The BEHAVIORAL MEASUREMENT *Letter*

Behavioral Measurement Database Services

Enriching the health and behavioral sciences by broadening instrument access

Vol. 6, No. 2
Fall 1999

Introduction to This Issue

I begin my introduction to this issue of *The* Behavioral Measurement *Letter* with an apology to both its readers and contributors. This issue is very much behind schedule due mostly to me, more specifically, a series of illnesses I suffered and the resultant accumulating backlog of various items of business, The BML included. I want to thank all of you for your patience and understanding, and especially our regular contributor, Fred Bryant, for his expert assistance in editing this issue.

One of the most interesting yet least investigated topics at the intersection of health care and social science is the relationship of spirituality to health and illness. In this issue, Bruce Frey and Timothy Daaleman discuss their current work to define and measure the construct spirituality. Their review of prior work found that existing studies of spirituality in health have two fundamental flaws: (a) these studies mostly examine religiosity rather than spirituality, or they otherwise confound these two different constructs, and/or (b) the studies examine spirituality from the perspective of the health care provider or researcher rather than that of the patient/subject in whom the spirituality of interest resides and from whom such is expressed. Given the lack of valid studies and measures of spirituality, Drs. Frey and Daaleman have initiated qualitative studies to define the construct of spirituality so that it then may be operationalized.

Understanding a culture, one's own or a foreign culture, requires knowledge of how persons in that culture view their world. This, in turn, requires tools to obtain such knowledge. John Gatewood, a cognitive anthropologist, discusses three measurement methods for measuring similarities, differences, and relationships among products of a culture or items in its environment -- free listing, pile sorting, and triadic comparison. Using examples from our own culture and environment, he shows how these tools may be used to begin to gain an understanding of a culture. One drawback of using these tools, however, is that data analysis can be cumbersome and time-consuming. Thus, at the end of his column, Dr. Gatewood provides an Internet website address of a manufacturer of data analysis software to be used in processing data obtained by these methods.

Once again, Fred Bryant writes about measurement modeling. In this installment of a continuing series on the topic, he discusses the use of measurement modeling to determine if what is measured by the Affect Intensity Measure (AIM) is one construct or more than one construct, that is, if the AIM taps one factor or multiple factors. The work Dr. Bryant reports indicates that the AIM actually measures three factors, and that the construct "affect intensity," as measured by the AIM, therefore consists of three factors. As in preceding columns in the series, this piece clearly demonstrates how measurement modeling is useful in defining the "structure" of a measurement instrument and thereby contributes to our understanding of instruments and of constructs as operationalized by instruments.

Address comments and suggestions to The Editor, *The* Behavioral Measurement *Letter*, Behavioral Measurement Database Services, PO

Culture . . . One Step at a Time

John B. Gatewood

There is a cross-disciplinary joke in fisheries management circles. A commercial fishery is in crisis, and its managers commission research to determine the causes and possible solutions. The *biologist* evaluates a handful of variables pertaining to fish stock growth and mortality, and writes a 3-page report concluding with a recommended policy action. The *economist* considers a dozen or so variables concerning the costs and benefits of the fishery based on alternative future scenarios, and submits a 15-page report with an either/or recommendation. The *anthropologist* considers the full multitude of factors contributing to the fishery's problem, and hands in a 200-page report with no recommendation.

Contrary to such folk humor about the idiosyncratic nature of anthropological research, in this column I review some ways in which anthropologists can and do use explicit, replicable methods to make headway understanding the cultural part of social life. I will focus on a few of the rather specialized data collection techniques being used in cognitive anthropology -- specifically, free-listing, pile-sorts, and triadic comparisons. Each of these topics is well presented in several recent anthropological methods books (Bernard, 1994, 1998; Borgatti, 1996; D'Andrade, 1995; Weller & Romney, 1988).

Free-Listing

One of the initial problems facing anthropologists interested in studying aspects of culture, or cultural domains, is how to ask questions in ways that are meaningful to natives, or people living within that culture. In particular, researchers need to phrase questions using natives' own cognitive categories (Frake, 1962, 1964). But what are these categories, what are the elements of the cultural domain, and how can a non-native discover them? The free-listing task is a good way to explore native vocabularies, and it provides interesting information about research informants.

A free-listing task is virtually the same as a free-recall task in psychology, except that the period of learning is not controlled by the researcher but rather consists of the previous (and variable) life experiences of the informants. Informants are simply asked to name (if non-literate) or write (if literate) all the items they can think of that match a given category. Examples include:

"Please make a list of all the contagious diseases you can think of."

"Please name all the parts of a human body you can think of."

"Please write down all the phases of the human life cycle you can think of."

Conceptually, such tasks are quite simple and are generally well understood by the informant, they do not impose preconceived response categories, and they work well with both non-literate and literate informants. Furthermore, although the ideal is to work with informants one at a time, the same free-listing task can be given to multiple literate informants at the same time or even as a communal task for focus groups. Still, there are a few choices and problems researchers should be aware of beforehand.

First, because the usual free-listing task (such as in the instructions above) is virtually unconstrained, it is not always clear when informants are finished. Typically, informants generate the first several items rather quickly, but then slow down dramatically. Depending on the cultural domain in question, informants' knowledge of the domain, and their motivation, the task can take from a few seconds to ten minutes before they run out of steam. Related to this problem of having no clear endpoint, informants seldom enjoy free-listing tasks. The instructions sound easy, and often informants are initially eager. But once they begin and realize the full open-endedness of the task, many begin to feel apologetic for the brevity of their lists or resentful at being made to encounter their own memory limitations.

