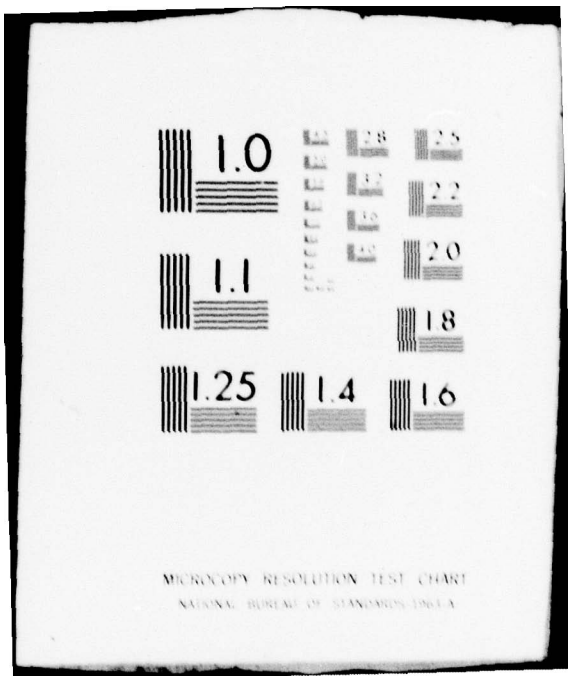
AD-A074 847 WEST VIRGINIA UNIV MORGANTOWN F/G 5/11
THE REVERSE SMALL WORLD EXPERIMENT II.(U)
SEP 79 H R BERNARD, P D KILLWORTH, C MCCARTY N00014-75-C-0441
UNCLASSIFIED BK-119-79 NL

| OF |
AD A074847

END
DATE FILMED
11-79
DDC


1.0
1.1
1.25
1.4
2.8
2.5
2.2
2.0
1.8
1.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: (14) BK-119-79	2. GOVT ACCESSION NO. (11)
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): (6) The reverse small world experiment II.	5. TYPE OF REPORT & PERIOD COVERED: (9) Interim rept.
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): (10) H. Russell Bernard, Peter D. Killworth and Christopher McCarty	8. CONTRACT OR GRANT NUMBER(s): (15) N00014-75-C-0441 2000 1
9. PERFORMING ORGANIZATION NAME AND ADDRESS: H. Russell Bernard, Dept. of Anthropology University of Florida, Gainesville, FL 32611	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: ONR-Code 452 Arlington, VA 22217 (11)	12. REPORT DATE: 1 Sept 1979
	13. NUMBER OF PAGES
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): West Virginia University Morgantown, WV 26505 (12) (52)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

LEVEL

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report describes an informant-defined experiment called "The reverse small world," or RSW. Our earlier RSW, we presented informants with a list of over 1200 potential targets in a small world experiment. Information was provided on targets' occupation, location, sex and race/ethnicity. Informants (or "starters" in a small world experiment) provided the name of a friend, relative, or acquaintance whom they felt would be most likely to know the target, or to know someone who might know the target, and so on. In the present experiment, informants were allowed to inquire about any aspect of a

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

374 100

JOB

20. target's life before selecting an intermediary. In addition, informants provided us with information about their choices. This allows us to test our earlier model of social structure, which we developed as a result of RSW. And it allows us to make much stronger statements about the global structure of (cognitively defined) communications in the U.S.

I. *Introduction*

In a recent paper (Killworth and Bernard, 1978, hereafter referred to as RSW) we began an experimental investigation of social structure. For our purposes, understanding social structure requires two essentially different kinds of information. First, we need to know, on average, how many people are known to each person in a group (such as the U.S.), and who they are. This would provide a *description* of social structure. Second, we need to know *how* people think they are related to the people they know. This would provide an *explanation* of the description. The small-world literature, dating from Pool and Kochen in 1958 (but published in 1978) and Milgram (1967) has advanced our knowledge of the numbers of people in a person's network, and of the degree of connectedness of individual networks. However, except for some subtle work by Lin et al. (1978), there has been very little investigation of how it is that individuals perceive that they are connected to the people in their network. (For a review of the small-world literature to date, see Bernard and Killworth, 1979.)

This led us, in RSW, to investigate experimentally how and why people think they know each other. In our experiment (the reverse small-world) we presented 58 informants ("starters" in the literature) with a long list of fictional people ("targets"). For each target, we provided some basic information: name, race/ethnicity, location and occupation. Starters provided the name of a choice whom they felt would be most likely to know the target, or to know someone who *might* know the target, and so on. In other words, informants provided the name of someone in their own network who would serve as the first link in a small-world chain to the target. Informants also provided some information about their choices. They told us their relationship with the choice (relative, friend, or acquaintance), and they checked a list of reasons for selection of a choice. The reasons were location, occupation, race/ethnicity and "other." For example, if an informant said a choice was picked on the basis of location, then something about the location of the target and/or the choice were somehow connected in the informant's mind.

Overwhelmingly, location and occupation were the important reasons for choice. Characteristics of informants (except for their sex) had little effect on the type of, or reasons for choices. However, characteristics of targets were highly correlated with both type of, and reasons for choices. For example, the most likely reason for choosing an intermediary for a given target could be predicted 81% of the time, based on the target's occupation and distance from Morgantown, West Virginia, where the experiment took place.

RSW yielded a lot of valuable information, but had two serious shortcomings. First, we had very little information about the choices. We knew their names (and thus, in most instances, their gender), and whether they were relatives or friends of the informant. Second, the RSW instrument was closed-ended; it provided only a few selected pieces of information about each target (as in traditional small-world experiments), and forced informants to choose intermediaries and provide reasons for their choices, based solely on these pieces of information.

