Consistency and Consensus: A Replication

DEVON D. BREWER, A. KIMBALL ROMNEY, and WILLIAM H. BATCHELDER*

School of Social Sciences
University of California at Irvine
Irvine, CA 92717, U.S.A.

ABSTRACT. This article reports a validation of the cultural consensus model (Romney, Weller, and Romney, 1986) and replicates and extends the findings of Weller (1984). In two domains, informants who agreed more with each other displayed greater internal consistency in their responses to a paired comparison task than informants who agreed less with others. Also, greater overall consistency among informants in a domain was associated with more overall agreement among informants in that domain. No relationship was observed between either individuals' cultural competences (amounts of agreement with others) or consistencies across domains. However, males exhibited greater consistency than females.

KEY WORDS: cultural consensus model, informant competence, internal consistency

Much of the basic raw material for anthropological research is generated through informant interviews. As early as 1938 Sapir observed that anthropologists arrived at conclusive ethnographic statements that were presumed true for a given society. He noted that "One was rarely in a position to say whether such an inclusive statement was a tacit quotation from a primitive "John Doe" or a carefully tested generalization abstracted from hundreds of personal observations or hundreds of statements excerpted from conversations with many John Does" Sapir (1938: 7).

All field workers encounter at least some disagreement among informants. The treatment of such disagreement is crucial to the definition of both culture and intra-cultural variability — concepts of central concern in anthropology. This research explores the validity of a recent method developed to: (1) measure informant accuracy and (2) provide an objective way to infer the existence of cultural beliefs. The method is called Cultural Consensus (Batchelder and Romney 1986; Romney, Weller, and Batchelder 1986; Romney, Batchelder, and Weller 1987; Batchelder and Romney 1988) and focuses explicitly upon the degree of agreement among informants.

When informants disagree we are faced with a sequence of critical questions. The initial question is whether our informants share and are reporting on a common underlying cultural tradition. On the assumption

Journal of Quantitative Anthropology 3: 195—205, 1991.

that the informants share the same culture, there remain two fundamental problems: first, what are the culturally appropriate answers to the questions asked by the ethnographer, and, second, how do we assess the competence or accuracy of the various informants concerning the cultural domain under consideration.

Boster (1985) observed that the information for answering these two questions is contained in the pattern of agreement among all possible pairs of informants. Under very general assumptions, informants who agree more with others tend to have more cultural knowledge than informants who agree less with others. This important generalization has been derived as a theorem of the cultural consensus model (Batchelder and Romney 1986; Romney, Weller, and Batchelder 1986; Romney, Batchelder, and Weller 1987; Batchelder and Romney 1988). The model provides, among other things, an estimate of how much each informant knows about a given domain of culture. This estimate is referred to as the cultural competence of the informant. The estimate is based on the amount of agreement that each informant has with each other informant.

Several researchers (Boster 1985; Weller 1984; D'Andrade 1987) have pointed out that the fruitfulness of the concept of competence would be enhanced if the competence of each informant could be validated with some independent information not based on between informant agreement. Boster (1985) provided one important validation when he retested six informants on his "easy manioc" identification task after a two-month interval. He looked at informant reliability by comparing test-retest, within-informant reliability with knowledge estimated from between-informant agreement and found a correlation of 0.92. He concluded that "... informants who agree with others were more likely to agree with themselves" (Boster 1985: 193).

Weller (1984) used a paired-comparison task to measure within-informant consistency and found that more competent informants as judged by between-informant agreement were more consistent (in a sense to be defined below). Weller commented that "This research represents only a preliminary investigation into the quantification of informant expertise across related tasks, and it is simply not known how these findings generalize to other domains" (Weller 1984: 973). In this research report we replicate her findings in two different domains, *animals* (animal domain) and *things students do to get good grades* (student domain). Our findings show substantial agreement with her main results. We also find that neither competence nor consistency transfers across domains, i.e., correlations between measures of competence and consistency are non-significant between the two domains. A secondary, and heretofore unreported, finding also emerged from our study, namely, that male informants display greater consistency than female informants.

