	6	6	6	5	6	6	6
Wisconsin	3	4	5	5	3	6	3	6	5	4

Let's assume that we know very little about NCAA American football and want to learn more about this domain. For example, which teams are considered the best, and which judges are the most informative? From looking at the data, we note that there are clear differences of opinion: The Sporting News places Auburn in the top tier, but Phil Steele places them in the 3rd tier; ATS places Michigan in the top tier, but

[4] We recoded the raw rankings into broader categories to partially address the lack of independence among these data. Our recoding scheme was as follows: Teams ranked 1-5=1st tier, 6-10=2nd tier, 11-15=3rd tier, 16-20=4th tier, 21-25=5th tier, others=6th tier. Without the recoding, a judge's choice of a particular rank for one team would (usually) preclude giving that rank to another team, violating the conditional independence assumption of the model.

College Football News has them in the 4th tier; Athlon and The Sporting News place Southern California in the 2nd tier, but ATS places them in the 6th tier. Without knowing much about the domain, it is difficult to determine the culturally correct ranking of each team, the expertise of each judge, and whether the judges belong to the same culture.

To address the variability in perceived team quality, we might use the modal tier placement of each team as an indicator of quality. For example, all 10 judges place Oklahoma and Ohio State in the top tier. We might feel confident that Oklahoma and Ohio State are considered among the best teams in the country. However, there are situations in which there are multiple modal values: four judges place LSU in the 3rd tier, and four place them in the 4th tier. Therefore, the modal approach can fall short because it cannot effectively address these discrepancies.

Another approach might be to average each team's tier placement from the 10 judges to calculate an aggregate quality score for each team. However, suppose one of the judges knows very little about NCAA football. The averaging method would weight this person's answers equally with all the others, creating an average value that is quite different from the majority. In addition, the averaging approach cannot distinguish respondents who do not know about the domain due to lack of knowledge from those who know a lot but are from a different culture. In the case of just a few outliers we could turn to more robust measures of central tendency, such as medians and trimmed means. But in the end, all of these methods have no way of discounting the data from judges who really don't know what they are talking about.

A third approach is to use consensus analysis. As discussed earlier, this method can be used to help us identify both the culturally correct quality of each team and the expertise of each judge. This method distinguishes between having low competence in a domain and having a different culture. To analyze these data we transpose the matrix in Table 4 to create a judge-by-team matrix in which x_{ij} equals the rating that judge *i* gave to team *j*. Table 5 gives output from running consensus analysis on these data using UCINET.

Table 5

Largest eigenvalue	3.82
Second largest eigenvalue	0.36
Ratio of largest to next largest	10.49

Table 5 indicates a good fit with the cultural consensus model in that the ratio of the largest eigenvalue to the second largest is large.[5] As discussed above, the pattern of eigenvalues is highly consistent with the assumption of a single set of right answers and conditional independence. Had there been a group of judges who provided rankings of academic prestige and not football quality, the ratio of the first to second eigenvalue would likely be less than 3 to 1, and the program would notify us that the data do not fit the cultural consensus model.

The high ratio of the first to second eigenvalue allows us to interpret the factor loadings as the competence scores of each judge in the domain of NCAA football. Table 6 indicates that CPA has the highest cultural competence, while Game Plan and ATS have the least cultural competence in this domain. If we were interested in learning more about the quality of NCAA football teams, journalists at

[5] A typical rule of thumb is that the first eigenvalue should be at least 3 times larger than the second.

CPA might be a good source of additional information. If there had been judges that provided responses very different from the cultural norm, they would have negative competence scores in this table.

Table 6

Judge	Competence
CPA	0.68
Sporting News	0.67
Athlon	0.67
CNN/SI	0.64
Lindy	0.62
Phil Steele	0.61
Street & Smith	0.61
College Football News	0.60
Game Plan	0.54
ATS	0.53

The consensus model also identifies the culturally correct quality of each team, as displayed in Table 7. The output provides an "answer" for every team and allows us to easily differentiate the top- and bottom-tiered teams. Note that the answer key identifies Southern Cal as a 2nd tier team despite ATS considering them a 6th tier team (we also note that ATS had the lowest competence score).