Informants' comments about the RSW experiment revealed both an occasional need for more information, and that a connection between a choice and target could be very indirect. For example, some informants asked about the religion

of the targets; many informants wanted to know the sex of targets whose exotic or foreign names concealed this piece of information.

Quite often, choices selected on the basis of location did not live (and had never lived) anywhere near the target. Nonetheless, on these occasions, informants insisted that the choice was associated with the target's location. The choice's children, for example, might have gone to college in the target's home state.

The comments by our informants regarding the indirectness of such associations were convincing. We attempted to build both direct and indirect associations into a model of the process by which informants made their choices (Killworth and Bernard, 1979). In order to test this model, we assumed that each link in a small-world chain belongs to one of a discrete set of classes or states in a Markov process. (We do not assume that the decision-making process is Markovian, only that the mechanics whereby the next choice is made are independent of the history of the small-world chain.) Many of the transition probabilities had to be guessed, lacking data about them, or even confirmation that all the states in our model existed. It could be argued, therefore, that the good fit between the model's predictions and known facts generated by small-world experiments was fortuitous or self-tuned.

In order to improve the credibility of such a model, we need to know what, if any, information informants need about a target (aside from location, occupation, sex, and race) to make their best choice. And we need to know how informants actually make theif choice, once they are armed with a collection of facts about a target. In order to study this, an open-ended reverse small-world experiment was conducted. We consider this a member of a genre we call INDEX, or "informant defined experiment." The idea is to study social structure experimentally, but to allow the subjects of the study to define the information which is collected.

We turn now to a description of the experiment, followed by a discussion about the coding of the data. The analysis of the data is organized as follows. There is a natural division into problems connected with one or more of: questions asked, choices made, reasons given, together with information about informants and targets. These topics are analyzed first singly, then in combination with one another, where appropriate, in order to find out the relationships between the variables.

II. *The Informant-defined Reverse Small-World Experiment*

A list of 50 mythical targets was constructed. Each target was given a name (and, therefore, gender), an occupation, a location, and a racial identity, as in traditional small-world experiments. Occupations were selected from the Duncan scale (Duncan, 1961) to represent a cross-section of life in the U.S. Three housewives were included, and were assigned husbands with occupations;[1] three students were included, and were assigned both fields of study and part-time jobs; three retired persons were included, but were assigned occupations prior to their retirement.

Location was rather more complicated. We divided the U.S. into six categories of location: 1) near-urban (i.e., Morgantown, WV); 2) near rural (i.e., the surrounding county); 3) "medium" urban (i.e., cities[2] within a 250 mile

radius of Morgantown); 4) medium rural; 5) "far" urban (i.e., cities further than 250 miles from Morgantown); and 6) far-rural. The first two categories were assigned five targets each; the other four were assigned 10 targets each. Five Black targets and 45 White targets were defined. Twenty-five males and 25 females were included.

In addition to these usual identifiers in small-world experiments, we assigned some additional information to the targets, based on informants' comments during RSW. Targets were assigned ages, ranging from 20-70 years; a religion; an education level, ranging from grade school to graduate degree (in six gradations); and a marital status. Table 1 summarizes the socioeconomic characteristics initially assigned to the 50 targets.

We explained the reverse small-world procedure to a group of six pretest subjects, and asked them to select a choice for each target. However, they were given no information whatever about any target -- not even a target's name. Informants were told that they could ask for any information they felt they needed about any target in order to make a choice of intermediary. This pretesting revealed that targets' organizational affiliations and hobbies were frequently requested by informants. The instrument was modified and targets were assigned a maximum of five hobbies and five organizations each. Several pretest informants asked whether targets were active or not in religion; and whether targets had children, how many, and of what ages. This information was added to the "personal history" of the targets.

Fifty informants provided us with the data reported in this paper. Informants were solicited by advertising, and were offered $20 each for their participation. Interviews lasted, on average, 2.5 hours. Table 2 summarizes some characteristics of our informants.

The reverse small-world procedure was explained to informants. We told them that we had complete life histories of 50 people from around the U.S., but that targets' names and characteristics had been shuffled in order to protect anonymity. Targets were presented to informants in random order. (We shall use the term "sequence" to mean when, from one to fifty, a target was presented to an informant.) Informants were to ask us questions about each target, until they felt able to make a choice. Informants were asked initially to explore any avenue they felt might be helpful, and to eliminate questions they found to be of no help as they went along. Many informants had difficulty in comprehending the task of matching a network member to a target. These people had to be taught how to "play the game." We used non-explicit examples for illustration, in order to avoid biasing the informant: "After you have asked questions you will have a set of information about the target. Try to think of associations (whatever you think an 'association' is) between the target and a friend or a relative, or an acquaintance of yours. You will probably want to pick the person you know who is somehow 'associated,' by your definition, with the target." With a few informants it was even necessary to illustrate the concept of the small-world problem graphically. This was simply not an easy experiment to conduct; in collecting these data we may have channeled our informants' behavior in subtle ways which we have no way to control for.

Figure 1 shows how informants rapidly adjusted to the experimental procedure, usually in the first seven targets. By about the eighth target, informants settled down to asking a reasonably steady number (though not type) of

questions. However, in spite of our exhortation to eliminate unhelpful questions, many persons continued to ask questions throughout the experiment which (by their own claim) were never helpful. (Periodically, we asked informants why they kept asking questions which they rated as unhelpful. The typical response was that they were developing a "feel for the target," which allowed them to "exclude" many choices from their consideration, thus making the task of choosing more efficient.)