Weller (1984) investigated the relationship between internal consistency and several measures of competence in four separate aspects of disease terms among rural Mexican villagers. The aspects of illness she used in her paired comparison tasks were: "(1) contagion (which is more able to pass from one person to another?); (2) severity (which is more severe?); (3) age (which is more an illness of adults and which is more of children?); and (4) the hot-cold concept (which needs hot remedies and which needs cold in order to get better?)" (Weller 1984: 968). She defined an informant's consistency as the number of contradictory or intransitive response patterns in their responses. Intransitive patterns occur when a circular triad is produced. For example, if three items X, Y, and Z are compared in terms of size, one of the two possible intransitive patterns for this set of items is: $X > Y$, $Y > Z$, and $X < Z$. One of the six possible transitive patterns between the same items is illustrated by: $X > Y$, $Y > Z$, and $X > Z$.

Weller found a modest relationship between the average consistency of informants and the degree of consensus, or agreement among informants, within each of the aspects of illness. The associations between competence and consistency on the individual level that Weller observed were somewhat variable with correlations of $r = 0.07$, 0.21, 0.57, and 0.75. The two low correlations ($r = 0.07$ and 0.21) were in dimensions of the domain where the consistency and competence scores exhibited little variability. Furthermore, she observed that competence did not generalize across the several dimensions of the disease terms domain.

With cultural consensus analysis, an individual's competence may be estimated by the degree of concordance with the overall responses of all informants (Romney, Weller, and Batchelder 1986). The degree of consensus among informants in a domain may be measured by the average competence of informants. The culturally "correct" answers are derived from averaging informants' responses weighted by their competences.

The present research examines the relationships between: (1) average within-informant consistency and the average competence in both the animal and student domains (i.e. is there more overall consistency among informants in more coherent domains?); (2) consistency and competence on the individual level (i.e. are more competent informants also more consistent?); (3) competence across domains; (4) consistency across domains; and (5) selected demographic variables with competence and consistency.

DATA COLLECTION

Each informant received a randomized complete paired-comparison questionnaire for 10 items in each of two domains, *things students do to get good grades* and *animals*. The complete paired comparisons format pairs

each item with every other item in the domain (Guilliksen and Tucker 1961) and results in a questionnaire of 45 pairs. The instructions for the student domain asked informants to select the item in each pair that is more important for getting good grades. The instructions for the animal domain asked informants to pick the animal that is larger in overall size.

The items for the student domain were obtained from a free listing task in which 28 undergraduate students in two discussion sections of an introductory sociology course at the University of California, Irvine responded to the question "What are all the things a student can do to get good grades?" The ten most frequently named statements are used in this study.[1]

The items in the animal domain came from Henley's (1969) set of 21 animal terms that were derived from free lists. We chose a subset of 10 animals that were intermediate in terms of overall size, i.e, we discarded the smallest and largest animals.[2] We used animals of intermediate size to make the ranking task more difficult since they are the most similar to each other in terms of size.

Data from four separate and independent samples of student informants from University of California, Irvine, were collected. Samples A and B had one randomization of the questionnaire, while samples C and D had a different randomization. Sample A ($n = 17$) consisted of self-identified White (also reporting English as a first language) students and sample B ($n = 16$) consisted of self-identified Asian (also reporting some Asian language as a first language) students, while samples C and D (both $n = 16$) were selected without reference to any demographic variable.[3] Informants from all four samples answered questionnaires on both the animal and student domains.

We chose four groups in order to see how stable our results would be from one group to another. In practice, the average anthropologist would probably not interview groups larger than about 16, so the finding should generalize to small groups of this size. Finally, we picked a comparison group whose first language was Asian, to see whether the effect was specific to English speakers.

RESULTS

Each informant's responses in each domain were converted into a 10 × 10 square matrix whose rows and columns represent the 10 items in the domain. A "1" was placed in a cell if the column item was viewed by the informant as more important (in the student domain) or larger in size (in the animal domain). In contrast, a "0" was placed in a cell if the row item was selected as more important or larger.

