Field Methods (formerly *Cultural Anthropology Methods*) publishes articles on the methods used by researchers in all of the social and behavioral sciences for the collection, management, analysis, and visualization of data about human thought and human behavior in the natural world. Articles should focus on innovations and issues in the methods used rather than on the reporting of research or theoretical/epistemological questions about research. **Field Methods** also publishes reviews of books and software.

Manuscripts should be submitted for review by e-mail to the editor, H. Russell Bernard, at ufruss@ufl.edu.

Field Methods (ISSN 1525-822X) (J389) is published quarterly—in February, May, August, and November—by Sage Publications, Thousand Oaks, CA 91320.

Periodicals postage paid at Thousand Oaks, California, and at additional mailing offices.

Subscription Information: All subscription inquiries, orders, back issues, claims, and renewals should be addressed to Sage Publications, 2455 Teller Road, Thousand Oaks, CA 91320; telephone: (800) 818-SAGE (7243) and (805)499-9774; fax: (805)499-0871; e-mail: journals@sagepub.com; http://www.sagepublications.com. **Subscription Price:** Institutions: $690; Individuals: $131. For all customers outside the Americas and Asia, please visit http://www.sagepublications.com/journalscs for ordering instructions. **Claims:** Claims for undelivered copies must be made no later than six months following month of publication. The publisher will supply missing copies when losses have been sustained in transit and when the reserve stock will permit.

Indexing: This journal is indexed by Anthropological Literature, Family Studies Database, International Bibliography of the Social Sciences, Linguistic Abstracts, Psychological Abstracts, PsycINFO, PsycLIT, SRM Database of Social Research Methodology, and Anthropological Index Online (http://lucy.ukc.ac.uk/AIO.html), and is available on microfilm from ProQuest, Ann Arbor, Michigan.

Advertising and Reprints: Current advertising rates and specifications may be obtained by contacting the advertising coordinator in the Thousand Oaks office at (805) 410-7160 or by sending an e-mail to advertising@sagepub.com. To order reprints, please call (805) 410-7763 or e-mail reprint@sagepub.com.

Change of Address: Six weeks' advance notice must be given when notifying of change of address. Please send old address label along with the new address to ensure proper identification. Please specify name of journal. POSTMASTER: Send address changes to Field Methods, c/o 2455 Teller Road, Thousand Oaks, CA 91320.

Printed on acid-free paper

Cultural Consensus Theory: Applications and Frequently Asked Questions

SUSAN C. WELLER
University of Texas Medical Branch

In the ethnographic context, where answers to questions are unknown, consensus theory estimates the culturally appropriate or correct answers to the questions and individual differences in cultural knowledge. The cultural consensus model is a formal model of the process for asking and answering questions and is limited to categorical response data. An informal version of the model is available as a set of analytic procedures and obtains similar information with fewer assumptions. This article describes the assumptions, appropriate interview materials, and analytic procedures for carrying out a consensus analysis. Finally, issues that sometimes arise during the application of a consensus analysis are discussed.

Keywords: *cultural consensus model; measurement of beliefs; analytical methods; reliability of informants; Q analysis*

Culture is the set of learned and shared beliefs and behaviors, and cultural beliefs are the normative beliefs of a group. Cultural consensus theory (CCT) is a collection of analytical techniques and models that can be used to estimate cultural beliefs and the degree to which individuals know or report those beliefs. CCT estimates the culturally correct answers to a series of questions (group beliefs) and simultaneously estimates each respondent's knowledge or degree of sharing of the answers.

Group beliefs can be estimated from responses to a series of related questions. The simplest way is to aggregate responses and use the majority responses (for categorical type or qualitative responses) or the average responses (for ranked or quantitative responses) to estimate the answers. Agreement between the responses of individuals and the aggregate responses of the group can be used to estimate how well each individual corresponds to the group.

This project was partially funded by the National Science Foundation (SBR-9727322). I would like to acknowledge colleagues and reviewers for comments that I think have improved the paper with the goal of making the methods more understandable.

Field Methods, Vol. 19, No. 4, November 2007 339–368
DOI: 10.1177/1525822X07303502

CCT builds on these basic analytic approaches. First, estimates of individual knowledge or competency can be estimated from the agreement between people. Then, the culturally correct answers are estimated by weighting the responses of each person by their competency and aggregating responses across people. This article describes the purpose and assumptions underlying consensus theory, the different approaches to estimating the culturally correct answers and individual competencies, and the various issues that may arise when using CCT.

PURPOSE AND ASSUMPTIONS OF CULTURAL CONSENSUS THEORY

When the culturally correct answers are unknown, as in the ethnographic context, the purpose of CCT is to estimate the culturally correct answers and the cultural knowledge or accuracy of informants. To use CCT, informants must be asked a series of questions all on the same topic. Responses to questions are not corrected, recoded, transformed, or reflected as they are with knowledge tests and attitudinal scales because the purpose is to use the original responses to estimate culturally correct answers. Informant reliability or competency can be estimated from the pattern of agreement between individuals; the observed agreement is a function of shared knowledge. Answers are estimated by weighting individual responses by their competency and aggregating those responses across individuals.

To use CCT, at least three assumptions must be met. First, each informant should provide answers independently of all other informants. This means that answers should be provided by individuals and not groups and without consultation with others. The consensus methods are not appropriate for group interviews. Second, the questions should all be on a single topic and at the same level of difficulty. This assumption concerns the homogeneity of items and means that items should represent only one topic or domain of knowledge and that competency should be consistent across items so that if someone is knowledgeable on one subset of questions, they should also be knowledgeable on another subset of questions. Third, CCT is applicable only if there is a single set of answers to the questions. Essentially, this means that there must be a high level of consistency (agreement) in responses among informants. An aggregation of responses is not valid unless there is reasonable consistency in the underlying data. An initial step in applying consensus theory is to check whether there is a high degree of agreement among the informants (e.g., to verify that there is only one response pattern present). Note that a consensus analysis does not create consensus; it only assesses the degree of agreement that is present.

There are two general approaches to consensus theory. The formal cultural consensus model (CCM; Romney, Weller, and Batchelder 1986) is a process model in the sense that it models the process of asking and answering questions. The formal model only accommodates responses to open-ended questions (with a single word or short phrase response for each question) and multiple-choice type questions (including those with dichotomous true-false or yes-no responses). The model assumes that there is no response bias in informants' answers. An informal version of the model (Romney, Batchelder, and Weller 1987) is actually a set of analytic procedures that can be used to estimate the culturally correct answers and informant correspondence to the group answers without some of the restrictions of the formal model. The informal version of the model can accommodate ordinal, interval, and ratio-scaled responses (where numerical estimates are provided for each item). Both approaches provide estimates of the culturally correct answers and estimates of individual differences in the accuracy of reported information.

CULTURAL CONSENSUS THEORY AND RELATED ANALYTIC APPROACHES

Simple Aggregations

Given that a series of related questions has been asked, culturally correct answers to the questions can be estimated with a simple aggregation of responses. In spite of some degree of heterogeneity in beliefs, the most frequently held items of knowledge and belief (the majority or modal items) can be considered the culture of a group (D'Andrade 1987). Using a simple majority or averaging responses across respondents for each question provides a reasonable estimate of answers that is easy to understand and statistically sound. Such an aggregation of responses provides a best estimate of the central tendency for a single variable or, in this case, each question. Such aggregations of responses are intuitive and have long been used to estimate answers to questions.

The only difficulty with this method arises when we try to differentiate strong beliefs from those that might indicate no cultural preference or the presence of multiple beliefs. When response patterns show very strong agreement—for example, 99% say "true" and 1% say "false" to a particular question—it is easy to discern the cultural preference. When responses instead are 51% true and 49% false, there is technically a majority but one that does not really indicate a strong preference for either true or false. In this case, a statistical test can be used to identify the items with a strong majority, for example, where responses are significantly different from a 50–50 split. A

binomial test (Siegel 1956) can be used to identify questions that indicate a significant deviation from chance; items significantly different from 50:50 can be classified as true or false. The two-choice response case can be generalized to multiple response choices and tested against a $1 \times k$ chi-square goodness-of-fit test (Siegel 1956). Simple statistical tests, such as the binomial or chi-square, can be used to distinguish a cultural preference from chance.