Of course, we did not actually have complete life-histories for each target. The life histories were incremented as the experiment progressed. Whenever an informant requested information about a particular target which was not already contained in that target's dossier, either the informant was told that the information was not obtainable, or the information was made up on the spot. In the latter case the relevant information was added to the target's dossier for potential use by later informants. This resulted in some inconsistent target characteristics. (For example, one target wound up with a father who had two distinct birthplaces.) If a piece of information was added to a target on the tenth informant, this meant that the previous nine informants had not requested the information. An example of a target's dossier, after 50 informants, is shown in Table 3.

Pretesting revealed many questions which informants might ask. This was subject to some necessary concatenations. One informant asked if one target played the guitar. We interpreted this as a request for information about the target's hobbies, and we told the informant so. Some informants asked questions at the beginning of the interview which suggested that they did not understand their task. Such questions as "What color is the target's hair" or "What kind of car does the target drive" were handled by giving the informant an answer and letting him judge the information's usefulness. In most cases, informants needed further explanations and stopped asking such questions. Subsequently, the questions were not recorded. If an informant insisted that a piece of information was useful, however, it was recorded as a question. An example of this was an informant who asked the exact birth date of a target. When presented with a birth date she proceeded to make a choice on the basis of astrological signs. This question was then recorded as "used."

Each question was assigned a unique identifying number, with no connotation as to order. As new questions were asked in the actual experiment, each was assigned a number. Table 4 presents a list of all questions asked during pretest and test phases. For later reference, note that question 3 refers to target's occupation, and question 14 to target's location.

For each target the procedure was as follows: as each question was asked, its code number was recorded, preserving the sequence in which they were asked. When informants had asked enough questions, they stated their choice, defined to be someone who would act as an intermediary in the small-world process. Then they provided a "few sentences" which explained why they had selected a particular intermediary (i.e., "because he's a real estate agent," or "because his girl friend's father is a pharmacist," or "because she was a graduate student at -- and because he (target) would have been there two years ago when she (choice) left.") Next, informants ranked the questions they asked by the degree to which the answer had helped them make their choice. Each informant was required to select a first-ranked question for each target, and was given the opportunity to rank

other questions (if he or she had asked more than one) second, third, fourth, or fifth, stopping as they felt appropriate. We reminded informants of all questions they had asked (but had not ranked), and inquired whether each had been "helpful" or "unhelpful." Thus each question asked for each target was accompanied by a code indicating its degree of usefulness. The relationship (friend or relative) of each choice to the informant was also recorded.

After completing the test, each informant answered a questionnaire. This consisted of basic socioeconomic data, and a personal response to any question ever asked by the informant about any target. For example, if the informant ever asked where a target's spouse went to school, then (unless the informant were single) he or she provided equivalent information about his or her own spouse.

III. *Coding*

The experiment yielded four essentially different sets of data: 1) information (which we had created) about 50 targets; 2) information about informants; 3) information concerning the questions asked by informants; and 4) information about the informants' choices and those choices' relationships to the targets.

The target data were coded first, since they contain more information than the equivalent informant information. The informant data contain less information because informants were not asked to provide data about themselves on questions they never asked. By the end of the experiment, the known answers to each of the 98 questions ever asked about any target (Table 4) were coded in a format which left room for every possible answer. Informant data were then coded using the same format as for targets.

As noted above, questions were coded in the sequence asked by informants, followed by a ranking of each question's usefulness, which was also provided by the informant.

Finally, we developed a scheme to code the information about choices' relations to targets. This information was contained in the short (usually one or two sentences) explanations given by informants on why they made a particular choice. Four concepts were introduced, the "direct hit," the "associated hit," the "via," and "the intervening choice." If an explanation revealed that a characteristic of a choice matched exactly to a characteristic of the relevant target, this was a direct hit. For example, if a target lives in Los Angeles and the choice for that target also lives in Los Angeles, then this counts as a direct hit. If, on the other hand, a target lives in Los Angeles and the choice lives in San Francisco, then if, and only if, the informant said he selected the choice on the basis of location, this counts as an "associated hit." Associated hits can occur for a wide variety of reasons. If an informant says he chose a pharmacist in order to get to a physician because "they are both in the medical field," then this is an associated hit. Similarly, a farmer and a tractor salesman may be associated by occupation; a student choice may be associated with a college administrator; a choice who plays jazz trumpet as a hobby may be associated with a target who collects jazz records, and so on. The concept of "associated location" and "associated occupation" were introduced in our earlier model of the small-world decision process (Killworth and Bernard, 1979). Our experience in this experiment has broadened the concept to include associations such as hobbies, organizations, religions, etc.

In fact, our experience with these data has shown that simple associations are not enough to describe all the relationships which informants claim exist between their choices and the targets. This led to the "associated via" and "intervening choice" categories. Consider the case of a choice who is a coal miner linked, by an informant, to a target who lives in Kentucky. The coal miner choice may, in fact, live in Ohio. But if the informant says "I chose him because he is a coal miner and he could contact people in Kentucky where there are lots of coal miners," then we believe this is best described as "associated with target's location via choice's occupation." Some other examples include the following: "I chose her because she belongs to the Sierra Club and the target works for the Environmental Protection Agency," then this counts as "associated with target's occupation via choice's organizational affiliation." "I chose him because he does cross country skiing and the target lives in Vermont" is coded as "associated with target's location via choice's hobby." "I chose him because he collects rocks and the target is a geology student" is coded as "associated with target's field of study via choice's hobby."