There are two effective ways to bring closure to the free-listing task, each of which makes the task less onerous and more standardized across different informants. First, the instructions can include an explicit time-limit, e.g., "Please make a list of all the kinds of diseases you can think of *in the next 3 minutes*." Alternatively, the task can specify a maximum number of items to be identified, e.g., "Can you remember any

sponsors of last year's Super Bowl? If yes, please name up to three of them." Of course, the specific limitation chosen should be guided by a pilot test of the unconstrained version of the task to observe when informants bog down, by the research objectives, and by the size of the sample of informants. Specifying a maximum number of items to list works well enough, if the domain itself is the primary research objective and the task is given to hundreds of informants. On the other hand, I prefer the time-limit approach when working with small numbers of informants (e.g., 40 or fewer), and especially if I am interested in informant-level attributes. When working with US college students, 90 seconds seems to be a reasonable limit for many domains, such as kinds of kin, trees, mammals, fish, mixed drinks, hand tools, fabrics, or musical instruments. Less indoctrinated informants might prefer a somewhat longer time-limit.

The second main problem with free-listing arises after collecting the data and beginning the analysis. The aggregated findings of free-listing tasks are displayed as a table in which rows contain the name of an item followed by the number of lists in which the item appeared, and usually the table items are sorted by decreasing frequency of mention. Many cultural domains, however, are organized taxonomically (e.g., a red oak is a kind of oak, and an oak is a kind of deciduous tree; a Neon is a kind of Dodge, and a Dodge is a kind of car; etc.), and hierarchical relations among items in a domain are not captured by a listing task. Also, some items may be synonymous with one another. These problems can make it hard to determine how many *different* items appear in the sample's lists. Suppose, for instance, that you asked four informants (A--D) to list 'five kinds of American cars' and obtained the lists in Table 1. How many different kinds of American cars appear in the four lists?

Table 1.
Four Free-Lists of Five Kinds of American Cars

List A	List B	List C	List D
Ford	Ford Taurus	Taurus	minivan
Chevy	Chevrolet Corvette	Explorer	sport utility vehicle
Dodge	Jeep Cherokee	Corvette	sedan
Cadillac	Jeep Wrangler	Neon	convertible
Jeep	Dodge Neon	Seville	station wagon

Because we are very familiar with the cultural domain of American automobiles, we can see patterns in these responses. Apparently, Informant A interpreted the instructions to mean "car companies"; Informant B likes binomial nomenclature but, like Informant C, has interpreted the task as asking for "car models"; and Informant D has interpreted the question functionally rather than along brand lines. (This sort of diversity in response indicates that our original expression, "kinds of American cars," is ambiguous and needs refining.) On the other hand, if we did not know this domain well, then our initial tally would have to rely on linguistic differences to identify different items. And on this basis, there are 20 "differently named" items in the four lists.

There are two customary ways to deal with these sorts of item identification problems. Immediately after completing the task, each informant can be asked to list alternative names for each item in his or her list. Second, the researcher can compile an initial aggregate table, then ask several of the more knowledgeable natives to judge the distinctiveness of the items. Either way, the idea is to enlist native experts to eliminate redundancy, but even these potential solutions are often ineffective, as illustrated in Table 2.

These problems aside, free-listing tasks is an excellent research tool to explore cultural domains (Gate-wood, 1983), with the added benefit that their results enable interesting compar-isons both across domains and across informants (Gatewood, 1984). If a single sample

of informants is used, domains can e compared in terms of various indices, such as median length of list and number of different items generated. At the same time, informants -- either as individuals or grouped by age, gender, expertise, ethnicity, etc. -- can also be compared.

Table 2.
Two Alternative Aggregations of Free-Lists A, B, C, and D

Item	Freq.	Item	Freq.
1. Ford Taurus : Taurus	2	1. Ford : Ford Taurus :Taurus : Explorer	4
2. Chevrolet Corvette : Corvette	2	2. Chevy : Chevrolet Corvette : Corvette	3
3. Dodge Neon : Neon	2	3. Dodge : Dodge Neon : Neon	3
4. Ford	1	4. Jeep : Jeep Cherokee : Jeep Wrangler	3
5. Chevy	1	5. Cadillac : Seville	2
6. Dodge	1	6. minivan	1
7. Cadillac	1	7. sport utility vehicle	1
8. Jeep	1	8. sedan	1
9. Jeep Cherokee	1	9. convertible	1
10. Jeep Wrangler	1	10. station wagon	1
11. Explorer	1		
12. Seville	1		
13. minivan	1		
14. sport utility vehicle	1		
15. sedan	1		
16. convertible	1		
17. station wagon	1		

Pile-Sorts

After identifying the items in a cultural domain, we can begin to examine their meanings and interrelationships, or their similarities and differences. In the 1950s and 1960s, the preferred technique was componential or semantic feature analysis (Goodenough, 1956; Lounsbury, 1956). But because different feature analyses can be formulated to explain the same data (the so-called "psychological reality" problem), research interest has shifted to studying the most salient or important semantic features in a given domain (see D'Andrade, 1995, pp. 31-91). Pile-sorts are an easy way to collect data toward this end.