Using each informant's matrix in each domain, consistency was calculated. Consistency is defined as the number of intransitive response patterns (or circular triads) made by each informant. Fewer circular triads indicate higher consistency.

The square data matrices from the dichotomous model described above could not be directly run through cultural consensus analysis for dichotomous items because such data violate the axioms of the model. The pervasive, though not complete, pattern of transitivity in paired comparison data of this sort violates both the local independence and homogeneity of items assumptions of that model (Batchelder and Romney 1986). Whether an informant judges item A as more important (or larger) than item B is linked to the item's relationships with the other items. Thus, each paired comparison is not wholly independent of the other comparisons. Similarly, if an informant judges $A > B$ and $B > C$, the decision about $A > C$ is probably a relatively easy one to make, illustrating that the paired comparison items also vary in degrees of difficulty.

Therefore, to perform cultural consensus analysis, we used the rank model of Romney, Batchelder, and Weller (1987: 169—174). This model is a *data* model rather than a *process* model as derived in earlier papers (Batchelder and Romney 1986; Romney, Weller, and Batchelder 1986). The rank model simply generalizes the structural equations of the earlier models to accommodate rank order data. To prepare the data for the rank model, the square data matrices were transformed into an analytically acceptable form. For each matrix, column marginals were computed by summing all the "1's" (indicating a more important or larger item) a column item received. These marginals can be interpreted as approximate values of an item's relative importance or size and were analyzed as ranking data via cultural consensus analysis, without violating any assumptions of the model.

We can summarize the above by saying that individuals' competences, coherence of the domains, and estimated "culturally correct" answer keys to both domains were calculated via consensus analysis. Competence is related to the degree an individual's responses are in agreement with the group. A domain's degree of consensus is measured in this study by the mean competence of the individuals. The overall fit to the model is judged by the ratio of the first eigenvalue to the second.[4] The "culturally correct" answer keys (in this case, the consensus rankings of both domains) are estimated by pooling all individuals' answers, and weighting the answers by the competence of individuals.[5] Table I contains relevant summary statistics for consensus and consistency.

A comparison of the two domains provides a first aggregate test of our hypothesis that competence and consistency are associated. The animal domain shows higher average competence and higher consistency (fewer transitive errors) than does the student domain. Tests on the significance of the differences were made on data aggregated over all samples for

TABLE I
Basic statistics on competence and consistency for the four samples in the student and animal domains.

sample	n	1st/2nd factor	competence mean	s.d.	errors mean	s.d.
Student domain						
A	17	4.8	0.70	0.29	5.53	5.34
B	16	3.7	0.69	0.23	5.56	4.36
C	16	2.3	0.36	0.42	8.25	6.50
D	16	2.9	0.62	0.20	8.31	5.86
Total	65	4.0	0.60	0.31	6.89	5.78
Animal domain						
A	17	9.4	0.88	0.07	1.06	2.10
B	14	6.8	0.81	0.16	2.07	3.55
C	14	5.6	0.84	0.11	3.43	5.30
D	16	11.8	0.88	0.09	1.56	2.09
Total	61	10.4	0.86	0.11	1.97	3.56

each domain. The mean estimated competence for the animal domain ($n = 61$, mean = 0.86, s.d. = 0.11) was significantly greater than the mean competence of the student domain ($n = 65$, mean = 0.60, s.d. = 0.31) at the $p < 0.0001$ level. The consistency analysis showed significantly fewer intransitive response patterns for the animal domain ($n = 61$, mean = 1.97, s.d. = 3.56) than the student domain ($n = 65$, mean = 6.89, s.d. = 5.78) at the $p < 0.0001$ level.