Finally, many of our informants were apparently thinking two steps into the small-world problem when they said such things as "I chose him because his girlfriend worked at Kroger's grocery and the target owns a grocery store." This counts as "associated with target's occupation via intervening choice's occupation." The choice was not associated with the target by any characteristics of his own; but his girlfriend (whom the informant may not have known well enough to name as his choice) is associated with the target's occupation. For simplicity, we code the fact that the girlfriend is an intermediary choice, and that she is somehow associated with the target's occupation. Another example is the following: "I chose her because her father used to be a professional pool hustler. He could contact the target who likes to play pool." This was coded as "associated with target's hobby via intervening choice's occupation."

IV. *Questions*

As Table 4 shows, a total of 82 different questions were asked by informants. (This does not include six questions which were asked only once, each by one informant.) Obviously, some questions were asked more frequently than others. Figure 2 shows the probability that the most frequent questions are ever asked. Note the dominance of questions 3 and 14, occupation and location respectively, asked 92% and 90% of all occasions. Other questions were asked much less frequently. The most commonly asked of these, are questions 2 (age of target, asked 42% of the time), 29 (sex of target, asked 36% of the time), 30 (marital status, 24%), and 21 (hobbies, 21%). These probabilities can also be interpreted as a fractional contribution to the total number of questions ever asked, with 3 and 14 at 19% each, contributing 38% of all questions ever asked.

The final curve in Fig. 2, shows the cumulative effects of the contributions. Four questions (3-occupation, 14-location, 2-age, 29-sex) account for more than 50% of all questions ever asked. Ten questions account for more than 75% of all questions every asked; eighteen questions account for 90%; twenty-five questions account for 95%.

Figures 3a-3c show similar curves, restricting attention respectively to cases when questions were declared by informants to have been the most helpful, at all helpful, or unhelpful to them in making a choice. Figure 3a shows that

two questions (occupation and location) account for 64% of all "most helpful" (i.e., top ranked) questions. When questions 21 and 26 are added, over 75% are accounted for. Question 16 accounts for another 5%; question 4 accounts for 4%; and then the curve drops off.

The picture is changed subtly when we consider questions which are graded as at all helpful (not necessarily first-ranked) by informants (Figure 3b). Again, location and occupation dominate. However, eight questions are required in order to account for "all helpful questions." It is perhaps surprising that the distribution of questions graded as "unhelpful" is largely the same as for those graded as "helpful" (Figure 3c). This suggests that people tend to ask the same questions about all targets.

The number of questions asked by informants varied greatly, as shown in Fig. 4. The mean number of questions in a string was 4.8, s.d. 2.7; but note that one informant once asked a string of 21 questions before making a choice. The mean number of questions asked by informants differs significantly** between informants, from 1.4 to 9.6. (Henceforth, single asterisks denote 5% significance levels; double asterisks denote 1% or better.) Similarly, some targets required significantly** more questions than others, from a minimum (average) of 3.4 for a target in Youngstown, Ohio, to a maximum of 5.5 for a target in Morgantown.

The length of a given question string, of course, depends on the difficulty the informant had in making a choice. In fact, there is strong evidence that if neither location nor occupation are very useful in a given case, the informants ask many more questions in an attempt to find some basis for making a choice. Specifically, question strings in which questions 3 or 14 were ranked first or second most useful, on average, are about one question shorter than strings where this was not the case. These differences are significant** for each of the six combinations we examined: location was most useful, or it was not; occupation was second most useful, or it was not; etc. In many cases there were also significant differences between informants and/or targets, but this does not affect the conclusions.

When informants had difficulty making a choice, they tended (not surprisingly) to stop when they reached the most useful question. Question strings ending with the most useful question were significantly** shorter, by 1.2 questions, than strings in which the most useful question was asked before the end. A different analysis gives similar conclusions: long strings (those containing six or more questions) end with the most useful question significantly** more often than would be expected by chance. For example, of the 15 strings containing 13 questions, five terminated with the most useful question.

There is evidence that informants become set in their habits of asking questions, even when their own results suggest they should change. (This is an experimental phenomenon which obviously biases our results. Hence the randomizing of the order in which targets were presented.) The mean number of <u>different</u> questions generated by an informant, over all 50 targets, was 19.7, s.d. 8.9, although one informant asked only four different questions, and one asked 40. During interviews several informants were bothered by their inability to think of questions to ask. This led to the development of a basic set of questions which these informants used over the 50 targets. They often felt no reason to

change their set of questions for a different target, as a new set of probabilities for a match was presented each time. (Note that the mean number of different questions generated by all informants per target was 28.3, s.d. 3.1. This much narrower variation per target than per informant suggests that the total amount of information needed by any informant for any target is remarkably uniform.)

Similarly, the correlation between the "sequence" (see above) when a given question was first asked by an informant, and the percentage of time it was asked thereafter, is significantly** negative. In other words, questions asked about the first few targets tend to be used for most targets; questions which were asked for the first time for, say, the thirtieth target are not frequently used after that. This tendency remains even for questions which are perceived as useful the first time they are asked.

As expected, there is a slight, but significant learning effect. Let the number of times a given question is asked by an informant be n, and the number of times it is deemed at all useful be m. The correlation between the fractional time it was useful, m/n, and n is 0.14**. Hence there is a weak tendency to use questions more frequently when they are perceived as helpful.

V. *Sequence of Questions*

The position of a given question in a string depended heavily on the particular question. Figure 5 shows the probabilities that certain questions occur first, second, third or fourth in a string. Location (14) is highly likely to be asked first or second, but much less likely thereafter. Occupation (3) is almost equally likely to be asked first or second or third, but hardly ever after that. Questions such as target hobbies (21) or organizations (26) are almost never asked first, or second, but occur frequently further down the string.