The most commonly used variant of the pile-sort task is the *single* pile-sort, in which the informant is asked to group items based on their *overall similarity*. Informants are given a collection of stimuli -- either the items themselves, cards with names of items, or pictures of items. The basic task is to group the stimuli such that "similar" items are in the same pile. Informants are free to define similarity in their own terms, to make as many piles as they want, and to place very unusual items in piles by themselves. Indeed, the only constraints are (a) there must be more than one pile (extreme "lumping" is disallowed), and (b) every item cannot be in a pile by itself (extreme "splitting" is disallowed). The researcher then records each informant's pile-sort and asks why items were placed in their piles. For example, Fred's and Janine's sortings of 19 kinds of fish might be recorded as shown in Table 3.

Singleton items -- such as marlin and shad in Fred's sorting -- end up being scored as dissimilar from every other item. Given his rationales, Fred's two singleton items are properly separated. By contrast, Janine's fifth pile is problematic. She has lumped carp, marlin, and shad together *because* she doesn't know anything about the three of them, i.e., the only thing they have in common is that Janine doesn't know anything about them. The proper way to handle such "unknown" items is to treat each as a singleton; hence, when recording Janine's pile-sort, the researcher should split her residual fifth category into three singleton piles. Only by asking informants for brief rationales, however, can we catch such bogus groupings and prevent informants' lack of knowledge from distorting the results.

In doing a single pile-sort, each informant is essentially judging the similarity of every item vis-a-vis every other item, using a dichotomous scale -- any two items are either in the same pile (similar) or they are in different piles (not similar). Thus, when preparing these data for analysis, an informant's similarity judgments are represented as an item-by-item matrix in which each cell contains either 1 (items appear in the

same pile) or 0 (items appear in different piles), and each informant produces one such matrix. The aggregate judged similarity for any two items, $item_i$ and $item_j$, is calculated by adding the values in $cell_{i\,j}$ across all *N* informants and dividing by *N*. The resulting number is the proportion of informants in the sample who placed $item_i$ and $item_j$ in the same pile. The researcher can then examine the most salient similarities and differences among items in the domain by using such multivariate statistical techniques as multidimensional scaling or hierarchical clustering to analyze the aggregate similarity matrix or consensus analysis (Romney, Weller, & Batchelder, 1986) to analyze inter-informant agreement.

Table 3.
Two Informants' Single Pile-Sorts of 19 Kinds of Fish

Fred's Piles
1. barracuda, piranha, shark
2. bass, carp, pike, sunfish, trout
3. marlin
4. catfish, cod, flounder, herring, salmon, swordfish, tuna
5. goldfish, minnow
6. shad

Janine's Piles
1. piranha, shark
2. barracuda, cod, flounder, swordfish, tuna
3. bass, catfish, pike, sunfish, trout, salmon
4. goldfish, herring, minnow
5. carp, marlin, shad

Fred's Rationales
1. dangerous to humans
2. freshwater (mostly sport) fish
3. ocean sport fish
4. grocery store fish
5. weird, little fish compared to the rest
6. don't know what this is

Janine's Rationales
1. dangerous fish
2. ocean fish
3. fish found in lakes and streams
4. very small fish
5. don't know what these are

The principal advantages of the single pile-sort are that the task is easy to administer, informants enjoy doing it, and it can be done with a relatively large number of items (i.e., as many as 30-50). Also, one can achieve reliability coefficients of .90 or higher with respect to the aggregate similarity matrix using as few as 30-40 informants (Weller & Romney, 1988, p. 25). On the other hand, since informants are free to come up with different numbers of piles, single pile-sorts have limited utility for comparing individuals. In particular, the "lumper" versus "splitter" variation tends to overwhelm other characteristics that might differentiate informants. Other variants of the pile-sort task, such as multiple sorts and successive pile-sorts, obtain more information per informant and, by imposing uniform constraints, are better suited to comparing informants. Weller and Romney (1988, pp. 20-31) provide an excellent discussion of the strengths and weaknesses of various pile-sort tasks.

Triadic Comparisons

Another way to obtain overall similarity judgments is to present three items at a time, and to ask informants to pick the one that is most different from the other two. The procedure is repeated until each item has been presented in a triad with every pair of other items. To avoid uncontrolled order effects and response biases, however, it is important to randomize the presentations, both among and within the triadic sets. Using this method to obtain similarity judgment among four kinds of fish, for example, we might obtain results like those in Table 4.

Table 4.
One Informant's Triadic Similarity Judgments Among Four Kinds of Fish (Capitalization indicates the informant's "odd item out" judgments)

Triad 1:	BASS	salmon	trout
Triad 2:	tuna	BASS	salmon
Triad 3:	trout	bass	TUNA
Triad 4:	salmon	tuna	TROUT

Like the single pile-sort, such data can be represented as an item-by-item matrix. Each triad involves three pairwise comparisons, i.e.,

ABC breaks down into three pairs: AB, AC, and BC. Thus, for each triad, there are three similarity scores: the pair of items not chosen is judged similar (scored 1), and the two other pairings are judged not similar (scored 0). Following this scoring procedure, Table 5 shows the matrix representation of the information contained in Table 4. (Note: Because similarity data are symmetric, we display only the lower half of the matrix.)