Reliability Analysis on Respondents

A limitation of evaluating questions one at a time is that information on the pattern of agreement across the entire series of questions is not used, and there is no information on the relative performance or knowledge of individual informants. A reliability analysis on people incorporates agreement and can provide information on the performance of individuals. In general, reliability is the degree to which the same answer can be obtained under similar conditions. In anthropology, reliability can apply to individuals and the degree to which they supply similar responses to similar questions on different occasions (Young and Young 1962). Informants who provide similar answers to the same questions when asked multiple times are considered to be more reliable sources of information. Equally important is validity, the degree to which the estimated answers correlate with the true answers. Information is considered to be more reliable if the same information is provided by multiple informants. When the same information is provided by multiple informants, the information is also more likely to be valid.

An aggregation of responses is the best estimate of the true answers, but accuracy of that aggregation depends on the agreement among the respondents and the number of respondents. The validity of an aggregation of responses as an estimate of the correct answers is given by the Spearman-Brown prophesy formula from psychology (Nunnally 1978, also described in Weller and Romney 1988:72). The Spearman-Brown prophesy formula can be used to articulate the mathematical relationship between the number of people, the agreement among those people, and the reliability (and validity) of an aggregation of their responses. In contrast to classical reliability theory, the focus here is on the agreement among respondents (not items) and the aggregation of responses across respondents, where responses are not scored, recoded, or transformed. As the number of people increases and/or the agreement among people increases, the reliability and validity of their aggregated responses increases. So valid estimates of the answers to questions (answers that correlate highly with the true answers) can be obtained with a small sample of respondents—if there is high agreement in the responses—or from a larger sample of respondents if there is more heterogeneity in their responses.

To understand the relation between agreement, sample size, and validity, consider the following example. If you were walking in a city and trying to find a bus station, you might ask someone for directions. If you were concerned about the accuracy of those directions, you could ask another person the same question. If the second person gave exactly the same response, you might not ask anyone else for directions. However, if the second person's response did not agree with that of the first person, you probably would want to ask more people for directions. When variability is low, only a small sample is needed to converge on an answer; when variability is high, a larger sample is needed to discern the answer.

Over the past century, the effect that aggregation across informants increased the accuracy of estimated answers has been observed empirically. Studies compared individual judgments against the true or known ordering of objects to study accuracy of individuals and aggregations of their responses (Gordon 1926; Stroop 1932). Zajonc (1962) also observed the relation between the agreement among individuals and the validity of their aggregated judgments. With simulated data, Weller (1987) showed how quickly aggregates converge on the true answers by increasing either the number of informants or the agreement among the informants.

A reliability analysis on people's responses to a series of questions can provide an estimate of the culturally correct answers to the questions, an estimate of the reliability and validity of the estimated answers, and an estimate of each individual's accuracy in answering the questions. In a reliability analysis, first, the correct answers are estimated by averaging the responses across people. Second, the reliability of that set of answers is given by the reliability coefficient, sometimes called Chronbach's alpha, calculated from the number of people and the agreement among them:

$$Rel = n\bar{r} \,/\, [1 + (n-1)\bar{r}],$$

where Rel is the reliability coefficient, n is the number of people being combined, and $\bar{r}$ is the average Pearson correlation coefficient between all pairs of individuals. Third, the validity of the estimated answers is given by the square root of the reliability coefficient (Nunnally 1978). Finally, informant accuracy—or how well the answers provided by each individual correspond with those of the rest of the group—is provided by the "item-to-total correlations" (the responses of each individual correlated with the aggregated responses of all other individuals without that particular individual).

A reliability analysis of respondents has been used to study informant accuracy and to estimate beliefs. Romney and Weller (1984) used the reliability

model to predict informants' accuracy when reporting on social interactions. By analyzing reported social interaction data, they used the item-to-total correlations as an index of individual reliability and found that greater individual reliability was correlated with greater reporting accuracy when reported data were compared to observed interactions. In Weller, Romney, and Orr (1986), item-to-total correlations were used as an index of how well the responses of each informant corresponded to the overall group beliefs or norms about child discipline. Adolescents whose responses had lower correlations with the aggregated responses of the group (e.g., those with greater deviance from the group) were more likely to report corporal punishment.

The Cultural Consensus Model

The CCM is also an aggregative technique similar to the analytical methods described above. It is assumed that the investigator does not know the answers to the questions or the competency of the individuals answering the questions. In contrast to a reliability analysis, where the answers are first estimated and then individual correspondence to those answers is calculated, the CCM first estimates individual competencies and then estimates the answers and the confidence in each answer. It is a formal cognitive model (Romney, Weller, Batchelder 1986; Batchelder and Romney 1988), derived from axioms, that describes the processes and parameters involved in answering questions.

The model assumes that an individual is asked a series of questions requiring multiple-choice-type responses. When a question is asked, the participant answers with the correct answer if he or she knows it, but guesses if he or she does not know it. Guessing is assumed to occur without bias; in other words, if a respondent does not know the answer, he or she guesses as if using a coin flip or a roll of dice to answer. The model identifies certain parameters involved in the answering process: the competency of each individual, the number of response categories, and the proportion of culturally correct answers in each response category.

Agreement between any pair of individuals can be shown mathematically to be the result of their individual competencies. Cultural competence is the cultural expertise of each individual with regard to a set of questions. It indicates the proportion of items each person knows. It is not a moral judgment but simply a description of the fact that some people know more about, say, the rules of Major League Baseball and some people know more about gardening.

Cultural competence scores are estimated from the pairwise similarity in responses between all pairs of informants. Because the model only

accommodates categorical-type response data (true-false, multiple-choice, and fill-in-the-blank question formats), agreement between pairs of informants can be calculated with the proportion of identical answers (matches) between individuals. Two people may provide matching answers to a question through knowledge or guessing, so the proportion of matching responses must be corrected for guessing to estimate knowledge. For example, pigeons taking a true-false test by pecking randomly at possible answers would be expected, on average, to get 50% correct by chance even with no knowledge (0% competency) of the test's subject matter. Thus, in a testing situation, the total proportion of correct answers on a true-false test must be corrected for guessing (by subtracting 1/the number of response choices and then dividing by 1 minus 1/the number of response choices) to estimate the underlying knowledge level. In the case of pigeons, 50% correct – 50% correct by chance = 0% competence ([.50 – .50] / [1.0 –.50] = 0.0, or alternatively, 2 times the proportion correct minus 1.0). If someone got 75% of answers correct on a true-false test, their knowledge level would be 50% ([.75 – .50]/ [1.0 – .50] = .75 × 2.0 – 1.0 = 0.50). This type of adjustment is routinely used in scoring standardized tests.

Similarly, the proportion of matching answers from a true-false or yes-no questionnaire between respondents also is corrected for the proportion of items that would match because of guessing to estimate competence. A multiple-choice test score indicating the percentage correct based on questions with three response choices would be corrected for one third to estimate the knowledge level. If someone got 75% of answers correct with three response choices, their knowledge level would be estimated as 63% ([.75 – .33] / [1.0 – .33] = 0.63). And, in a consensus analysis on a series of questions with three possible responses, the proportion of matching responses between individuals is corrected for one-third (.33), to remove the effects of guessing. The match method assumes that the data do not contain any response bias. That is, the method assumes that informants do not respond preferentially with true (or false) when unsure of an answer. When bias is present, agreement is inflated, causing competence to appear higher than it really is (Weller and Mann 1997).

With dichotomous (true-false) responses, similarity between respondents can be calculated with the match coefficient or with the covariance method. The covariance method is not sensitive to the presence of response bias when estimating competence (Batchelder and Romney 1988). The covariance method is, however, sensitive to the proportion of true or yes answers in the set of answers. Although the answers are not known beforehand (and the purpose is to estimate them), careful construction of the set of questions with an attempt to balance positive and negative items can usually keep the proportion of true items somewhere between 30% and 70%. However, both

the covariance and match methods will have inaccuracies in the estimated answers when response bias is present (Weller and Mann 1997). It is not clear what effect the presence of response bias and a highly skewed answer key will have on the resulting competence and answer estimates, although the consensus model may simply not fit data (see below) with such extreme conditions.