It is straightforward to define the "most likely" question string. Suppose we consider only strings beginning with question 14. This may be useful or non-useful for the informant. In each case he will select a "most likely" second question; whether this is useful or not determines his third question, and so on. Figure 6a shows the sequences, and associated probabilities, for such strings. Sequences beginning with question 3 are almost identical, but with 14 and 3 interchanged.

A clear pattern emerges, with questions on hobbies (21) or organizations (26), etc., being asked when location or occupation are unhelpful. These strings are usually quite short, as indicated both in Figure 6a and in the previous section. Figure 6b shows a similar sequence for strings beginning with question 29 (sex). After finding out the target's age, the informant normally proceeds to occupation and location, before moving into similar sequences to Figure 6a. Finally, sequences beginning with question 1 immediately ask location and occupation, and again continue as in Figure 6a.

There is, of course, a strong causal link between certain questions and those immediately following: "Where does the target travel?" is almost always preceded by "Does the target travel?" and can only be asked if the latter question received an affirmative answer. The causal links can best be presented on a branching diagram (Fig. 7). We define question j to be causally related to question i if

two conditions are met: (a) the probability that i immediately precedes j in a string, given that j was asked, be high; (b) the probability that j not be asked, given i was not asked, be high. The level of causality is then the product of these two probabilities; unity therefore implies certainty in cases (a) and (b). Figure 7 traces each question back down the causal chain to its most likely predecessor (in cases of ties, the lowest numbered question is used for clarity). Thus question 60, if it occurs, always follows question 59, which, if it occurs, follows 35 95% or more of the time, and so on. Note that 36 also, if it occurs, follows 35 95% or more of the time; hence 36 and 59 are almost never asked together. The connections between 2, 29, 30, 3, 14, and 1 are all negligible, but are included for completeness.

Examination of the data shows that the weakness of some of the causal links is caused by informants shuffling the orders of some of their questions. This may be due partially to informants' attempts at breaking the monotonous routine of the experiment. Some felt that part of the "game" was to skip needless questions ("why ask if a target has children when I can just ask the ages of children and get more information?"). Some informants even intentionally varied the order of their questions in order to avoid "getting in a rut." To allow for this, the requirement that i immediately precede j was relaxed; instead, i need merely precede j, at some position in a string. Fifty-one questions were found to be 100% related to 1, 2, 3, o4, or (in one case) 30, and usually to several.[3] Sixteen more were related to 3 or 14 at the 90% level or above. Of course, this reflects the strong probability of 3 and 14 being asked near the beginning of each string.

If we remove any reference to the order of the questions, and concentrate instead on the "packaging" by informants of entire question strings, some systematic patterns emerge. For example, the 2,500 (50 x 50) question strings were treated as lists of 99 integers. The jth integer in a string was one or zero, depending on whether question j was asked in that string or not. Factoring these strings produced the questions which tend to occur together. Nineteen factors were found; several of these had only one question with a high (varimax) loading on that factor. Define a group to be those questions with a factor loading of 0.2 or more on one factor (the "typical" loading is 0.02). The groups found were: 21, 30, 39, 41, 42, 53, 54 (children); 2, 21, 29, 30 (socioeconomic status); 37, 38 (travel); 35, 36 (family); 6, 63 (spouse); 23, 24, 49 (schooling); 8, 22, 40 (more socioeconomic questions); 49, 59 (social life); 7, 17 (previous location, occupation); 21 , 26 (spare time); 56, 64, 11, 52 (precise details); 1, 4 (more precise details); 10, 74 (more precise details).

It is interesting to note that 3 and 14 do not occur in these groups, although 14 has a high loading on factor 2 but with the opposite sign (i.e., when 14 is asked, group 2 tends not to be, and vice versa). A similar factoring but on strings with -1 (question asked but not useful), 0 (question not asked), and +1 (question asked and useful) yielded very similar groups.

Finally, we chose to examine the pattern of co-occurrence of questions by a slightly modified form of cluster analysis, operating on groups of questions. Initially each question was allocated a unique group. The data to be clustered were the probability that a question in group j would co-occur with one from group i, given that at least one from group i was asked in a string.[4] Groups i and j were merged if the relevant probability was above a cutoff level (usually 100%,

but reduced by 5% as many times as necessary to ensure further merging). After each set of merging, the probabilities were re-computed; the process halts when all questions are in one large group. Unfortunately, one large group, containing 3, 14, 2, 29, etc. was created early in the procedure; because at least one of these questions was always asked in a string, this group always co-occurs with other questions, so that this group then swallowed up all the other questions.

To modify this, it was decided not to allow merging with any group whose (weighted average) probability of co-occurrence with any other group exceeded an arbitrary figure of 50%. This yielded the nine question groups shown in Table 5. Apart from the rare questions in groups 3 to 5, the clustering seems to have yielded meaningful groups of questions; it is particularly enlightening that virtually all questions, other than basic socioeconomic ones, are together in one group.

The strong patterning of questions which we have seen throughout this section confirms the strategy we adopted in RSW of supplying location, occupation, and sex of target to our informants. We also found in RSW that the race and/or ethnicity of targets was unimportant to informants, and this result is confirmed here. Race was only asked 1% of the time in the present experiment, and ethnicity was asked only ½% of the time. Based on data from this informant-defined experiment, future RSW-like experiments should include the age, organizational memberships, hobbies, and marital status (including children) of targets.

VI. *Accounting for Questions*

The previous sections described the questions asked by informants. We now seek to explain the variation in questions by referring to differences between both informants and targets. At the simplest level, there are 2,500 question strings, each asked by an informant (with known background data) about a target (also with background data). It is logical, therefore, to attempt to account both for the number of questions asked, and for whether a given question was asked, on the basis of individual strings.