Table 5.
Matrix Representation of Data from Table 4

	Bass	Salmon	Trout	Tuna
Bass	.....	.....	.....	.....
Salmon	0	.....	.....	.....
Trout	1	1	.....	.....
Tuna	0	2	0	.....

Informants are quick to understand this triadic comparison task, whether it is administered in written or oral form. Another advantage of the method is that, because all informants perform the same task, the resulting data enables comparisons among individuals. And as long as there are only a few items in the domain, informants find the task mildly amusing. Unfortunately, as the number of items increases, the number of triads required rises dramatically -- the combination of n things taken three at a time, or $n! / [3!(n-3)!]$ -- and informants quickly lose patience when confronting a large number of triads. For example, 8 items require 56 triads, 10 items require 120 triads, and 19 items require 969 triads.

Although one can use a balanced incomplete block design (BIB) to reduce the number of triads given each informant, this forces the researcher to decide whether to focus on differences among items within the domain or on informant differences with respect to the domain. If the objective is to compare informants, then the same subset of triads should be given to each informant; whereas if the focus is more on the items themselves, then each informant should get a different randomly selected instance from the BIB (see Borgatti, 1996, for a fuller discussion of this point). Thus, practical concerns dictate that the number of items be no more than 8-10 for complete designs, and no more than about 25 for BIB designs (Weller & Romney, 1988, p. 37).

In a complete design, such as the example in Table 4, each pair of items occurs in n-2 triadic sets. For example, salmon and tuna were judged similar in both of the triads in which they occurred, whereas salmon and trout were judged similar in only one of their two co-occurrences. If we had eight items, then each pair would co-occur in six triadic sets, and so forth. Following this logic (and modified appropriately for balanced incomplete block designs), the aggregate similarity for any two items can be expressed as a proportion, i.e., the number of triads in which the two items are judged similar by all informants divided by the total number of triads in which the two items are presented to all informants. Thus, aggregated data from triadic comparisons are similar in form to the results from pile-sorts and amenable to the same kinds of statistical analyses.

Conclusion

My main purpose in this essay has been to explain enough about free-listing, pile-sorting, and triadic comparisons to motivate readers to learn more. I might also mention that medical anthropologists and other applied researchers are increasingly using these techniques (e.g., Boster & Weller, 1990; Garro, 1986; Mathews, 1983; Ryan, Martinez, & Pelto, 1996; Weller, 1984; Weller & Mann, 1997). Although I have not covered current approaches to the statistical analysis of these kinds of data, I will do so in a column to appear in the next issue of the newsletter.

In closing, let me note that these approaches to data collection and analysis are much less laborious than they once were thanks to ANTHROPAC (Borgatti, 1998), a PC-based software package. Whereas tabulating a free-list task with 40 informants used to take hours or days of uninterrupted work, ANTHROPAC reads in text files, tallies items, and computes the appropriate descriptive statistics in seconds. What's more, ANTHROPAC not only reads and analyzes pile-sort and triadic comparison data, but can also generate randomized triads questionnaires using many different BIB

designs. It even includes a variety of analytical tools, such as consensus analysis, multi-dimensional scaling, and hierarchical clustering. To learn more about this worthwhile software package, including pricing, see Analytic Technologies, web page: http://www.analytictech.com/.

References

Bernard, H.R. (1994). *Research methods in anthropology, (2nd Ed.).* Thousands Oaks, CA: Sage.

Bernard, H.R. (Ed.) (1998). *Handbook of methods in cultural anthropology.* Walnut Creek, CA: Altamira Press.

Borgatti, S.P. (1996). *ANTHROPAC 4.0 methods guide.* Natick, MA: Analytic Technologies.

Borgatti, S.P. (1998). *ANTHROPAC, Version 4.95X.* Natick, MA: Analytic Technologies.

Boster, J.S., & Weller, S.C. (1990). Cognitive and contextual variation in hot-cold classification. *American Anthropologist,* 92, 171-179.

D'Andrade, R. (1995). *The Development of cognitive anthropology.* New York: Cambridge University Press.

Frake, C.O. (1962). The ethnographic study of cognitive systems. In T. Gladwin and W.C. Sturtevant (Eds.), *Anthropology and human behavior* (pp. 72-85). Washington, DC: Anthropological Society of Washington.

Frake, C.O. (1964). Notes on queries in ethnography. *American Anthropologist,* 66 (3, Part 2), 132-145.

Garro, L.C. (1986). Intracultural variation in folk medical knowledge: A comparison between curers and noncurers. *American Anthropologist,* 88, 351-370.

Gatewood, J.B. (1983). Loose talk: Linguistic competence and recognition ability. *American Anthropologist,* 85, 378-387.

Gatewood, J.B. (1984). Familiarity, vocabulary size, and recognition ability in four semantic domains. *American Ethnologist,* 11, 507-527.

Goodenough, W.H. (1956). Componential analysis and the study of meaning. *Language,* 32, 195-216.

Lounsbury, F.G. (1956). A semantic analysis of Pawnee kinship usage. *Language,* 32, 158-194.

Mathews, H.F. (1983). Context-specific variation in humoral classification. *American Anthropologist,* 85, 826-847.

Romney, A.K., Weller, S.C., & Batchelder, W.H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist,* 88, 313-338.

Ryan, G., Martinez, H., & Pelto, G. (1996). Methodological issues for eliciting local signs/symptoms/ illness terms associated with acute respiratory illnesses. *Archives of Medical Research,* 27, 359-365.