The cultural competence scores are found by factoring an agreement matrix containing the corrected match or covariance coefficients. The agreement matrix (with observed agreement between individuals *i* and *j* in cell row *i* and column *j*) can be used to solve for the unknown competence of each individual (the corresponding main diagonal cells row *i* column *i* and row *j* column *j*) when there are three or more individuals. Cultural competence estimates are provided by a principal axis (minimum residual) factoring method that solves for the unknown main diagonal values of the matrix. The competence scores appear as factor loadings on the first factor. To use the agreement between individuals to solve for competence of each individual, the solution should have only a single factor. Thus, a first diagnostic test is to examine the ratio of the first to second eigenvalues to determine the dimensionality of the solution. This indicator helps the researcher determine if the data conform to an assumption of the model, namely, that there is only one set of answers present in the data. When responses are homogeneous across informants, there will be a single pattern of responses across the questions and the eigenvalue ratio will indicate that there is only a single dimension present in the data.

A general rule is that the ratio should be 3 to 1 or greater. If this ratio is approximately this large or larger, then the consensus model may be used to represent the group's responses with a single set of answers. Also, because the competence scores are interpreted as the proportion of answers that each person knows (and does not guess), the scores must range between 0 and 1, inclusive. Values less than 0 or greater than 1 are undefined.

Answers to questions are estimated by weighting the responses of individuals by their competence scores and then combining the responses. This is accomplished by adjusting the prior probabilities by the observed individual competencies and the pattern of responses to arrive at Bayesian posterior probabilities. This means that the likelihood of each possible answer to each question is considered and calculated as the sum of the likelihood that each person gave a correct answer. The posterior probabilities provide a confidence level for each possible response to each question. Usually, the answer with the highest likelihood is the majority or modal response. Sometimes, however, if more knowledgeable informants are the minority, it is possible for the results to indicate the minority response as the best answer to a question.

Applications of the formal model have used open-ended and multiple-choice questions. Boster (1986) used open-ended questions to study community patterns of plant knowledge. He planted a garden of plants and then asked people to name each plant. Dichotomous (yes-no) questions have been used to study beliefs about AIDS (Trotter et al. 1999), the common cold (Baer et al. 1999), diabetes (Weller et al. 1999), asthma (Pachter et al. 2002), and folk illnesses (Weller et al. 1994; Weller et al. 2003; Baer et al. 2004). In each study, a series of questions was asked about the illness, and respondents answered with yes-no responses. In the folk illness study about *empacho* (Weller et al. 1994), the results of the CCM classification were compared to those from a binomial test for each question. The CCM was able to classify many more questions than a simple question-by-question statistical test.

The Informal Cultural Consensus Model: A Factor Analysis of People

There is another version of the CCM that makes fewer assumptions about the data (Romney, Batchelder, and Weller 1987). This model is essentially a factor analysis (principal components analysis) of people. It does not require correcting for guessing, because it is not a model of how questions are asked and answered. Competence scores are not interpreted as the proportion of answers that an individual knows. Rather, the competence scores tell how well the responses of each individual correspond with those of the group. This model is subject oriented in that the analysis focuses on differences between respondents and not variables and is similar to a Q analysis (McKeown and Thomas 1988). The model is distinct from general Q-analytical approaches, however, in that a specific single-factor, single-group model is specified, and the respondents' factor loadings correspond to their correspondence with the shared group beliefs.

The informal CCM is a set of statistical procedures for estimating answers to a series of questions and estimating respondent accuracy for answering those questions. In this regard, the model is similar to a reliability analysis, except that this model provides a weighted answer key. In the reliability model, answers are estimated with an average of the responses across people. Here, an agreement matrix is factored and the competence scores are used to weight the responses of each individual (by multiplication) and then are summed together. This model can accommodate fully ranked, interval, or ratio-scaled response data, and the Pearson correlation coefficient is used to estimate the similarity or agreement between each pair of respondents.

As with the formal consensus model, the agreement matrix is factored with the minimum residual factoring method (without rotation) to solve for

the missing main diagonal elements in the agreement matrix. The preferred method is actually the maximum likelihood factoring method, although it may fail to reach a solution (Romney, Batchelder, and Weller 1987). The competence scores are the first factor loadings, and the answers (a weighted, linear combination of responses) are found on the first set of factor scores. The method assumes that there is only a single factor solution, so the ratio of the first two eigenvalues is used to determine if the solution is unidimensional.

Applications of the informal model have used ranked data to study perceived mortality, causes of cancer, social class, and social support. In the Romney, Batchelder, and Weller (1987) article on the informal version of consensus analysis, an example is shown where students ranked the causes of death from the most to least frequent, and then their answers were compared to mortality statistics. Chavez et al. (1995) had women rank-order potential causes of cancer from most to least likely to understand women's perceptions of the causes of breast and cervical cancer. Magana, Burton, and Ferriera-Pinto (1995) identified a single, shared ordering of occupational prestige from informants' rankings of occupations in terms of their prestige. Dressler, McBalieiro, and Dos Santos (1997) used the informal consensus model to estimate cultural patterns in seeking social support.

An interesting application done before consensus theory was formalized was a study of height by Dawes (1977). Dawes had respondents rate the height of colleagues and then used factor analysis to combine the ratings. The first set of factor scores (like the consensus answer key) correlated extremely well (.98) with measured height. Romney, Batchelder, and Weller's (1987) study of causes of death and Dawes's (1977) study of height provide important evidence about the validity of model estimates.

ISSUES THAT ARISE IN APPLICATIONS

Development of Interview Materials

We turn now to issues that arise in applying consensus theory to real data. The first step in a study of cultural beliefs is the development of appropriate interview materials. Questions may be developed from information obtained from a variety of sources. New items may be generated from open-ended, structured interviews with individuals or small groups. Items also may be taken from existing sources such as scientific publications, archival records, or previous questionnaires and tests. The items should be reasonable indicators of the concept being measured (content validity). Because abstract concepts require more questions to estimate the concept, the goal is to obtain twenty or more questions/items, at the same level of difficulty, on a single topic.

Questionnaire development is similar to that for attitudinal studies in that items are generated from a sample of individuals similar to those whom you wish to study, the items should be reasonable indicators of the concept, and the items should be balanced in terms of their positive and negative aspects. The questions should have enough variation so that there will be variation in the responses. This means that to the extent possible, questions should elicit both positive and negative responses. (For more information on item generation and interview development, see Weller and Romney [1988] and Weller [1998]).

For example, if we wanted to learn about people's beliefs about AIDS, we might begin with open-ended questions eliciting descriptions of AIDS. Interviewers might ask people to describe causes, symptoms, treatments, experiences, and so forth concerning AIDS. Responses from these interviews and other sources would then be used to create a set of questions designed to systematically study beliefs and variation in beliefs about AIDS (see Trotter et al. 1999). True-false questions might be: Is AIDS inherited? Can you get AIDS from a public bathroom? Can you get infected with AIDS when you donate blood to someone else? Are people with AIDS more susceptible to getting other illnesses? Is a cold that will not go away a symptom of AIDS? Do you have diarrhea with AIDS? If you have a positive attitude, can you help cure AIDS? Is a doctor the best person to treat AIDS?

Writing good items is not a trivial task. Questions and statements need to be clear and need to be understood in the same way by all the respondents. Questions can cover various aspects of the topic. For an illness, questions can cover causes, symptoms, treatments, and so on. Balancing positive and negative items is also important. For example, some questions should be reasonably answered as no and some as yes to avoid a response bias set where respondents simply answer yes to all or most of the questions. To do this, some items need to be reversed and/or supplemented. If an item is elicited in open-ended interviewing concerning transmission of AIDS—that you can get it from having sex with a prostitute or by using a contaminated needle—you may want to further explore the notion of contagion by asking if you can get AIDS by being near someone with AIDS or if you can get AIDS by using the utensils of someone infected with AIDS. Similarly, symptoms identified in open-ended interviews can be supplemented with questions about various body systems: Do you have a runny nose and cough? Do you have a fever? Do you have a stomachache? Do you have aches and pains?

So of the ideas expressed in open-ended interviews, some need to be expressed in a negative form (using supplemental negative questions if necessary). Batchelder and Romney (1988) noted that there should be about an equal number of positive and negative items. With careful attention to item content, between 30% and 70% should be reasonably answered yes and between 70% and 30% answered no. This is similar to the need to balance positive and negative items when designing interview materials for the development of attitudinal scales (Nunnally 1978).