More important, the association between competence and consistency is also evident on the individual level within each domain. Table II shows that six of the eight samples (four student and two animal) show negative correlations between individual competence and number of intransitive response patterns.[6] Due to small sample sizes and low variances we do not expect all groups to reach significance. While only one of the student samples was significant by itself, the aggregated samples for both domains were significant (both $p < 0.01$). Further evidence of the association between competence and consistency was provided by combination of z-scores from the transformed correlations of each independent sample (see Fisher 1948). The combination of the z-scores involves summing the transformed z-scores and dividing by the square root of the number of independent tests (Mosteller and Bush 1954). The cumulative z-score for the student domain indicated a significant relationship ($p < 0.05$) between competence and consistency, although the animal domain only approached significance ($z = -1.008$).[7]

Weller (1984) raised the question as to whether individuals who were

TABLE II
Pearsonian correlations between individual competence and number of transitive errors for each sample for the student and animal domains.

Sample	Correlation between competence and transitive errors		Transformed z-scores
Student domain			
A	−0.26		−1.010
B	−0.48[a]		−1.875[a]
C	−0.23		−0.829
D	−0.20		−0.721
Total	−0.30[b]	aggregate Z	−2.218[a]
Animal domain			
A	0.06		0.224
B	−0.33		−1.128
C	−0.38		−1.327
D	0.06		0.216
Total	−0.34[b]	aggregate Z	−1.008

[a] $p < 0.05$.
[b] $p < 0.01$.

competent on one type of ranking would be competent on other types. She found no carryover from one task to the other. Similarly, our data do not show a significant relationship between individuals' competences across the two domains. The correlations between competence in the animal domain and competence in the student domain in the four sample were −0.14 (*ns*), −0.19 (*ns*), 0.57 ($p < 0.025$), and −0.24 (*ns*) for samples *A*, *B*, *C*, and *D*, respectively. For the aggregated samples, the correlation between competence across domains was 0.12 (*ns*). Likewise, the combined z-score of the transformed correlations from the four samples was 0.04 (*ns*). Similar results were observed for the association between individuals' consistencies across domains. The correlations between consistency scores in the animal domain and errors in the student domain were −0.04, 0.45, −0.05, and −0.16 (all *ns*), for samples *A*, *B*, *C*, and *D*, respectively. Though the correlation between informants' student and animal consistency in the aggregated samples was significant ($r = 0.31$, $p < 0.01$), the cumulative z-score of the four samples' transformed correlations did not approach significance ($z = 0.73$, *ns*).

There was no association between competence or consistency with ethnicity, academic class level, or grade point average. However, we observed gender differences in consistency. Because of small numbers and unequal proportions of males and females in the four samples, we were

able to test for gender differences only by aggregating all cases into a single pooled sample with 19 males in total in each domain. Males showed significantly fewer ($t = -3.754$, $df = 51$, $p < 0.001$) intransitive response patterns (circular triads) than females in the student domain. Although the differences were not significant in a t-test in the animal domain ($t = -1.671$, $df = 59$, $p = 0.1$) the differences were in the same direction. No significant gender differences, however, were observed in competence (student: $t = 0.093$, $df = 45$, $p = 0.93$, animal: $t = 1.068$, $df = 45$, $p = 0.29$, showing that competence was not an intervening variable between gender and consistency.

DISCUSSION

The results from this study offer support to Weller's (1984) finding that consistency and competence are moderately associated at both the individual and aggregate (domain) level. While Weller compared the consistency-competence relationship on several dimensions of a single domain (illness), the present study compared two different domains. The results demonstrated that informant responses display greater consistency in the more coherent domains. In other words, in a domain where agreement among informants is high, informants as a whole make fewer transitive errors. Likewise, a similar relationship exists on the individual level.

The aggregated samples for both domains demonstrated significant negative correlations between competence and the number of circular triads. The significant cumulative z-score of the consistency by competence transformed correlations in the student domain and the cumulative z-score in the animal domain in the same direction give further support for the link between consistency and competence. Hence, in a given domain, more competent individuals are also more likely to be more consistent than low competence individuals. We interpret inconsistency in responses to be a by-product of lack of knowledge in a domain. From this perspective, when informants are not sure of or do not know an answer, they guess, and informants' guesses are more likely to be inconsistent with their other responses.