Weller, S.C. (1984). Consistency and consensus among informants: Disease concepts in a rural Mexican town. *American Anthropologist,* 86, 966-975.

Weller, S.C. & Romney, A.K. (1988). *Systematic data collection:* Qualitative research methods, Volume 10. Newbury Park, CA: Sage.

John B. Gatewood is Professor of Anthropology at Lehigh University. He has roughly 35 professional publications in the areas of cognitive anthropology, fisheries ethnography, tourism studies, and linguistic anthropology. In addition, Dr. Gatewood has considerable experience doing marketing and advertising research for a variety of firms. His current research interests include distributive models of culture, network analysis of organizations, motivations of heritage tourists, and theoretical issues concerning the 'units' of culture.

DOUBLE ISSUE

The BEHAVIORAL MEASUREMENT *Letter*

Behavioral Measurement Database Services

Enriching the health and behavioral sciences by broadening instrument access

Vol. 7, No. 1
Winter 2000

Introduction to This Issue

This issue of *The* Behavioral Measurement *Letter* contains a very large volume of material, so large, in fact, that it is a double issue. And although I was again assisted by our regular contributor, Fred Bryant, there was a related delay in its publication. I hope that both the quantity, and especially the quality of its contents will be adequate compensation for the wait.

In this issue of The BML, Dr. Frank Baker, Director of Research at the American Cancer Society, reviews instruments of various types used to measure the quality of life (QOL) of cancer survivors and research designs for evaluating cancer survivors' QOL, and then discusses challenges and issues in QOL research. Due to the implementation of cancer screening programs and lifestyle changes (e.g., in diet, cigarette smoking, alcohol consumption), improved diagnostic methods, and the use of new and more effective cancer treatments in the last quarter of the 20th century, cancer, still associated with high levels of morbidity and mortality, is no longer the dreaded killer it once was. These advances have resulted in large and growing numbers of cancer survivors, many of whom live for years (even decades) beyond the time of initial diagnosis. Thus, instruments and methods are needed to assess the quality of life of cancer survivors in order to obtain baseline measures, to identify factors that contribute to a good QOL and those that contribute to a poor QOL for survivors, to design and improve ways and means for QOL enhancement, and to determine the effectiveness of attempts to improve cancer survivors' QOL.

Racial, ethnic, and class bias are common, often substantial sources of error variance in measurement. Such bias may be introduced at any stage of instrument development and use, including definition and operationalization of variables to be measured, item construction, and instrument administration. These sources of measurement bias are discussed in a column by Drs. Mildred Ramirez, Marvella Ford, and Anita L. Stewart, from the Research Centers for Minority Aging Research -- Measurement and Methods Cores. They point out that measures administered to various racial/ethnic groups and/or persons of low socio-economic status that do not account for racial/ethnic/class differences can produce results that are not generalizable to these groups. This, in turn, leads to flawed social policies and ineffective services designed using such research. The column strongly reminds us that 1) measurement bias of various types exists and must be addressed effectively to assure validity, and 2) the type of measurement bias due to insensitivity to racial, ethnic, and/or class differences has consequences not only for the corpus of research-based knowledge, but for applications of such invalid knowledge.

Also in this issue is the second of a two-part column, "Culture . . . One Step at a Time," by cultural anthropologist John Gatewood. As the reader may recall, the first part (The BML (6) 2:5-10, Fall 1999) dealt with means to gain an understanding of a culture by learning how persons within the culture view their world. He presented three methods to discover how persons see similarities, differences, and relationships among products of a culture or items in its environment -- free listing, pile sorting, and triadic comparison. In the second

Culture . . . One Step at a Time (Part 2)

John B. Gatewood

Editor's Introduction to Part 2: In the first part of Dr. Gatewood's piece, published in the previous issue of The BML (6(2):5-10), he discussed three techniques by which one can begin to gain an understanding of a culture -- free-listing, pile-sorting, and triadic comparison. In this second of two parts, Dr. Gatewood discusses various techniques to analyze data generated through free-listing, pile-sorting, and triadic comparison.

Consensus Analysis

So far, I have reviewed somewhat unusual methods for collecting data. By contrast, I now consider consensus analysis, an interesting technique that cognitive anthropologists have devised to analyze these kinds of data. Although many of the ingredient ideas have been percolating in various literatures for decades, Romney, Weller, and Batchelder (1986) were the first to propose the formal theory and mathematical procedures of consensus analysis. Here I can do no more than sketch the basics, and I must refer readers interested in more details to the original source.

The essential problem that consensus theory addresses is as follows. Given that members of a culture do not uniformly agree with one another in their beliefs about what is true or proper, how can an outsider tell if there is a 'common culture' underlying their diverse opinions? The key to answering this question lies in realizing that (a) no one knows all of his or her group's culture and (b) agreement is a matter of degree. In particular, experts in a cultural domain should agree with one another more than nonexperts do (see Boster, 1985). Following this intuition, consensus theory assumes that "the correspondence between the answers of any two informants is a function of the extent to which each is correlated with the truth" (Romney et al., 1986, p. 316). Consensus theory focuses precisely on the extent to which informants converge on the same answers to systematically asked questions as a measure of cultural knowledge.