Can I do a consensus analysis on responses to only four questions? Although this can be done, it is not advisable. All the methods described above rely on the agreement between people across questions (questions are the unit of analysis). To estimate the agreement between each pair of informants, a greater number of questions provides a more stable (and thus, better) estimate. At least twenty questions are recommended to obtain reasonable estimates.

If the questionnaire is a true-false or a yes-no set of questions, can "I don't know" be a separate response choice? In general, it should not. Respondents should be encouraged to answer every question. However, if there are some missing responses, responses can be "guessed" for the respondent by flipping a coin. Remember that the formal model assumes that people will guess when they do not know an answer. Thus, a missing answer can be imputed by randomly inserting a 0 or a 1. "I don't know" should not be considered as a third response category, because the model has a built-in correction for guessing. In general, I allow up to 10% missing responses per person and impute random answers for those that are missing, although this is a fairly arbitrary rule.

Can the consensus model be used with open-ended questions with responses such as free-recall lists or with narratives? Responses to open-ended questions can be analyzed with the CCM, only if informants are asked a series of questions and give only one answer for each question. For example, the consensus model can handle responses to questions such as "Who owns this land?" and every time the question is asked, the informant is shown a new plot of land. The same holds true if informants are asked, "What plant is this?" (Boster 1986) for each of fifteen different plants, and each informant responds to each question with only one word, one name, or a single short phrase. The CCM can handle responses to open-ended questions, but only for single-word or short-phrase responses.

The CCM cannot accommodate response data where more than one answer is given for a single question (as is the case in free-recall listing; Weller and Romney 1988:chap. 2). However, such data may be analyzed with a related model, such as the reliability model. For example, if all informants are asked one question, "Tell me [or write one page] on why you decided to become an anthropologist," the responses may be analyzed for the presence or absence of particular themes. After collecting the narrative statements, a cumulative list of the unique themes mentioned by all the informants is compiled. Then, for each narrative, the presence or absence of each theme is noted, specifically, in a table where each row indicates an informant, each column refers to a specific theme, and cells in the table indicate whether a person mentioned a theme (1) or did not mention the theme (0). This type of problem and data coding can be handled with a reliability analysis. The presence or absence of themes can be analyzed to see the main themes (mentioned by a majority of informants), the agreement among informants, and the correspondence between each individual's themes and the main themes mentioned by the group.

If narrative responses are recorded for each question, can they be categorized into two or more categories by content? When categorizing responses into mutually exclusive categories, instead of using verbatim responses, standard coding techniques for handling of qualitative data must first be used. Rules for coding passages must be made explicit, and two or more coders should independently code the data. Because the categories no longer represent verbatim responses of informants, the formal consensus model cannot be used. If all response categories are dichotomous, a reliability analysis of respondents can be used. If multiple categories are used, the answers can be estimated by using the modal response for each question (D'Andrade 1987).

Can a written or "paper and pencil" questionnaire be used? Although most applications have used face-to-face oral interviews, the analytic techniques would apply as well to written materials. A disadvantage of written responses is that respondents must know how to read, they must understand what is expected of them, and they should not skip questions. An advantage of written materials, however, is that they may be administered in a group setting without violating the independence assumption as long as respondents do not consult one another. The original application (Romney, Weller, and Batchelder 1986) used an interview in a testlike format.

Can the consensus model be used with rating scale responses? Rating scales are a little problematic. If the ordinal information is ignored and the

responses are considered as categories, then the formal CCM can be used. If the ordinal information is used, then the informal model must be used. Although the informal model specifies fully ranked data or interval/ratio estimates, rating scales are often used in factor analyses and are probably appropriate. However, rating scales would most likely perform best with a greater number of ranks. Thus, when designing interview or questionnaire materials, 7-point rating scales are better than 3-point rating scales. Because the similarity between pairs of individuals is measured with the Pearson correlation coefficient, it is not necessary to standardize respondents' rating scales prior to analysis.

When using rating scales or ranking, should there still be positive and negative items? As for questions with dichotomous response categories, questions with rating scale responses also should have a balance between positive and negative questions. For rating scales, this is like considering responses on a 6-point rating scale where responses that are 3 and above would be considered as 1s and those below 3 would be considered as 0s.With dichotomous responses, some questions should be answerable as true or yes and some as false or no. With rating scale response, some questions should be answerable on the high end of the scale and others on the low end of the scale. Ideally, there should be roughly an equal number of each type of question or statement to ensure that there is some variability between respondents.

An exception to the balancing of positive and negative items, however, can occur for ranked items. Part of the reason for this is that a maximum amount of variance is forced into the data of each individual by using all ranks from 1 to *k*. Items should represent a range of possibilities but should be expressed in a similar way. A difference between desirable or good aspects and undesirable or bad aspects can dominate a rank order, with good aspects consistently being ranked higher than those considered to be bad. Good and bad features should be explored separately, or all items need to be stated in either a positive or negative form. For example, when asking individuals to rank the desirability of neighborhood attributes, an item might be "Houses are not painted, and yards are untended." Among a set of positive attributes, it might be better to say, "Houses are painted regularly, and the yards are well kept." If a subgroup of items is expressed in a negative form and a subgroup in a positive form, the negative items may be consistently ranked low and the positive items consistently ranked high, which may artificially inflate agreement. For example, in a study of personal attributes, Romney et al. (1979) and Freeman et al. (1981) divided the items into the two subsets' descriptors of success and failure, because detail within each subgroup was overwhelmed by the difference between the two groups.

Can the consensus model be used with judged similarity data, such as pile sorts or triads? Currently, the formal consensus model cannot accommodate this type of data. Remember that the formal model models the process of how questions are asked and answered, and similarity data do not fit the assumptions. Triad similarity judgments could be analyzed with the formal consensus model with three choices for each set of items, but the answers for each item are not completely independent of the answers for other sets. A preferable way to analyze triadic similarity data is to transform the triad choices into the pairwise similarity vector for each individual and then use reliability or factor analysis. For k items, this would result in information on $k(k-1)/2$ pairs. Thus, a ten-item triad task would result in a factor analysis on people across the forty-five pairs. Pile sort data could also be analyzed with a reliability or factor analysis, if and only if each person has the same number of piles. As for triads, individuals can be compared in terms of their vector of responses (representing each pair of items), but for pile sort data, 0s and 1s indicate whether the pair is alike or not.

Sample Size Estimation

Because the relationship between agreement, the number of informants, and the validity of the aggregated responses is formalized, sample size requirements for such studies can also be calculated. Using the Spearman Brown prophesy formula (Nunnally 1978; Weller and Romney 1988:72), sample size can be calculated for various levels of agreement (expressed as the average Pearson correlation coefficient between all pairs of respondents or as cultural competency) and for different levels of validity (the correlation between the estimated answers and the true answers). The average Pearson correlation coefficient between respondents is equivalent to the squared average group competency (Weller 1987). So if the average correlation between respondents is .25 (average cultural competency is .50) and the sample size is twenty-eight, the averaged responses (estimated answers) would correlate with the true answers at about .95 (see Table 1). Similar estimates can be calculated (Table 2 for the consensus model; Romney, Weller, and Batchelder 1986): At a comparable level of agreement (cultural competency = .50), a sample size of twenty-three would correctly classify 95% of the answers at the .99 confidence level.

How do I know what sample size I need before I do a study and I do not know the level of agreement or competency I will find? In general, when the level of shared beliefs reaches the 50% level (average cultural competency is .50), there is sufficient agreement that the consensus model can identify

TABLE 1
Sample Size and Validity Estimates for Different Levels of Agreement

	Validity[a]				
Agreement[b] *(competency)*	*.80*	*.85*	*.90*	*.95*	*.99*
.16 (.40)	10	14	22	49	257
.25 (.50)	5	8	13	28	148
.36 (.60)	3	5	8	17	87
.49 (.70)	2	3	4	10	51

SOURCE: Weller and Romney (1988:77); table values from Spearman Brown prophesy formula.
a. Validity is the correlation between the aggregated responses and the "true" answers.
b. Agreement is expressed as the average Pearson correlation coefficient (and average competency).