Multiple regression of the number of questions in a string with both informant and target data accounted for only 16% of the variance. Although this amount is significant**, many correlations later in the paper account for far larger variances. To save space, we henceforth choose to ignore any regression accounting for less than 40% of the variance.[5] Similarly, the number of different questions per informant or per target was not well predicted by personal characteristics (although target with children, not surprisingly, had significantly* more different questions asked about them). Discriminant analyses were conducted on frequently asked questions, in an attempt to predict for which informants and targets any given question would be asked. Target data are of little use in this matter: only the background of the informants have much bearing on whether they ask a question. For example, question 16 (where's the target's location near) is asked more* by informants who have lived in places other than Morgantown than by those who have always lived in Morgantown (7.3 to 2.8 respectively). Additionally, most questions are hardly ever asked, and even after discriminant analysis, the optimal prediction remains that these questions are never asked.

However, for three questions it was possible to improve the quality of this simple prediction. For questions 2 (age), 14 (location), and 29 (sex), the discriminant function correctly predicts whether the question was asked 67%, 93% and 73% of the time respectively. These are to be compared with the (null hypothesis) prediction of 58% not asked, 90% asked, and 64% not asked. In all three cases the structure of the discriminant function is similar: the higher the informant's occupation and education levels, the more organizations they belong to and the more hobbies they have, then the more likely age and sex are to be asked, and the less likely is location to be asked.

Restricting attention to questions which were designated "most useful" destroys all predictive capability except for question 14. Since the "most useful" designation depends on the type of target, both informant and target date affect whether 14 is most useful or not. This can now be predicted 72% of the time. The higher the informants' age, the lower their education level and the number of organizations they belong to; the higher the target's occupation level and education, and the larger the target's town, the more likely is question 14 to be most useful. (This is one of the few inconsistencies with RSW; a higher target occupation level should be paired with a higher probability of occupation being the most useful question. However, the prediction from RSW is confirmed if we examine the discriminant function for question 3, where the higher the target's occupation level, the more likely is 3 to be the most useful question.) Similar results are found for questions which were designated to be at all useful, save only that target's occupation now has no significant effect upon whether location is useful.

Having found that little variation in questions could be accounted for on a string-by-string basis, we analyzed the question strings first averaged over all targets (i.e., retaining only informant data with which to explain the variation) and then averaged over all informants (retaining target data).

There are a great many questions which could be asked of either of these data sets. In searching for signals in the data we chose to examine whether differences in dichotomous variables produced significant differences in question usage. For example, do male informants ask a specific question more than females? Then we attempted to account for differences in question usage by regressing question usage against characteristics of informants or targets.

Table 6 presents all 13 examples of significant** differences in question usage between informants, split into two subgroups by various criteria. There were 13 additional significant* findings which are not reproduced. The reason for their suppression is that, of the 700 tests we carried out, we would expect 35 to be significant by chance alone at the 5% level. However, only 7 significant comparisons at the 1% level would be expected by chance; the 13 findings in Table 6 should not, therefore, be a chance occurrence. (In fact, the total of 13 is itself significant at the 5% level.)

The most interesting findings are the persistent use of question 38 ("where does the target travel?") by informants who report that they do not travel. Presumably those who do travel are more likely to have connections with the target's location, and therefore have no need to enquire further. The other finding, not in Table 6, is that males ask location less* than females and find location to be most useful, when asked first in a string of questions,

more* than females. Combined with the fact that females find location to be most useful when asked last more** than males, there is a clear difference in usage of location between the sexes. We have no idea why this consistent difference emerges from the data.

Multiple correlation of question usage by informant characteristics yielded very little additional information. No informant characteristics accounted for: the mean number of questions asked; the number of different questions asked; the probability of a given question being asked;[6] the probability of a given question being most useful; or the probability of a given question being not useful. Three multiple correlations did produce acceptable results (i.e., more than 40% of the variance accounted for, over 50 cases; nearly all the 125 multiple correlations were statistically significant). For example, the probability that marital status was first or second most useful increases as occupation level, number of hobbies and organizations, education, and degree of activity in religion of the informant increases. This makes intuitive sense: the more connections an informant has with the rest of the world (as indicated by organizational affiliations, for example) the easier it is to "connect" with a target; but if this fails, then it may still be possible to "connect" with the target's spouse, especially if the target is a housewife.

Next, the probability of question 14 being most useful when asked first in a string is higher for male informants and for those who are active in religion, and those who have children; it also increases with age, income, and decreases with number of organizations and hobbies, education and occupation level of the informant. It would seem that informants with few links to the world may ask other questions after location, but there is less chance that these questions will be useful. Finally, the probability that question 7 (religion) be most useful when asked last decreased with informants' occupation level, income, education, and number of organizations; it increases with age; and it is lower when the informant is active in religion or has children.

Target characteristics apparently play a much larger role in accounting for question usage, although there are again only 13 significant** differences in question usage for different targets, as split into two subgroups by various criteria. Table 7 shows several interesting features. The sex of the target is asked more often** for female targets (do informants somehow get clues to a target's sex from other questions and then seek to confirm their feeling?) Questions 6 and 63 (relating to spouse's occupation) are asked more often** about female targets, and conversely question 3 (actual target's occupation) is asked more often** about males. The split of targets into those in urban and rural areas confirms the findings in RSW, namely the strong tendency to use occupation as a reason for rural targets and location for urban targets.