In this study, competence in one domain did not correlate with competence in the other. This seems to be in accord with D'Andrade's assertion that "... there is a major division of labor in who knows what" (D'Andrade 1981: 180). Competence in one domain does not necessarily carry over to competence in another. Moreover, the relationship between consistency in different domains is not clearly exhibited in this study.

It was also found that male informants exhibited greater consistency than females, even though no significant gender differences existed in competence. This is a little surprising since it was long ago noted that

women tended to have a higher correlation between performance and aptitude than men. This may be considered one kind of consistency. For example, Romney (1950) reported correlations between English achievement in a college class and aptitude as measured by the ACE for 566 males ($r = 0.69$) and 519 females ($r = 0.84$). Further research would be required to determine if the gender effect on consistency is stable and if so, whether cultural, biological, or other reasons account for the difference.

Further research relating consistency and competence is needed to conclusively confirm the findings of Weller (1984) and the present study. To increase the external validity of these studies, the consistency-competence relationship should be compared among several additional domains. This would indicate whether the consistency-competence relationship (individual and aggregate) is specific to certain types of domains.

NOTES

* We collected the data for this paper in Winter, 1990 with the cooperation of Shampa Mazumdar and Jim Egan. We also would like to thank Kathy Brewer for help in data coding and typing the manuscript and Mike Migalski for computing assistance. Linton Freeman gave helpful comments. Research on consensus theory has been supported by NSF Grant No. SES-8320173.

[1] The 28 informants gave 328 responses to the free list task, with a mean of 11.7 responses per informant. Although a few of the 328 responses were identical, most were different in some way, thus the list was reduced and standardized in a fashion similar to Romney, Smith, Freeman, Kagan, and Klein's (1979) procedures, so as to make the questionnaire manageable. The criteria for reducing the number of statements were to eliminate those statements that:

(1) pertained to only one kind of class or field of study;
(2) referred to ascribed characteristics for student success;
(3) were vague and/or ambiguous.

The criteria for standardizing the statements were:

(1) put in standard grammar and spelling;
(2) simplify responses such that unnecessary articles are eliminated and each statement is limited to a single idea;
(3) put in active tense and use unconjugated verbs, e.g. "go to class always;"
(4) replaced "professor" and/or "TA (teaching assistant)" with "instructor";
(5) replace "discussion section" and/or "lecture" with "class";
(6) replace "tests", "quiz", etc., with "exam".

These procedures yielded 212 separate behavioral statements, or "behavior tags" (Roberts 1964). At this point, very similar statements that differed only in the exact wording were combined, e.g. "attend class always" and "go to class always." The wording of the final combined statement was that of the statement that had been most frequently listed by informants. These statements and their relative frequencies are: go to class always (0.75), keep up with the reading (0.54), ask questions (0.54), form study groups (0.36), review class notes (0.36), go to office hours (0.32), manage time wisely (0.32), go to a tutor (0.29), pay attention in class (0.29), and get enough sleep (0.29).

[2] The ten terms selected are: tiger, lion, zebra, gorilla, deer, antelope, goat, sheep, chimpanzee, and dog.

[3] All White informants reported English as a first language and identified ethnically as White or Caucasian. All Asian informants in sample B reported a first language other than English and listed their ethnicity as some Asian nationality or Asian-American group. Aside from these intended ethnic and language differences between samples A and B, no significant demographic differences were found among the four samples in terms of gender, major, ethnicity, first language, G.P.A., or class level.

[4] A 3:1 ratio of the first eigenvalue to the second is usually considered adequate (Romney, Batchelder, and Weller 1987: 174). Although samples C and D in the student domain have low ratios of the first to second factors (see Table I), their estimated answer keys are extremely similar to other samples keys for the student domain.

[5] The following table shows the aggregate rankings, using the rank model, of each domain (the rankings within samples differ only slightly from these aggregates).