For example, suppose Mr. Smith gives a multiple-choice test to his class, but arriving home discovered that he has lost the answer key. Could he grade the students' answer sheets anyway? Yes, he could (Batchelder & Romney, 1988). According to consensus theory, if students did not know the correct answer to a question, then they would just guess, and such guessing should produce predictable proportions of agreement across the available answers. On the other hand, if students know the correct answer, then they will converge on the same answer (the "correct" one) more frequently than expected just by chance. Knowledge -- cultural competence in a domain -- produces deviations from equiprobability, and more knowledgeable individuals will agree with one another more often than less knowledgeable individuals do.

The ingenuity of consensus analysis is that it provides a way to estimate the cultural competence of individual informants from the patterning of their agreement. The formal model rests on three central assumptions (Romney et al., 1986, pp. 317-318):

1. *Common Truth.* The informants all come from a common culture, such that whatever their cultural version of the truth is, it is the same for all informants.
2. *Local Independence.* Informants' answers are given independently of other informants, i.e., there is no collusion or influence among informants.
3. *Homogeneity of Items.* Questions are all of the same difficulty, such that the informant's cultural competence is equal across all questions.

Certain kinds of statistical criteria serve as checks on the validity of these three critical assumptions, though a detailed description of these technicalities is beyond the scope of this column. If these technical criteria are met, then it is considered reasonable to compute a relative "competence score" for each informant, as a measure of how well that particular individual represents the entire sample's answers to the questions asked. Under these conditions, a competence score can be interpreted as the proportion of questions an informant answered "correctly." Conversely, if these statistical checks on the validity of underlying assumptions are violated, then one or more of

the three critical assumptions must *not* be true of the data. For example, it may well be that there is no common culture within the sample, but rather many different subcultures, or systematically different ways of responding.

Consensus analysis works well with many kinds of data: true-false, check lists, belief-frames, multiple-choice, rankings, ratings, and even proximity matrices (such as similarity matrices). There are, however, two important limitations: (a) the battery of questions must be of a single type, such as all multiple-choice, all similarity matrices, etc., and (b) the questions must ask informants for conventional truths or judgments, *not* their personal preferences or histories. For instance, consensus analysis is well-suited for questions asking informants to check all diseases on a list that are serious, contagious, result in a high fever, or for which one should see a doctor. But consensus analysis makes little sense if informants are asked to check all the diseases that they have actually had in their lives. Likewise, consensus analysis is appropriate for pile-sort data of mixed drinks based on their similarities, but not for pile-sorts based on whether the drinks taste good. The reason is simple: agreement with respect to preferences or histories is not knowledge-driven.

An Integrative Example

Results from a small class project will illustrate how these different methods and analyses can be used in concert. The data come from 14 college students, whose knowledge of a single cultural domain (kinds of fish) was probed via free-listing, single pile-sort, triadic comparison, and several ranking and rating tasks.

Free-listing was the first task. The instructions were to list all the different kinds of fish they could think of in 90 seconds. Collectively, the 14 informants produced lists containing 226 items totaling to 115 different kinds of fish. Table 6 shows the aggregated results for the 43 kinds of fish mentioned by at least two people.

Table 6
The 43 Most Frequently Listed Kinds of Fish (N=14)

Goldfish	10	Steelhead	2
Trout	8	Rock bass	2
Salmon	8	Beta	2
Flounder	7	Great white shark	2
Catfish	7	Bullhead catfish	2
Sunfish	7	Brown trout	2
Bass	7	Brook trout	2
Tuna	7	Pink salmon	2
Piranha	6	Barracuda	2
Shark	5	Bluegill	2
Swordfish	5	Striped bass	2
Rainbow trout	5	Blue catfish	2
Carp	4	Remora	2
Guppy	4	Jellyfish	2
Cod	3	Channel catfish	2
Muskie	3	Dog salmon	2
Red salmon	3	Smallmouth bass	2
Bluefish	3	Clams	2
Largemouth bass	3	Minnow	2
Pickerel	3	Yellowfin tuna	2
Blowfish	2	Scrod	2
Walleye	2		

From the 115 kinds of fish the students free-listed, I chose 19 varieties for further study. Given the large number of items in this cultural domain, single pile-sorting was the most obvious method for obtaining similarity judgments. I could have chosen more items for this task, but I wanted students to see that similar results could be obtained via triadic comparisons. By using a balanced incomplete block design for the triads task and the same items for the pile-sort, the results of both methods would be directly comparable without overburdening my captive students. Thus, students did the single pile-sort task one day and the triadic comparisons the following class period.

Figure 1 is a nonmetric multidimensional scaling of the aggregate similarity matrix from the single pile-sort data, and Figure 2 is from the triads data. (In these plots, the closer items appear to one another, the more similar students judged them to be. The items are displayed in relation to horizontal and vertical axes, whose underlying meaning the researcher must interpret subjectively. "Stress," a measure of

how well the particular configuration fits the data, ranges from 0-1, with lower values reflecting better fit.) Visual comparison of the two figures indicates the two methods produced similar results. Indeed, the Pearson *r* between the two similarity matrices is .74. Normally, we would expect to find greater reliability between these two methods, but bear in mind that we had only 14 students doing the single pile-sort rather than the recommended 30-40. Also, when implementing the BIB (balanced incomplete block design) triads design, I chose to give each student the same subset of triads (so I could better compare informants), rather than using different instances of the design for each student (the domain-focused approach). Given these shortcomings of the class project, the obtained reliability of .74 is not bad.