Table 8 demonstrates that target characteristics account for most of the frequently asked questions. (In fact, multiple regressions of any question usage or question-related topic on target characteristics almost invariably yield significant* amounts of variance accounted for.) Several clear signals, again, confirming the findings of RSW, occur in these fits. For example, the probability that the frequent questions (other than 3 and 14) are asked, increases with the distance of the target from Morgantown. Occupation tends to be most useful for targets far from Morgantown, in a rural location, with high occupation

level. Conversely, location tends to be the most useful reason when targets live in urban locations; however, high occupation level and distance also tend to increase the probability that location be most useful. Note also that question 1 (name) is most useful, given that it was asked, only for targets near Morgantown, as might be expected. For such targets, the likelihood of the target's name being most useful increases as the target's socioeconomic status decreases. (Name was used in one of three ways: primarily as an identifier for the next person in the [nonexistent] chain; secondly, if a choice had the same name as the target; thirdly. if the target's name implied ethnicity which could be used as a criterion for making a choice.) Questions relating to hobbies (21) and organizations (26) also yield plausible results: the more hobbies or organizations a target has, the more likely is the relevant question to be useful; the likelihood is also increased for targets living further from Morgantown. However, hobbies are more frequently useful for male targets, and organizations for female targets, although the latter tendency is very weak.

In summary, characteristics of targets control most of the question which informants ask; and characteristics of informants do not appear to have much influence on which questions are asked. Of course, on a one informant-one target basis, this is untrue (witness the discriminant functions). The signals only emerge upon averaging over all informants or targets.

VII. *Choices*

On average, informants used 40.7 different choices for the 50 targets (s.d. 4.9). This number is significantly** higher than the 34.7 different choices for the first 50 targets in RSW. The difference is of course due to the inclusion of ten very local targets in the current experiment. In fact, on average, 9.2 different choices (s.d. 0.9) were used for the ten local targets, suggesting that each informant has a large number of choices for local targets, as expected from intuition. Only two of the remaining 40 non-local targets, on average, had one of the "local" choices used for them. If locality did not matter to informants, this low number would occur by chance less than one in 10^{20} times. Hence "local" choices are only used for local targets.

Informants made male choices 67% of the time (which is significantly** higher than the 60% found in RSW, but reflects the same tendency to choose males). Informants made choices who "knew a lot of people" only 7% of the time (s.d. 15%). The distribution of these choices across targets is significantly** less scattered, suggesting that the decision to use someone who "knows many people" is a function solely of informants, and not of targets. Similarly, the number of intervening choices used per target varies significantly less** than per informant, so that the decision to use an intervening choice depends only on the informant.

Friends and acquaintances account for 80% (s.d. 11%) of all choices made, with family members accounting, of course, for the other 20%. These figures are almost identical to the 82% (s.d. 10%) found in RSW.

Again, qualities of informants and targets account for some of the variation in the above values. As in RSW, sex is the most important factor accounted for. Male informants make significantly** more male choices (37.9) than female informants (31.1); male choices are made significantly** more

often for male targets (38.6) than for female targets (27.9); female informants used relatives as their choice significantly** more often than did males (11.6 to 7 respectively). In RSW we noted that females used family more than males did, but we had no way to know whether this would still hold when the target population included 20% local targets (10 out of 50).

The number of different choices made can be fitted well (43%) by a linear combination of informant characteristics. The number increases with informants' occupation and education levels, number of organizations and hobbies, income; it decreases with age, and if the informant is active in religion and/or has children. The number of male choices made decreases with the informant's age, occupation and education levels, and number of organizations; it is higher if the informant is male or has children (46% of variance accounted for).

The number of male choices per target may also be accounted for by target characteristics (53% of variance). The number increases with target's occupation and education levels, number of hobbies and distance from Morgantown; it also increases for male targets with children; and it decreases with age of target. Finally, the number of family members used per target increases with target's distance from Morgantown and age; decreases with occupation and education levels and number of target's organizations; and increases if the target lives in a rural area or is female (44% of variance).

Given this brief discussion of choices, we turn now to the reasons they were chosen.

VIII. *Reasons for Choices*

It is difficult to separate reasons totally from questions or choices, so that the degree of usefulness of a question, for example, has frequently been a feature of the previous sections. However, we can now extend the concept to include features of the choices discussed in the coding section: namely the direct hits, associated hits, vias, and intervening choices.

We found in RSW that, for any target, the most popular reason for choice was always location and occupation (but only "ethnicity" and "other" were the other possible reasons). The current data permit testing of this finding. Over the 50 targets, location was the most popular reason for choice 23 times, and occupation 25 times. Only twice were there any other most popular reasons: once question 6, once question 4.

The finding is repeated if we consider the most popular reason for choice per informant. Twenty-one informants used occupation most often, twenty-six used location most often, and three informants used one each of age, hobbies and organization most often. Hence the dominant role of location and occupation as overriding reasons for choice is confirmed.

The data also permitted an independent test of the discriminant function computed in RSW to predict the most likely reason for choice. However, its success rate on the current data was a mere 50%. We tried removing all local targets (since RSW contained none) but the predictions were not improved.

The mean number of direct hits (per informant-target combination) was 0.9, s.d. 1.1. Certain questions are the most likely to be those which are direct hits. These are the basic collection which recur throughout this paper: 3,14 (0.9 times per target over the 50 informants); 2 (0.46); 21 (0.34); 26,29 (0.26); and 30 (0.14). Virtually nothing in the informant or target data accounts for any variation in quantities connected with direct hits.