TABLE 2
Sample Size and Validity Estimates for Different Levels of Agreement

	Proportion of Items Classified Correctly at .99 Confidence Level				
Cultural Competency	*.80*	*.85*	*.90*	*.95*	*.99*
.50	15	15	21	23	*
.60	10	10	12	14	20
.70	5	7	7	9	13
.80	4	5	5	7	8
.90	4	4	4	4	6
	Proportion of Items Classified Correctly at .999 Confidence Level				
Cultural Competency	*.80*	*.85*	*.90*	*.95*	*.99*
.50	19	21	23	29	*
.60	11	13	13	17	23
.70	7	8	10	10	16
.80	6	6	8	8	12
.90	4	4	5	5	7

SOURCE: Table adapted from Romney, Weller, and Batchelder (1986).
*requires more than thirty informants.

a single response pattern. A strong cultural pattern appears to be present when this level reaches about 67% (.66). When beginning a study, it is best to estimate the agreement conservatively, that is, to (1) assume that there will only be a low level of agreement, say 50% sharing of beliefs (competency), and (2) to require

a high accuracy of the answers (.95 validity). The minimum sample size, given these stringent criteria, is about thirty people (per group). If only twenty people are interviewed and average competency is higher than .50 (say it is .60), then the answers can be used with the same degree of accuracy or confidence. However, if only ten people are interviewed and the average competency is .50, then the answers have much lower accuracy (validity between .85 and .90).

Can I do a consensus analysis with responses from only four people? The number of people necessary is determined by the agreement observed in their responses and the accuracy with which you hope to estimate the answers. For extremely high consensus data, four people may be sufficient. After a study is completed and the level of agreement (competency) is known, the validity of the estimated answers can be determined.

Doing the Analysis

What software is available to do these analyses? The formal consensus model is currently available in ANTHROPAC (Borgatti 1996) and UCINET (Borgatti, Everett, and Freeman 2002). Software for the informal model is available in most statistical packages. The informal model can be run with a factor analysis procedure, the minimum-residual method that solves for the missing diagonal without rotation. However, when statistical procedures such as reliability or factor analysis are used for consensus applications, the data must be analyzed with the dataset transposed from its usual structure so that questions become the unit of analysis (the rows in a data matrix) and people are the variables (the columns in a data matrix).

A reliability analysis can be conducted with standard statistical software such as SPSS or SAS. Before you begin, the data must be transposed so that people are the columns and the questions are the rows. Dichotomous, ranked, and interval responses can be used. The software should automatically provide the reliability coefficient, and you can take the square root of that value to obtain the validity coefficient (the correlation between the aggregated responses and the true answers). The estimated answers to the questions are generally provided as the average score for each question. To get an accuracy estimate for each respondent, request the item-to-total correlations. These numbers are the correlation between the responses of each individual and the aggregated responses of the group (minus that individual). You can also request the average Pearson correlation coefficient.

A factor analysis of respondents can be used to obtain the estimated answers and individual accuracy estimates for ordinal- or interval-scaled responses. For example, a sample of respondents might be asked to estimate the heights of each of twenty people, or a set of twenty items might be ranked on a single

concept from 1 to 20. Again, the data must be transposed from the usual structure so that respondents are the columns in the dataset and items are the rows. A factoring procedure is used to solve for competency from agreement. In general, select the minimum-residual or principal factor algorithm (no rotation); this is the factoring method that assumes that the main diagonal of the correlations matrix between respondents is missing and must be estimated. Similarity between respondents will be calculated with the Pearson correlation coefficient (usually the default method). The solution should not be rotated. In the output, look for the eigenvalues, factor loadings, and factor scores. The eigenvalues help determine whether the solution is unidimensional (the first value divided by the second should be 3 or greater). The factor loadings are the estimated individual competence values. The factor scores are the estimated answers; they are the weighted, aggregated responses. Factor scores are readily provided in some programs (SPSS) and not so readily in others.

Factor scores are usually provided as standardized variables (mean of zero, standard deviation of one) but may be transformed back to your original units of data collection. To transform the first set of factor scores so that the answers look like your original units, multiply each value by the original standard deviation and then add the original mean:

$$A_i = (F_i \cdot s_o) + M,$$

where A_i is each answer in original units, F_i is the factor score on the first factor for the ith answer, s_o is the original standard deviation, and M is the original mean. For example, if the task involved ranking 20 items from 1 to 20, the original mean (M) is 10.500 and the standard deviation (s_o) is 5.916. To begin to transform standardized factor scores from the computer output, a standardized factor score of 1.60 would be 19.97 ($1.60 \times 5.916 + 10.500$) in the original units and a value of -1.60 would be 1.03. (The factors scores can also simply be ranked from 1 to 20.)

The formal CCM (Romney, Weller, and Batchelder 1986) can only be run in ANTHROPAC (Borgatti 1996) and UCINET (Borgatti, Everett, and Freeman 2002). This model only accommodates categorical response data. For ANTHROPAC, data may be in the more common form, where respondents are rows and questions are columns, since the program is smart enough to use the respondents as the units of analysis. Select "analysis" and then "consensus." With dichotomous response data (only two response choices for each question), similarity between respondents may be assessed with either the match method or the covariance method (see more below). With three or more response choices, only the match method may be used. ANTHROPAC will automatically provide all information: the eigenvalue ratio (to assess goodness of fit), the individual competency scores (to assess variation in knowledge about the questions), and the estimated answers to the questions (the answer "key") with the likelihood for each response that it is the correct response.

With the formal consensus model, how do you decide whether to use the match or the covariance method when you have dichotomous response data? As mentioned previously, the match method assumes no response bias, and the covariance method assumes that the proportion of true or yes answers is .50. Romney (1999:S109) advocates using both methods and comparing the competency estimates. If the positive (true or yes) and negative (false or no) items are balanced (around .50 but between .30 and .70), the difference between the competency scores estimated by the match and covariance methods can indicate if there is response bias present and how much there might be. The match estimates are inflated when response bias is present. I typically try to keep an equal proportion of positive and negative items and use the covariance method. Note, however, that if response bias is present, the answers with either method will be less accurate (Weller and Mann 1997). The new models of Karabatsos and Batchelder (2003) now attempt to model the bias parameter for each individual.

How can you tell if consensus theory is appropriate for your data; how do you determine whether the model fits the data? An assumption of consensus theory is that there is a single, shared set of culturally appropriate answers for the questions asked. This means that there is high agreement with consistency in responses. A first step, therefore, is to test whether this is true. If it is true, then consensus analysis may be used. If not, some other approaches are necessary. The ratio of the first and second eigenvalues tells whether there is a single response pattern present in the data. In spite of variation among individuals in answers, if there is a single pattern of answers, the ratio of the first and second eigenvalues will be large. If a plot were made of the dimensions/factors (horizontal axis) by the eigenvalues (vertical axis), there should be a large relative drop in value from the first to the second value. This plot is called a "scree plot." By custom, if the first value is three times larger than the second, it can be reasonably assumed that there is only a single factor present in the response data.

Homogeneity of items (equal item difficulty) is also an assumption but not often checked. To check for the homogeneity of the questions, questions can be randomly divided into subparts and competence calculated for the respondents on each of the subsets. Competence for respondents is then correlated between the two subsets of questions. The correlation between the two sets of competency estimates should be high, meaning that those who

have high competence on one set of questions should also have high competence on another set of questions.

What if the model does not fit? If the initial analysis of pairwise agreement between informants indicates that there is more than one factor or dimension in the solution, other methods must be sought. One way to explore this issue is to examine the competence scores (the factor loadings from the first factor) and see if they correlate with any other information about the people in the sample. When there is more than one group of respondents present (e.g., more than one answer key or set of beliefs), the factor loadings on the first factor may have a large proportion of negative scores. Do women score high (positive scores) and men low (negative scores)? Or are high (positive) scores associated with urban residents and low (negative) scores with rural residents? This would indicate the presence of more than one subcultural group. In the case of two such groups, the competence scores may be positive for one group and negative for the other. In that case, the groups should be analyzed separately.

Sometimes, it may appear that the model does not fit, but what actually happened is that there was insufficient variation in responses. In fact, respondents can report that they agree strongly with all items, but the analysis suggests that there is no consensus. This is because of a statistical artifact that occurs when there is little variation. For example, consider a series of dichotomous questions where all respondents answered all questions with yes or if items were rated from 1 *strongly disagree* to 7 *strongly agree*, but all respondents rated each item as 7 *strongly agree*. There would be high agreement and no variation. What can you conclude in the face of perfect agreement? The conclusion is that yes, indeed, there is high agreement that these items are important. What would happen if these hypothetical data were passed through a consensus analysis? The analysis would blow up, because a correlation coefficient cannot be calculated without variance and is thus undefined. Note that this occurs if informants all answer "agree" or if they all answer "disagree."