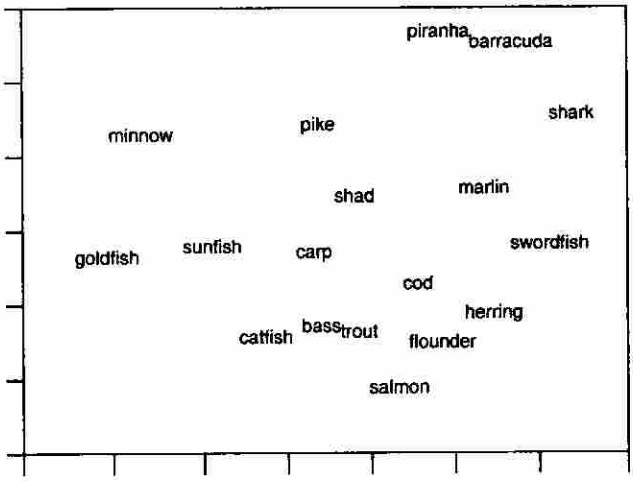

Figure 1. MDS from Single Pile-Sorts (stress = .15)

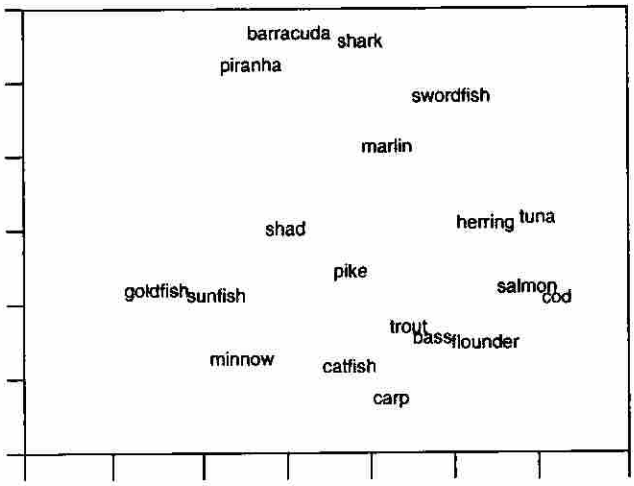

Figure 2. MDS from Triadic Comparisons (stress = .20)

Because I was most interested in using a variety of methods to compare informants with respect to their knowledge of fish, I included three auxiliary rating and ranking tasks after we had completed the free-listing, pile-sort, and triadic comparisons. The first of these asked informants to rate their own knowledge of each of the 19 fish varieties using a 5-point scale, where 1 indicated "never heard of it before" and 5 indicated "know quite a bit." And, because there had been some class discussion concerning kinds of fish following the pile-sort and triads tasks, the second and third auxiliary tasks asked each informant to rate all the students on a 5-point "novice to expert" scale, then to rank order everyone in the class with respect to how much they knew about fish.

In all, the project garnered four direct measures of informants' knowledge of the domain (length of free-list, self-rating of knowledge, social-rating of expertise, and social-ranking of expertise) and three indirect measures (competence scores from consensus analyses). Table 7 presents these informant-level measures.

Note that Table 7 displays three competence scores: one for the pile-sort data, and two for the triads data. This is because there are two ways to handle consensus analysis for the triads task. One can use either (a) the lower half of each individual's similarity matrix, as is done for pile-sort data (resulting in $n(n\text{-}1) / 2$, or 171, "multiple-choice" items), or (b) the non-randomized tallies of the triadic sets that were actually given to each informant (114 "multiple-choice" items, given the BIB design that I used).

Incidentally, all three consensus analyses meet the statistical criteria of the formal model. Hence, the patterning of agreement among the 14 students indicates there is a culturally correct way of doing the pile-sort, and likewise for the triads task. Kevin best exemplifies the sample's "correct" way of pile-sorting (competence score of .89), whereas Jennifer (.67 and .68) and Lisa (.66 and .69) best represent the sample's "correct" responses to the triads task.

The last points I will make from this example concern the correlations among different individual-level measures. Table 8 shows the correlation matrix of all seven measures, which includes a couple of surprising findings. (Again,

Table 7
Informant-Level Measures

Students	Items in Free-List	Mean Self-Rating of 19 fish	Mean Soc-Rating of Expertise	Mean* Soc-Ranking of Expertise	Pile-Sort Comp. Score	Triads Comp. Score	Triads Comp. Score
Josh	34	4.95	5.00	14.00	.78	.43	.38
Matt	30	3.84	3.50	10.00	.32	.41	.37
Sara	26	3.05	2.64	7.79	.80	.62	.63
Judy	21	3.68	3.00	8.07	.74	.61	.62
Corey	19	3.47	3.07	8.21	.55	.56	.63
Hassan	17	4.16	3.93	11.50	.82	.37	.24
Nora	15	2.95	2.43	6.00	.25	.57	.61
Jennifer	13	2.53	1.86	3.71	.03	.67	.68
Kevin	12	3.37	3.50	10.07	.89	.56	.59
Derek	10	4.00	3.43	10.14	.77	.47	.46
Lisa	10	2.84	1.86	2.36	.70	.66	.69
Beth	7	2.79	1.93	4.50	.74	.52	.53
Serene	7	2.84	1.64	3.14	.68	.62	.64
Hope	5	2.58	2.14	5.50	.43	.63	.67
M	16.14	3.36	2.85	7.50	.61	.55	.55
SD	8.60	.67	.93	3.32	.25	.09	.13

Note: *The scores for expertise ranking (fourth column) have been inverted to keep the meaning of correlation coefficients' signs clear in what follows. For example, Josh was uniformly regarded as the most knowledgeable in this domain; hence, his inverted score is 14 rather than 1.

note that we display only the lower half of the matrix because it is symmetric.)