Table 7

	Answer Key
Oklahoma	1
Ohio State	1
Miami-Florida	1
Texas	1
Auburn	1
Michigan	1
Kansas State	2
Virginia Tech	2
Pittsburgh	2
Southern Cal	2
Georgia	3
Washington	3
Tennessee	3
LSU	3
Florida State	3
N. C. State	4
Maryland	4
Notre Dame	4

Virginia	4
Purdue	4
Wisconsin	5
Oklahoma State	6
Alabama	6
Florida	6
Arizona State	6
Nebraska	6
Oregon State	6
TCU	6
Arkansas	6
CSU	6
Penn State	6
Colorado	6
Texas A&M	6
Missouri	6
UCLA	6
Fresno State	6
West Virginia	6
Mississippi	6
Oregon	6
Minnesota	6
Southern Miss	6

This dataset also allows us to investigate the predictive accuracy of each judge by comparing the preseason predictions with actual outcomes. At the end of each season the Associated Press ranks teams based on their on-the-field outcomes (i.e., win-loss record, margin of victory, etc). This outcome measure allows us to compare the accuracy of the perceived quality scores provided by the 10 judges and the "culturally true" quality scores derived from consensus analysis. To do this we recoded the final AP rankings using the same method used to recode the preseason predictions and correlated the results. See Table 8.

Table 8

Judge	**Correlation with Actual Outcomes**
Consensus Analysis Answer Key	0.64
CPA	0.58
Lindy	0.52
Phil Steele	0.50
Sporting News	0.47
Game Plan	0.46

Athlon	0.44
CNN/SI	0.42
Street& Smith	0.41
College Football News	0.40
ATS	0.23

Findings indicate that the answer key derived from consensus analysis was the most accurate predictor of actual outcomes (r = 0.64). In other words, the consensus truth output was a better predictor of actual performance than any of the 10 judges individually. In summary, this finding provides additional evidence for the "wisdom of crowds" and highlights the utility of using consensus analysis to get at this "wisdom" and identify competent individuals.

We now turn to an application of consensus analysis to social network analysis. Consider the following example taken from a consulting project in which we studied the relationships among 14 executives forming the top management team of a local organization. Part of this project involved interviewing each executive and identifying how he or she perceived the network connections of each of his or her colleagues. Network ties were evaluated on a 1 to 5 scale, where 5 indicated a stronger tie. One of the driving research questions was whether executives who accurately see relationships among others are more effective leaders than those who are less accurate. The data collection process involved asking each executive to provide their perception of the relationship between every pair of executives, including themselves. In other words, each executive reported the strength of his or her relationships with other executives as well as his or her perception of the strength of relationship between every possible pair of executives. The resulting data was a collection of fourteen 14-by-14 matrices of perceptions, one for each executive, a type of data known as cognitive social structure data (Krackhardt, 1987).

We used consensus analysis to identify the consensus view of the network of perceived relationships and determine each individual's competence in perceiving the ties around them.[6] Table 9 shows that the largest eigenvalue is more than three times the second largest, suggesting that all of the executives see the network in a fundamentally similar way. Table 10 gives the competence scores for each person, showing which individuals have the best understanding of the network around them.[7] We can see that Farhill has the highest score, making him a good choice as a key informant for an ethnographic study (Johnson, 1990), and arguably someone who is well positioned to get things done because he understands who is allied with whom. We also note that Butler, Trout and Agachar have significantly lower competence scores, indicating that they have little idea of who is connected to whom. Interestingly, Trout is the CEO of the organization and Butler is 2nd in command. It is possible that their unique job responsibilities separate them from other executives and might influence their ability to accurately view the network of relationships. Nevertheless, their lack of understanding of the relationships around them is a potential source of serious management problems.[8] Agachar's low score

[6] Technically, data of this type potentially violate the second assumption of the model, because each judge's data is a matrix in which all the cells in a given row correspond to the judge's perceptions of the row person's relationships. If the judge thinks the person is an odd duck, it will influence the perceptions of the person's relations to all others. Maher (1987) has shown via simulation that the model is quite robust to violations of the third assumption, but to date no study has examined violations of the second assumption.

[7] All names are pseudonyms.

[8] It is presumably no coincidence that the organization was experiencing some problems, which is why consultants were brought in.

is also interesting because a separate analysis (not shown) shows that he occupies a highly peripheral position in the informal communication network, which could explain why he knows so little about who is connected with whom. As an aside, if we remove these three individuals with low competence scores and rerun the analysis, we find that the ratio of 1st to 2nd eigenvalue increases to 10.72, and competence scores for the remaining executives are virtually unaffected (see Tables 11 and 12).