Although the above example is extreme, the same problem occurs when there is very low variance across respondents—if, for example, the average response were 6.0 for each of the items, but some individuals choose 5, some choose 6, and some choose 7 as their responses. There would still be high overall agreement, since no one has reported that they disagreed about these propositions (e.g., no one has chosen the rating scales values of 1, 2, 3, or even 4). However, the average correlation coefficient between respondents would approach 0. The problem is not whether there is agreement but that there is agreement with little variation across items so that statistical

measures of agreement tend toward 0 or toward being undefined. The problem of insufficient variance can usually be avoided by having a balance of positive and negative items.

Do I need to use other methods to explore patterns in responses, such as MDS? If data fit the CCM, then the data are unidimensional. The respondents have provided responses that are sufficiently homogeneous that their patterns of responses in the agreement matrix can be captured with the single array of competence scores. Subgroups of informants (that is, subcultures) may be compared using the competence scores (or loadings for the first factor). Multidimensional scaling is generally not necessary and in fact may artifactually display the unidimensional data as a horseshoe shape in two dimensions (Kruskal and Wish 1990).

Could there be subgroup variation in beliefs in addition to the shared or consensual beliefs? When looking for beliefs in addition to those described by the CCM, loadings on the second factor may be used for comparisons. To date, only a few researchers have explored this variation. Boster (1986) identified variation between kin groups in addition to the overall agreement on plant naming. Boster and Johnson (1989) examined variation in addition to the consensual pattern in fish identification. Handwerker (2002) studied parent-teacher activities and identified a subsample of respondents that appeared to have a slightly different organization of the activities. Recently, Berges et al. (2007) identified an overall pattern for seeking social support and identified variations between minority and nonminority preferences in addition to the general pattern.

Although the first factor solution provides the unidimensional ordering of respondents according to their competence or accuracy in reporting (values on the first factor loadings) and an estimate of the answers, there may be additional information in subsequent factors. There are a few points to consider when searching for additional patterns. First, any known information about respondents (age, ethnicity, etc.) may be compared to the factor scores to see if there are subgroups of people that may have some systematic differences in responses. Those variables allow for the sample to be divided into specific subgroups. Second, analysis of the subgroups separately provides information about the agreement and coherence of beliefs within each group and provides an estimate of the answers for each group. Finally, it is important to assess the correlation between the answers/solutions between the subgroups and to estimate the validity of the solutions (given the sample sizes) to ensure that differences between the groups are not simply caused by sampling variation. If a subgroup has fewer than twenty people

or the average agreement drops within a subgroup, the answers may not be reliable. Reliability and validity of the answers can be estimated with Table 1 or 2. A correlation calculated between the answer keys for each subgroup can help determine how similar or different they are.

In the study of social support, the general pattern of social support shared across demographic subgroups was for kin and then nonkin sources of support. However, a comparison of demographic information on the respondents (such as marital status, ethnicity, and level of education) indicated some variations in the model between minority (African American and Mexican American) and nonminority (white) respondents. Finding the difference is relatively easy: Statistical tests can be used to compare demographic and other information on respondents with the loadings on the second and subsequent factors. Interpreting the difference is more complex. Berges et al. (2007) conducted consensus analyses separately for the minority and white respondents as well as for everyone together and compared all three answer keys. There was sufficient agreement and a sufficient sample size, given the level of agreement, to obtain stable answers for each subgroup. It was in this detailed comparison that it was evident that whites, although preferring kin to nonkin sources, would also consult nonkin "others" when they had a problem with their spouse or if they needed to borrow money.

Does a high eigenvalue ratio mean that there is high agreement? A high ratio between the first and second eigenvalues indicates that the solution has only a single dimension. When a single dimension is present, then by definition, there is moderately high agreement present in the response data. The actual value of the ratio, however, is not directly related to the average level of agreement present in the data. The average competence score provides information directly about the level of agreement present in the data, because the square of the average competency level of a group is approximately equal to the average Pearson correlation coefficient between all pairs of respondents (Weller 1987).

If competence scores can only range between 0 and 1, inclusive, what do I do if I have one competence score of −.01? If there is only one negative score and it is very close to 0, you might assume that it is 0. However, if there is more than one negative score (or score greater than 1.0) or the negative (or greater than 1.0) scores are not close to zero, then the model does not fit, at least for those individuals.

Can I just drop or omit people that don't fit the model? Eliminating people from the sample can be unethical, so this can only be done under very careful circumstances and with full disclosure of the process and rationale. For example, if one person had a moderately sized negative competence score, this would be reported; the person's demographic variables could be used to see if other such people maybe scored low (younger vs. older, low vs. high education). If only one person appeared unusual, this would be reported and the analysis might be tried again without that person. In general, omitting part of a sample is considered unethical because the results do not describe the whole sample. When a sample is selected to be representative of some group, all of the sample must be accounted for in the analysis. If subsamples are to be analyzed separately, the rationale and explicit criteria defining subgroups (such as demographic categories) need to be stated, and then the groups may be analyzed separately. This was the case in the study by Berges et al. (2007) that found ethnic differences in social support patterns.

Can this model be used in participant observational research, where the data are observations or behaviors instead of beliefs? Because the formal CCM represents the process of asking and answering questions, it is not appropriate for behavioral or observation data. To study normative behaviors from observations of behaviors (and not reports), a similar approach could be used by using a related analysis with fewer assumptions and requirements. Either a simple aggregation of behaviors (estimating the modal action) or a reliability analysis of such behaviors would provide analogous information.

Does the method create agreement? Although there are group processes whose purpose is to bring a group of people to consensus (such as nominal and Delphi group processes), the CCM is not such a group consensus-building technique. Instead, it is a method to evaluate the level of agreement within responses by a sample of informants. The CCM assumes that responses of informants are independent of one another and not reported in a group setting or with consultation with others. The CCM does not create agreement or make assumptions about how agreement may have occurred among the informants; it only tests for and describes agreement that is there.

Are there other statistical methods to assess consensus? As with most statistical techniques, there are inferential and descriptive methods. Inferential tests determine whether the observed agreement is significantly greater than chance. Significance, however, is driven by the amount of agreement and the number of respondents. As mentioned previously, a binomial test can be used to test whether agreement is greater than chance for a single question. Two tests are available for measuring agreement across a series of items that

have been ranked. One is Kendall's test of concordance, W, and another is Friedman's test (see Siegel 1956 for both). Descriptive measures of the amount of agreement are the average Spearman's rho (in Kendall's concordance) and the average Pearson correlation coefficient used in reliability analysis. An inferential test might indicate that agreement is greater than chance, but a consensus analysis (a principal components analysis of respondents) can indicate if there is one or more patterns in that agreement.

Interpreting Results

Confidence levels and the culturally correct answers. If the consensus model fits the data, then both the formal and informal models provide the best estimates for the answers to the questions, given the responses. For the informal model, these appear as the standardized factor scores and may be rescaled to look more like the original units. With the formal CCM, a probability value is associated with each response category. These are the Bayesian-adjusted posterior probabilities. A Bayesian adjustment is a revision of the probability, given the evidence of the pattern in responses; it is the likelihood that each possible answer is correct, given the competency of the individuals and the pattern of responses. Before responses are collected, each dichotomous question has a 50% chance that the answer is (a) and a 50% chance that the answer is (b).

After estimating the competency of the respondents and the pattern of their responses, these probabilities may be revised so that the probability that the answer is (b) is 99.9%, and the probability that the answer is (a) is 0.01%. These latter, posterior probabilities are used to determine the set of answers. Because these values are sensitive to sample size, it is important to use a criterion that considers sample size. For similar levels of agreement, larger sample sizes classify items at higher confidence levels: Compare .90, .95, .99, and .999 confidence levels and sample sizes in Romney, Weller, and Batchelder (1986) and .99 and .999 confidence levels in Table 2. This means that if one moves between applications with different sample sizes, different confidence levels may be appropriate. A very basic introduction to Bayes's conditional probabilities may be found in introductory statistics books that cover probabilities (Harnett 1975:64–68). Romney, Weller, and Batchelder (1986) detail the calculations to show how the confidence level is found.