The seven measures form into three logical groupings. First, length of free-list and self-rating of knowledge with respect to the 19 fish are both related to how much informants really know about this cultural domain. But as Don Campbell was fond of saying, "All measures are fallible"; and the correlation between these two is somewhat lower than one would like ($r = .68$). Length of one's free-list is affected by motivation, and self-ratings presume shared understanding of the response scale. Second, the two social evaluations differ only in the manner in which informants are allowed to express their opinions -- rating versus ranking -- and are highly correlated ($r = .98$). And third, because the instructions for both the pile-sorting and triadic comparisons tasks asked for overall similarity judgments, we might expect all three competence scores to be positively interrelated. But as inspection of Table 8 reveals, this is not the case.

The two ways of assessing triadic competence are highly correlated ($r = .97$), as expected, but competence doing the pile-sort task is largely independent of these two measures, and even tends to be *inversely* related to triadic competence ($r = -.27$ and $-.28$, respectively). How can this be? What does it mean? Recall that the two methods did produce convergent results with respect to their aggregate item-by-item similarity matrices (see Figures 1 and 2); hence, the negative correlation does not bear on the question of intermethod reliability. Rather, competence scores indicate the degree to which individual's responses represent the patterning of agreement in the entire sample -- how well an individual "measures" the group consensus. Thus, the "surprising" negative correlation simply means that students whose pile-sorts resembled those of other students tended to be more idiosyncratic in the ways they answered the triadic comparisons, and vice versa.

Another interesting finding in Table 8 is the ability of students to evaluate one another's expertise. That is, the correlations between the social evaluations of expertise and the self-generated indicators of knowledge (i.e., free-listing and self-rating) are quite high ($rs = .67 - .95$). Having been in the classroom and observed the brief occasions when anyone demonstrated knowledge or lack thereof, I marvel at the students' sensitivity to and convergent interpretations of very subtle social clues.

Lastly, perhaps the most puzzling and important finding in Table 8 is the lack of correspondence between the direct measures of informants' knowledge of fish and their cultural competence scores on the pile-sort and triads tasks. Indeed, the correlations between the two types of measures are either nonsignificant or actually negative. The more knowledgeable informants (as determined by the four direct measures) are slightly more typical of the group with respect to their pile-sorts. But contrary to other studies (Brewer, 1995), they are very *atypical* with respect to their triadic judgments. Indeed, for the triads task, the most domain-knowledgeable informants are actually the poorest representatives of the sample's common culture. Their greater knowledge of fish did *not* produce greater agreement; rather the sample's consensus was formed by relatively ignorant informants who agreed among themselves.

This latter finding provides a general caution to those who would use consensus analysis uncritically. There are situations where 'knowledge of the common culture' means being fairly ignorant, and sometimes ignorance produces its own patterning of agreement. For some tasks, less knowledgeable people see only one way of responding, whereas experts are aware of alternatives (see also, Boster & Johnson, 1989). In such cases, the more informants know, the less likely they are to agree with others. As a way of reminding ourselves that such possibilities exist, it might be wise to relabel "competence scores" as "representative scores" in conducting consensus analysis.

Table 8
Pearson Correlation Coefficients Among Informant-Level Measures

	Free-List	Self-Rating of 19 Fish	Soc-Rating of Expertise	Soc-Ranking of Expertise	P-S: Comp	Tri: Comp. 1
Self-Rating	**.68**	– –	– –	– –	– –	– –
Soc-Rating	**.70**	**.95**	– –	– –	– –	– –
Soc-Ranking	**.68**	**.91**	**.98**	– –	– –	– –
P-S: Comp.	.04	.43	.38	.38	– –	– –
Tri: Comp. 1	-.46	**-.80**	**-.78**	**-.79**	-.27	– –
Tri: Comp 2	-.46	**-.79**	**-.75**	**-.76**	-.28	**.97**

Note: **Boldfaced** values are significant at $p \leq .05$, two-tailed

References

Batchelder, W., & Romney, A.K. (1988). Test theory without an answer key. *Psychometrika*, 53, 71-92.

Boster, J.S. (1985). "Requiem for the omniscient informant": There's life in the old girl yet. In J. Dougherty (Ed.), *Directions in cognitive anthropology* (pp. 177-197). Urbana, IL: University of Illinois Press.

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

Brewer, D.D. (1995). Cognitive indicators of knowledge in semantic domains. *Journal of Quantitative Anthropology*, 5, 107-128.

Romney, A.K., Weller, S.C., & Batchelder, W.H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist*, 88, 313-338.

John B. Gatewood is Professor of Anthropology at Lehigh University. He has roughly 35 professional publications in the areas of cognitive anthropology, fisheries ethnography, tourism studies, and linguistic anthropology. In addition, Dr. Gatewood has considerable experience doing marketing and advertising research for a variety of firms. His current research interests include distributive models of culture, network analysis of organizations, motivations of heritage tourists, and theoretical issues concerning the 'units' of culture.