Table 9. Eigenvalues

Largest eigenvalue	5.92
2nd largest eigenvalue	1.81
Ratio of largest to next	3.27

Table 10. Competence Scores

Informant	Competence
Farhill	0.90
Black	0.78
Mechanic	0.76
Andrews	0.76
Gold	0.76
Godfrey	0.74
Westminister	0.69
King	0.68
Jones	0.67
Long-Rong	0.62
Pyre	0.60
Butler	0.08
Trout	0.06
Agachar	-0.23

Table 11

Largest eigenvalue	5.85
2nd largest eigenvalue	0.55
Ratio of largest to next	10.72

Table 12

Informant	**Competence**
Farhill	0.90
Black	0.79
Mechanic	0.77
Andrews	0.76
Gold	0.76

Godfrey	0.74
Westminister	0.69
King	0.69
Jones	0.68
Long-Rong	0.62
Pyre	0.59

Finally, consensus analysis also produces an inferred "true" network of ties based on all respondent points of view. This "true network" is different from simply taking the majority answer for each dyadic relationship because it takes into account the varying competencies of the judges. It also differs from a network constructed by considering only the responses of the two members of any dyad – an approach Krackhardt (1987) refers to as locally aggregated structures (LAS). Figure 2 displays the inferred "true" network of perceived relationships derived from consensus analysis[9]. Figures 3 and 4 display Farhill's and Agachar's views of the same network. Note that Farhill's view is more similar to the inferred truth than is Agachar's, in keeping with their respective competence scores, which is very high for Farhill and very low for Agachar.

In summary, consensus analysis can be used with cognitive network data to reveal very powerful findings not easily obtained from other analytic approaches. These findings can be used to identify potential leaders and to better understand organizational performance and conflict.

Figure 2: "True" Network of Perceived Executive Relationships Derived from Consensus Analysis (only displaying ties of strength greater than 4)

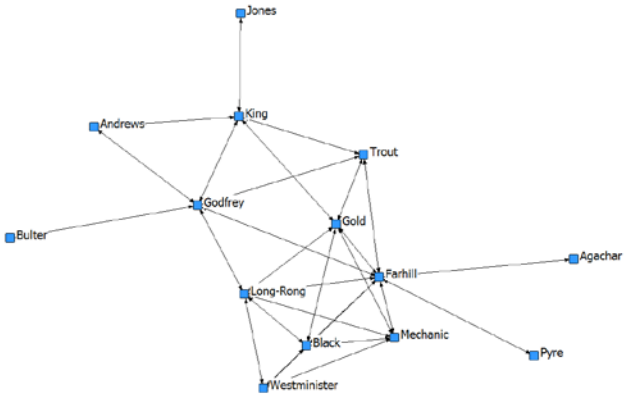


Figure 3: Farhill's View of Executive Relationships (only displaying ties of strength greater than 4)

[9] The networks displayed in Figures 2, 3, and 4 were dichotomized at tie strengths greater than 4.

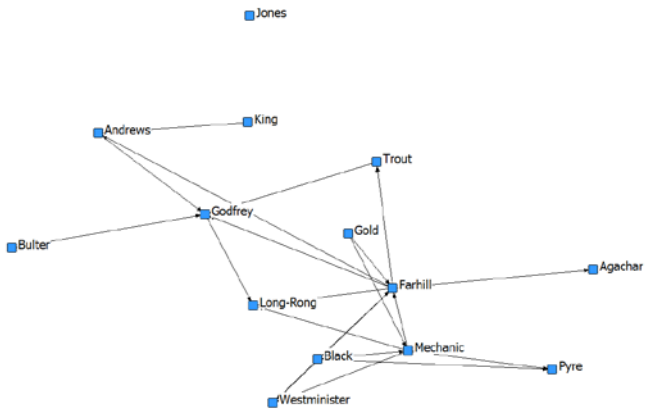
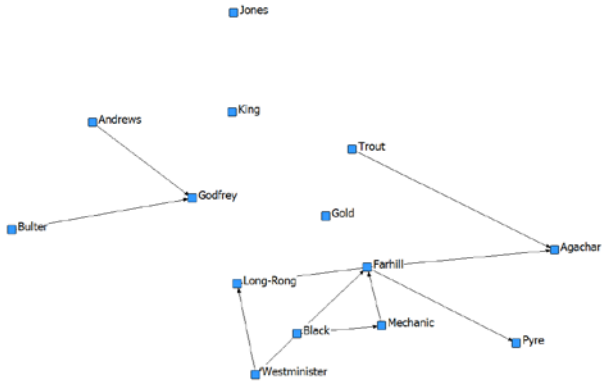