Aggregate Rankings

Student domain	Animal domain
1. pay attention in class	1. gorilla
2. go to class always	2. zebra
3. keep up with reading	3. lion
4. review class notes	4. antelope
5. manage time wisely	5. tiger
6. get enough sleep	6. deer
7. ask questions	7. sheep
8. go to office hours	8. goat
9. form study groups	9. chimpanzee
10. go to a tutor	10. dog

The reliability of these ranking keys is further demonstrated by results obtained from two other samples of informants. Two separate samples of 14 student informants at California State University at Long Beach directly ranked the items in the student and animal domains, respectively, and the answer key rankings produced were very similar to the ones reported above.

[6] In the animal domain, sample A and D's correlations between animal consistency and animal competence ($r = 0.06$ for both) may be distorted by the low number of errors (mean = 1.06 and 1.56, respectively) and lack of variation in intransitive response patterns (s.d. = 2.10 and 2.09, respectively) and in competence (s.d. = 0.07 and 0.09, respectively).

[7] When Weller's (1984) early data-based "competence" measures were calculated using the square data matrices, we observed higher correlations between consistency and "competence" for the eight samples (from −0.18 to −0.63) which were commensurate with Weller's findings.

REFERENCES

Batchelder, W. H. and A. K. Romney 1986 The Statistical Analysis of a General Condorcet Model for Choice Situations. In Information Pooling and Group Decision Making. B. Grofman and G. Owen (eds.) Pp. 103—112. Connecticut: JAI Press, Inc.

Batchelder, W. H. and A. K. Romney 1988 Test Theory Without and Answer Key. Psychometrika 53: 71—92.

Boster, J. S. 1985 Requiem for the Omniscient Informant: There's Life in the Old Girl Yet. In Directions in Cognitive Anthropology. J. Dougherty, Ed. Pp. 177—197. Urbana: University of Illinois Press.

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

D'Andrade, R. G. 1987 Modal Responses and Cultural Expertise. American Behavioral Scientist 31: 194—202.

Fisher, R. A. 1948 Statistical Methods for Research Workers. New York: Jafner Publishing Co., Inc.

Guilliksen, G. and L. R. Tucker 1961 A General Procedures for Obtaining Paired-Comparisons from Multiple Rank Orders. Psychometrika 26: 173—183.

Henley, N. M. 1969 A Psychological Study of the Semantics of Animal Terms. Journal of Verbal Learning and Verbal Behavior 8: 176—184.

Mosteller, F. and R. R. Bush 1954 Selected Quantitative Techniques. In G. Lindzey, ed. Handbook of Social Psychology: Volume 1: Theory and Method. Cambridge, MA: Addison-Wesley Publishing Co., Inc.

Roberts, J. M. 1964 The Self-Management of Cultures. In Explorations in Cultural Anthropology, W. H. Goodenough, ed., pp. 433—454, New York: McGraw-Hill.

Romney, A. K. 1950. The Kuder Literary Scale as Related to Achievement in College English. Journal of Applied Psychology. 34: 40—41.

Romney, A. K., T. Smith, H. E. Freeman, J. Kagan and R. E. Klein 1979 Concepts of Success and Failure. Social Science Research 8: 302—376.

Romney, A. K., S. C. Weller and W. H. Batchelder 1986 Culture as Consensus: A Theory of Culture and Informant Accuracy. American Anthropologist 88: 313—338.

Romney, A. K., W. H. Batchelder and S. C. Weller 1987 Recent Applications of Consensus Theory. American Behavioral Scientist 31: 163—177.

Sapir, E. 1938 Why Cultural Anthropology Needs the Psychiatrists. Psychiatry, 1: 7—12. Reprinted in Selected Writings of Edward Sapir. D. G. Mandelbaum, ed. Pp. 569—577. Berkeley: University of California Press.

Weller, S. C. 1984 Consistency and Consensus among Informants: Disease Concepts in a Rural Mexican Village. American Anthropologist 86:966—975.