Does pooling of informants' responses "idealize" cultural beliefs? It is incorrect to think that CCT is somehow different from other statistical techniques or theories (for a full discussion, see Romney 1999). CCT is not really different from other statistical methods. Methods have assumptions about the level of measurement (categorical, ordinal, interval) appropriate for analysis and a way to determine whether a model fits the data (descriptive measures of goodness of fit, such as a correlation coefficient). Statistical methods are not synonymous with the theories and hypotheses that they test; they are tools to test ideas.

Pooling or aggregating responses is not a new technique. We know from statistics (the Central Limit Theorem) that an average of responses is an unbiased estimate of the population mean when representative sampling is used. We also know that although an aggregation of responses is the single best estimate, the relative accuracy of that estimate diminishes as variance in responses increases. So for a single item/question, if the range of responses is large and responses deviate greatly from the mean, accuracy of the mean is less than when the responses cluster closely about the mean. For example, if seven individuals were asked to report the age at which a particular ritual should be performed for girls and respond with 18, 20, 20, 21, 22, 22, and 24 and when asked the same question for boys they respond 6, 13, 18, 21, 24, 29, and 36, the summary response for both questions would be 21, but the estimate of 21 serves as a better summary measure for the first question than for the second.

The relation between the aggregate and within-group variance is also true for dichotomous data. If seven people are asked if infant boys should be circumcised at birth and they respond yes, yes, no, yes, yes, yes, yes and then they are asked if girls should have their ears pierced at birth and they respond yes, no, no, no, yes, yes, yes, then we see that there is lower variance in responses about circumcision (86% said yes) than there is for ear piercing (57% said yes). The modal response in both cases is yes, but circumcision is more strongly supported than is ear piercing for this sample.

What CCT does, in essence, is to use an aggregation of responses to estimate the answers to questions and to see how much each person's responses deviate from it. The first test, however, is to check the heterogeneity in responses. The ratio of the first to second eigenvalues provides an indicator of whether the responses are homogeneous or heterogeneous across people. When the first eigenvalue is large, relative to the second eigenvalue, then there is a single response pattern present in the data. When this indicator of goodness of fit is not met, then responses are heterogeneous, and there may be multiple or no patterns present in the responses. This threshold level occurs at approximately an average competence (shared knowledge level) of less than .50; the equivalent to an average interperson correlation coefficient of .25 (Weller 1987). Thus, consensus theory tends only to be applicable to data with high agreement.

Can I generalize to a whole culture from a single sample? One of the most important issues in interpretation and generalization of sample results is how the sample was selected. If the sample is not representative of a larger group, then generalizations are limited to the sample. The methods of selecting respondents should parallel the degree to which you want to generalize. If the sample of informants is representative of a larger group or if several samples represent population variation, then results may be generalizable from the sample to a larger population.

Can agreement vary across different samples of respondents and different samples of questions? First, it is important to select respondents according to the purpose of the study and degree to which one would like to generalize from the findings (Johnson 1990). Similarly, items/questions should be representative, if not exhaustive, of the set of items on a particular topic or domain. Items should be homogeneous in terms of difficulty. When items come from a coherent domain and are homogeneous, the relative competence of respondents should be stable across different subsets of questions.

Aunger (2004:78) expressed concern that results could indicate that the sample as a whole might agree when, in fact, some subgroups might have little or no agreement. Some cultural knowledge is highly and evenly distributed across culture members. Some is not. When knowledge is unevenly distributed, one subgroup may have expertise while others have little or none. This variation is evident in the set of competency scores. As long as those with less expertise do not have a distinctly different response pattern, results would indicate that there is a single model and that one group knows it better than another. For example, when young adults judge the relative effectiveness of an array of contraceptive methods, women tend to have higher competence (meaning they rank the methods similarly), and men tend to have lower scores (their responses do not indicate a different pattern; they are just less consistent in their rankings). If analyzed separately, it is possible that the men's responses might not have sufficient agreement to indicate consensus. These results would indicate that there is one model (not two separate models) and that one group knows it more than another. When one group has little or no knowledge, it does not detract from finding the predominant pattern of shared cultural knowledge; it just takes a larger sample to find it, since that group adds "noise" to the data. The average agreement for the whole group will be between the high value of the experts and the lower value among the novices.

Can you really say that consensus finds the "beliefs" of the respondents (e.g., what they actually believe)? Responses may not actually estimate beliefs; instead, results are a summary of what people say. Results assume, as do most techniques that rely on interview data, that respondents, for the most part and to the best of their ability, tell the truth. Respondents, of course, vary in what they know and in their ability to recall and report what they know. The effect of individuals who might purposively mislead an investigator are mitigated by the agreement of others. Only wide-scale collusion, with systematic bias across most of the respondents, would affect results. Systematic bias, on the other hand—the tendency to answer in a particular way when unsure of the answer—can affect the accuracy of results (Weller and Mann 1997).

SUMMARY

Currently, CCT has two basic forms: a formal cognitive model and a set of analytic procedures that serves as an informal version of the model. Both estimate the group normative answers and individual competence in reporting on those answers. The informal consensus model is a principal components-type analysis of people. With this analysis, the degree to which individuals correspond to the group or the proportion of beliefs shared between an individual and the group is estimated with factor loadings, and the answers are estimated with the factor scores. A reliability analysis of people can provide similar information. These analyses are available in most major statistical software packages but must be run with people as the column variables and the questions as the units of observation (rows). The informal consensus model can handle response types that are appropriate for the Pearson correlation coefficient: dichotomous, interval, or ratio-scaled response data. Ranked data are also often used.

The formal version of the CCM is more than a set of analytic procedures. Rather, it is a family of models that represent the answering process (for example, with guessing) and estimate various parameters in the answering process. The simplest model and the one described in this article is the original or basic consensus model. In the basic model, parameters that must be estimated are the amount of guessing and the proportion of true or yes responses in the answer key. The model assumes item homogeneity, no response bias, and one answer key. It currently accommodates only categorical response data (dichotomous, multiple-choice, and open-ended short answers). Agreement is measured with the match method when there are two or more response categories or the covariance method when there are only two response categories. The answer key is estimated by weighting the responses of individuals by their cultural competency scores, and Bayesian posterior probabilities express the confidence level for each possible answer.

More consensus models are being developed. Because fitting the basic model to real-life field data can stretch assumptions of the basic model, new extensions test and model additional parameters in the answering process. Both the covariance and match method estimates of answers are affected by response bias and become inaccurate when response bias is present (Weller and Mann 1997). A new model (Karabatsos and Batchelder 2003) tests for the presence of response bias and whether items vary in their difficulty (and are thus not homogeneous) as well as estimating competency and the answers. Another version under development (Batchelder and Romney 1989) tests the assumption that there is only one answer key. This latter version will be able to handle individuals or groups with different beliefs (e.g., different sets of answers) and classify the individuals into separate classes. This model will help solve the problem of determining whether samples have meaningful differences in response patterns.

REFERENCES

Aunger, R. 2004. *Reflexive ethnographic science.* Walnut Creek, CA: AltaMira.

Baer, R. D., S. C. Weller, J. G. de Alba Garcia, M. Glazer, R. Trotter, L. Pachter, and R. E. Klein. 2004. A cross-cultural approach to the study of the folk illness *nervios. Culture, Medicine, and Psychiatry* 27 (3): 315–37.

Baer, R. D., S. C. Weller, L. M. Pachter, R. T. Trotter, J. G. de Alba Garcia, M. Glazer, R. Klein, L. Deitrick, D. F. Baker, L. Brown, K. Khan-Gordon, S. R. Martin, J. Nichols, and J. Ruggiero. 1999. Cross-cultural perspectives on the common cold: Data from five populations. *Human Organization* 58 (3): 251–60.

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

———. 1989. New results in test theory without an answer key. In *Mathematical psychology in progress*, edited by E. E. Roskam, 229–48. Heidelberg, Germany: Springer-Verlag.

Berges, I. M., F. J. Dallo, A. DiNuzzio, N. Lackan, and S. C. Weller. 2007. A cultural model of social support. *Human Organization* 65 (4): 420–29.

Borgatti, S. 1996. ANTHROPAC 4.0. Columbia, SC: University of South Carolina Press.