Figure 4: Agachar's View of Executive Relationships (only displaying ties of strength greater than 4)



Conclusion

The consensus model provides both a powerful theoretical perspective and a very useful analysis methodology. A key theoretical contribution is the distinction made between cultural and competence sources of inter-informant variability. Cultural variability is manifested as systematically different patterns of answers across a survey (or multiple answer keys in the language of the model) that apply to different clusters of informants. Competence variability is the remaining variability among informants who are drawn from the same culture, but with perhaps differing access to that culture, leading to some knowing more about some domains than others.

As a methodology, the consensus model can determine whether a given set of informants have consensus, meaning that they share an underlying culture. Given that they do, the model can then provide highly accurate estimates of how much knowledge each person has about the domain. Finally, the model can then use the pattern of who responded in what way to each question to infer the culturally correct right answers, yielding a more accurate understanding of the culture than possessed by any one informant.

Of course, the power provided by the consensus procedure does not come without cost (Borgatti and Carboni, 2007). The method must be coupled with a qualitative first stage in which the researcher elicits the beliefs of a group in a given domain. The result of this ethnographic work must be a set of discrete propositions or statements of fact that can be used as the basis of a multiple-choice questionnaire. This is a time-consuming process that requires some skill. It also requires a domain that lends itself to being atomized in this way. Boster (1987) has gone so far as to refer to consensus analysis as being premised on a "particle theory of culture".

There is also some question about its applicability to cross-cultural research. In some ways, it is ideal for cross-cultural research in that it explicitly distinguishes between systemic, presumably cultural, variability, and competence-based variability. Applied to a dataset in which sufficient members of different cultures all responded to the same questions, it should easily detect the multiple cultures. However, this assumes that the set of questions employed in the survey make sense in each culture. If they don't, separate surveys must be used, which means we cannot make comparisons across the cultures except to determine which if any has more consensus. In addition, if we edit our surveys so that only the questions intelligible in both cultures are present, this may in effect remove the cultural differences, again making it difficult to use consensus analysis.

Despite these very real limitations, consensus analysis is exceptionally useful. It is also an excellent exemplar of model-based science in which the researcher can capture a social or cognitive process as a simple mathematical model from which he or she can then derive some powerful results, and at the same time specify the conditions under which these results are valid.

Finally, it should be clear that consensus analysis is based on a view of culture that is cognitive – culture is in the mind, not in the artifacts or the enacted behaviors. However, it is not too much of a stretch to apply the model to behavioral data, on the assumption that behavior follows cognition. For example, we might make a list of health-related behaviors and record, for each informant, which they do and which they don't. This is similar to a true/false test of health beliefs and the results can be interpreted similarly.

As a final note, the consensus model provides a path of reconciliation between scientific and post-scientific epistemologies. The model utilizes a mathematical approach that seems clearly based on a view of knowledge as objective truth, but ends up providing a way to identify and assess culturally relative knowledge.

References

Batchelder, W.H., Kumbasar, E., and Boyd, J.P. 1997. Consensus analysis of three-way social network data. *Journal of Mathematical Sociology*. Vol. 22: 29-58.

Batchelder, W.H., and Romney. A.K. 1988. Test theory without an answer key. *Psychometrika*. Vol. 53 (1): 71—92.

Bernard, R. 2006. *Research Methods in Anthropology, 4th edition*. Oxford, UK: Altamira Press.

Borgatti, S. P. 1992. *ANTHROPAC 4.983*. Natick, MA: Analytic Technologies.

Borgatti, S.P, and Carboni, I. 2007. Measuring individual knowledge in organizations. *Organizational Research Methods.* Vol. 10(3): 449-462.

Borgatti, S.P., Everett, M.G. and Freeman, L.C. 2002. UCINET for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies.

Boster, J.S. 1987. Introduction. In Intracultural Variation. James S. Boster (ed.) Special Issue of the *American Behavioral Scientist.* Vol. 31(2):150-162.

Boster, J.S. and Johnson J.C. 1989. Form or function: A comparison of expert and novice judgments of similarity among fish. *American Anthropologist*. Vol. 91:866–889.

Caulkins, D.D. 2001. Consensus, clines, and edges in Celtic culture. *Cross-Cultural Research*. Vol. 35(2): 109-126.

Caulkins, D.D, Offer-Westort, M., and Trosset, C. 2005. Perceiving ethnic differences: the use of consensus analysis and personhood in Welsh-American populations. *Mathematical Anthropology and Cultural Theory: An International Journal*.

Caulkins, D.D. and Hyatt, S.B. 1999. Using consensus analysis to measure cultural diversity in organizations and social movements. *Field Methods*, Vol. 11(1): 5-26.

Chavez, L.R., Hubbel, F.A., McMullin, J.M., Martinez, R.G., and Mishra, S.I. 1995. Structure and meaning in models of breast and cervical cancer risk factors: A comparison of perceptions among Latinas, Anglo women, and physicians. *Medical Anthropology Quarterly*. Vol. 9(1): 40-74.

Comrey, A.L., 1962. The minimum residual method of factor analysis. *Psychological Reports.* Vol. 11, pp. 15–18.

D'Andrade, R. G. (1987) A folk model of the mind. In D. Holland and N. Quinn (Eds.) *Cultural Models in Language and Thought*. Cambridge: Cambridge University Press.

De Munck, V., de Alba, G., Guadarrama, V. and Garro, T. 1998. Consensus Analysis: High Blood Pressure in a Mexican Barrio. In *Using Methods in the Field: A Practical Introduction and Casebook*. V. de Munck and E.J. Sobo, Eds (pp.197-211). Walnut Creek: Altamira Press.

Hruschka, D.J., Sibley, L.M., Kalim, N., Edmonds, J.K. 2008. When there is more than one answer key: cultural theories of postpartum hemorrhage in Matlab, Bangladesh. *Field Methods*. Vol. 20(4): 315-337.

Jameson K. and Romney, A.K. 1990. Consensus on semiotic models of alphabetic systems. *Journal of Quantitative Anthropology.* Vol. 2:289–303.

Johnson, J.C. 1990. *Selecting Ethnographic Informants*. Newbury Park: Sage Publications

Johnson, J.C. and Griffith, D.C. 1996. Pollution, food safety, and the distribution of knowledge. *Human Ecology.* Vol. 24(1):87–108.

Krackhardt, D. 1987. Cognitive social structures. *Social Networks.* Vol. 9: 109-134.

Kumbasar, E. 1996. Methods for analyzing three-way cognitive network data. *Journal of Quantitative Anthropology.* Vol. 6**:** 15–34.

Lounsbury, F. 1964. "The structural analysis of kinship semantics." In H.G. Lunt (Ed.) *Proceedings of the ninth international congress of linguists.* The Hague: Mouton.

Maher, K.M. 1987. A multiple choice model for aggregating group knowledge and estimating individual competencies. Doctoral dissertation, University of California, Irvine, 261 pages; AAT 8724745.

Miller, M.L., Kaneko, J., Bartram, P., Marks, J., and Brewer, D.D. 2004. Cultural consensus analysis and environmental anthropology: yellowfin tuna fishery management in Hawaii. *Cross-Cultural Research*. Vol. 38(3): 289-314.

Moore, R., Brodsgaard, I., Miller, M., Mao T., and Dworkin, S. 1997. Consensus analysis: reliability, validity, and informant accuracy in use of American and Mandarin Chinese pain descriptors. *Annals of Behavioral Medicine.* Vol. 19(3): 295-300.

Novick, M.R. 1966. The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*. Vol. 3(1):1-18.

Romney, A.K., Batchelder, W.G., and Weller, S.C. 1987. Recent applications of cultural consensus. *American Behavioral Scientist*. Vol. 31(2): 163-177.

Romney, A. K., Weller, S. and Batchelder, W.H. 1986. Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist* 88(2): 313–38.

Shafto, P., and Coley, J.D., 2003. Development of categorization and reasoning in the natural world: novices to experts, naïve similarity to ecological knowledge. *Journal of Experimental Psychology*. Vol. 29(4): 641-649.

Spradley, J. 1979. *The Ethnographic Interview.* NY: Holt, Rinehart & Winston.

Surowiecki, J. 2004. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations.* New York: Doubleday.

Weller, S.C. and Baer, R.D. 2001. Intra- and intercultural variation in the definition of five illnesses: AIDS, diabetes, the common cold, empacho, and mal de ojo. *Cross-Cultural Research.* Vol. 35(2):201–226.

Weller, S.C., and Romney, A.K. 1988. *Systematic Data Collection*. Newbury Park: Sage Publications.