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. 1986. Exchange of varieties and information between Aguaruna manioc cultivators. *American Anthropologist* 88 (2): 428–36.

Boster, J. S., and J. C. Johnson. 1989. Form or function: A comparison of expert and novice judgments of similarity among fish. *American Anthropologist* 91 (4): 866–89.

Chavez, L. R., F. A. Hubbell, J. M. McMullin, R. G. Martinez, and S. I. Mishra. 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* 9 (1): 40–74.

D'Andrade, R. G. 1987. Modal responses and cultural expertise. *American Behavioral Scientist* 31 (2): 194–202.

Dawes, R. 1977. Suppose we measured height with rating scales instead of rulers. *Applied Psychological Measurement* 1 (2): 267–73.

Dressler, W. W., M. C. McBalieiro, and J. E. Dos Santos. 1997. The cultural construction of social support in Brazil: Associations with health outcomes. *Culture, Medicine, and Psychology* 21 (3): 303–35.

Freeman, H. E., A. K. Romney, J. Ferreira-Pinto, R. E. Klein, and T. Smith. 1981. Guatemalan and U.S. concepts of success and failure. *Human Organization* 40 (2): 140–45.

Gordon, K. 1926. A study of aesthetic judgments. *Journal of Experimental Psychology* 1 (1): 36–40.

Handwerker, W. P. 2002. The construct validity of cultures: Cultural diversity, culture theory, and a method for ethnography. *American Anthropologist* 104 (1): 106–22.

Harnett, D. L. 1975. *Introduction to statistical methods.* 2nd ed. Menlo Park, CA: Addison-Wesley.

Johnson, J. C. 1990. *Selecting ethnographic informants.* Qualitative Research Series, Vol. 22, Newbury Park, CA: Sage.

Karabatsos, G., and W. H. Batchelder. 2003. Markov chain Monte Carlo estimation theory for test theory without an answer key. *Psychometrika* 68 (3): 373–89.

Kruskal, J. B., and M. Wish. 1990. *Multidimensional scaling.* Newbury Park, CA: Sage.

Magana, J. R., M. Burton, and J. Ferreira-Pinto. 1995. Occupational cognition in three nations. *Journal of Quantitative Anthropology* 5 (2): 149–68.

McKeown, B., and D. Thomas. 1988. *Q methodology.* Newbury Park, CA: Sage.

Nunnally, J. 1978. *Psychometric theory.* New York: McGraw Hill.

Pachter, L. M., S. C. Weller, R. D. Baer, J. G. de Alba Garcia, R. T. Trotter, M. Glazer, and R. E. Klein. 2002. Variation in asthma beliefs and practices among mainland Puerto Ricans, Mexican-Americans, Mexicans, and Guatemalans. *Journal of Asthma* 39 (2): 119–34.

Romney, A. K. 1999. Culture consensus as a statistical model. *Current Anthropology* 40 (Supplement): S103–15.

Romney, A. K., W. H. Batchelder, and S. C. Weller. 1987. Recent applications of cultural consensus. *American Behavioral Scientist* 31 (2): 163–77.

Romney, A. K., T. Smith, H. E. Freeman, J. Kagan, and R. E. Klein. 1979. Concepts of success and failure. *Social Science Research* 41 (8): 302–26.

Romney, A. K., and S. C. Weller. 1984. Predicting informant accuracy from patterns of recall among individuals. *Social Networks* 6 (1): 59–77.

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

Siegel, B. 1956. *Nonparametric statistics.* New York: McGraw-Hill.

Stroop, J. R. 1932. Is the judgment of the group better than that of the average member of the group? *Journal of Experimental Psychology* 15 (5): 550–62.

Trotter, R. T., S. C. Weller, R. D. Baer, L. M. Pachter, M. Glazer, J. E. Garcia de Alba Garcia, and R. E. Klein. 1999. A consensus theory model of AIDS/SIDA beliefs in four Latino populations. *AIDS Education Prevention* 11 (15): 414–26.

Weller, S. C. 1987. Shared knowledge, intracultural variation, and knowledge aggregation. *American Behavioral Scientist* 31 (2): 178–93.

———. 1998. Structured interviewing and questionnaire construction. In *Handbook of methods in cultural anthropology*, edited by H. R. Bernard, 365–410. Walnut Creek, CA: AltaMira.

Weller, S. C., R. D. Baer, L. M. Pachter, R. T. Trotter, and J. G. de Alba Garcia. 2003. Regional variation in Latino descriptions of *susto. Culture, Medicine, and Psychiatry* 26 (4): 449–72.

Weller, S. C, R. D. Baer, L. M. Pachter, R. T. Trotter, M. Glazer, J. E. Garcia de Alba Garcia, and R. E. Klein. 1999. Latino beliefs about diabetes. *Diabetes Care* 22 (5): 722–28.

Weller, S. C., R. D. Baer, L. M. Pachter, R. T. Trotter, M. Glazer, R. E. Klein, J. E. Garcia de Alba Garcia, M. Glazer, and Z. Castillo. 1994. Empacho in four Latino groups: A study of intra- and inter-cultural variation in beliefs. *Medical Anthropology* 15 (2): 109–36.

Weller, S. C., and N. C. Mann. 1997. Assessing rater performance without a "gold standard" using consensus theory. *Journal of Medical Decision Making* 17 (1): 71–79.
Weller, S. C., and A. K. Romney. 1988. *Systematic data collection.* Newbury Park, CA: Sage.
Weller, S. C., A. K. Romney, and D. P. Orr. 1986. The myth of a sub-culture of corporal punishment. *Human Organization* 46 (1): 39–47.
Young, F. W., and R. C. Young. 1962. Key informant reliability in rural Mexican villages. *Human Organization* 20 (3): 141–48.
Zajonc, R. B. 1962. A note on group judgments and group size. *Human Relations* 15 (2): 177–80.

SUSAN C. WELLER is a professor of preventive medicine and community health and director of research in the Department of Family Medicine at the University of Texas Medical Branch in Galveston. She has worked as a medical anthropologist and epidemiologist since receiving her PhD in social science from the University of California, Irvine. Her research has focused primarily on the measurement of cultural beliefs. She was one of the codevelopers, with A. K. Romney and W. H. Batchelder, of the cultural consensus model. Using the consensus model, she has published on beliefs about AIDS, diabetes, asthma, the common cold, and folk illnesses (empacho), susto, *and* nervios.

Accessing Married Adolescent Women: The Realities of Ethnographic Research in an Urban Slum Environment in Dhaka, Bangladesh

SABINA FAIZ RASHID
BRAC University

This article reports on the problem of obtaining reproductive histories from women in the slums of Dhaka, the capital of Bangladesh. Access to women in these slums is controlled by several gatekeepers. The gatekeeper problem is common in all field research, but the problem is particularly difficult when the research involves interviewing young Muslim women on the sensitive issue of reproductive health and family planning.

Keywords: *ethnography; urban slums; adolescent women; gatekeepers; reproductive health*

It is 9:00 a.m., May 9th, when we arrive at Phulbari slum. People are running everywhere. We notice blood splattered from Nur Islam's [a leader in the slum] home right up to the health clinic. The entire mud path is covered in blood. We quickly enter the clinic. Sufia Khala [a health worker who lives in the slum] says, "The leaders are fighting with each other about money, and rival gangs are also involved. The gangs are looking for Mostafa [another leader] to kill him. They didn't find him, but they found Kala Sayeed [another leader], and they cut his hands." Sayeeda Apa [another

This article is drawn from my PhD dissertation in medical anthropology/public health, titled "Worried Lives: Poverty, Gender and Reproductive Health Needs of Married Adolescent Women Living in Urban Slums in Bangladesh" (2004), National Centre for Epidemiology and Population Health, Australian National University, Canberra. This study was supported by the Special Programme of Research, Development and Research Training in Human Reproduction, Department of Reproductive Health and Research, World Health Organization, Geneva, Switzerland. I am grateful to the young women in the slums for their patience and hospitality and for the time they gave me so generously despite the constraints in their lives. I thank Nipu, my colleague, research assistant, and friend, for her support during fieldwork. I am also grateful for the critical feedback, editing, and guidance of the editor and reviewers of this article.

Field Methods, Vol. 19, No. 4, November 2007 369–383
DOI: 10.1177/1525822X07302882