There was a similar number of associated hits (mean 0.95, s.d. 0.82). Again, occupation and location dominate: 3 was an associated hit 18.1 times and a via 13.9 times per target; 14 was an associated hit 21.3 times, a via 13.2 times. All other questions occurred only about once per target at most. Although many of these can be accounted for by regression with target or informant characteristics, there is little point in explaining very infrequently occurring phenomena. However, question 3 occurs as an associated hit significantly** more for male targets (21) than for female targets (15). Question 14 occurs as an associated hit significantly* more for targets with children (22) than for childless targets (15); and as a via significantly less* for targets who live in an urban area (11) than for those in a rural area (17). These three findings tend to agree with both with other findings in this paper and those in RSW.

Finally, there were only 0.1 (s.d. 0.4) intervening reasons, on average; again questions 3 and 14 dominate the usage in this category (10.3 intervening hits per target for question 3, 19.9 for 14, 4 for 16 and under 1 for other questions; and only 3,14 ever occur more than once per target as an intervening via).

IX. *The "Tag" Concept and Its Use in Predicting Choices*

If we are to understand how an individual makes a choice when presented with limited information about a target, we need to model the decision process in some way that allows testing. The model we present here is very simple, and surprisingly successful in predicting the choices made by each informant.

We shall assume that an informant selects a choice for a given target because, in some sense, the informant perceives the choice to be "similar" to the target. Furthermore, we assume that if there are several choices which are similar to a target, the informant chooses the most similar such choice (however this may be evaluated by the informant). In other words, the dissonance between the choice and what is known about the target is minimized.

How is similarity between choice and target to be measured? Clearly, the actual decisions involve highly complex cognitive processes about which we understand little. As a simplification, therefore, we assumed that a choice and a target are perceived as similar if and when some facet of the choice (e.g., where the choice went to school) and some facet of the target (e.g., where on of the target's children attends school) are either connected or, at best, identical. In some cases, of course, we had to suggest this concept to informants; this inevitably must weaken the following case slightly.

We shall term each facet of a target's personal history a "tag." Although targets began the experiment with very few tags (see Section II), as

the experiment progressed and more data were invented for each target, the number of tags grew. Of course, the nature of the tags differed widely, as did their coding within the data. In order to count and catalogue the target tags, we again simplified the problem. We treated all tags coded in location style as location tags (i.e., target's location, previous location, family location, where the target travels, etc. are all location tags). Overlaps were removed (so that if a target still lives where he was born, only one tag is created). Similar tags were counted for occupations, hobbies, organizations, age, sex, and religion (the latter three, for targets, consisted of one tag apiece, of course).

Targets develop many tags. Figure 8 shows the probability of a target possessing 11, 12, ..., or 23 tags (no target possessed a number of tags outside this range). The mean number of tags was 15.7, s.d. 2.7. Splitting into the various types of tags yields Figure 9, showing that occupation, hobby and organization tags are all similarly distributed, with means of 2.5 to 3, s.d. about 1, while the mean number of location tags was 4.7, s.d. 2.2.

We did not possess directly comparable data about choices; collection of such data would have presented enormous complexity and was not attempted.[7] Instead, we chose to deduce choice data by using the reasons informants gave for each choice. Whenever a choice achieved a direct, associated, or intervening match with the target it was chosen for (possibly several targets), that piece of target data was added to the list of tags for that choice.

The number of choice tags (again, only distinct ones are counted) is much fewer than for targets. Figure 10 shows the probability of any choice having 0, 1,...,12 tags. The mean number is 1.8, s.d. 1.2. Split into categories (Figure 11), the dominance of location and occupation tags is clear (mean numbers of 0.75, 0.62, compared with at most 0.14 for all other tag types).

We can now test the simple hypothesis that, for any given target, an informant will choose the choice that has the largest number of matching tags with that target. (This procedure is of course biased by the way we obtained the choice tags: the correct choice, for a given target, almost invariably possesses some tags in common with that target. However, we will allow for this statistically below.) We define two non-location tags to match only if they agree completely; in other words, an occupation tag of "symphony orchestra conductor" does not match with "symphony orchestra player." Location tags match if the choice and target location tags correspond to positions in the U.S. less than some cutoff distance apart. These cutoffs were taken to be 444 km, 222 km, 111 km, etc., down to 7 km, and a final cutoff of 0 km (corresponding to a location match only if the two locations are identical).

Thus, for each distance cutoff, and for each informant-target combination, we can nominate the 'optimal' choice(s) as being the choice(s) which has (have) the most tags in common with the target, and then compare the optimal with the actual choice. We have defined two ways to measure the accuracy of this procedure, which we term the "easy and difficult scores." The easy score is defined as unity whenever the actual choice is among the optimal choices, and zero otherwise; the difficult score as 1/(no. of optimal choices) if the actual choice is among the optimal choices, and zero otherwise. In other words, the easy score counts how often the actual choice was correctly (but not necessarily

uniquely) predicted; the difficult score counts how often we would be correct if we chose at random from among the optimal choices.

The results of this are shown in Table 9, averaged over all informant-target combinations. Two things are immediately obvious. The first is that maximal accuracy is obtained when the location matches exactly, although there is little degradation if two location tags are up to 28 km apart. Henceforth, we shall treat location tags like all others, and require an exact match. The second is the high rate of accuracy: the actual choice is among the optimal choices 89% of the time, and predicted 60% of the time if the choice is selected at random from among the optimal choices.

The high success rate, combined with the simplistic approach, suggest that we might improve the accuracy still further if we weighted the tags in some fashion before counting the matches. We examined eight different weighting combinations, shown in Table 10. Note that no weighting achieves the accuracy of the simplest counting scheme; also, the overall importance of location and occupation is confirmed by scheme 3's scores of 0.52, 0.82. We are forced to conclude that if the model has relevance to the decision process, then all tags should be counted, independent of their directness or lack thereof. In linear programming terms, then, all tags have an equal utility.

Informants occasionally asked about schooling. In our coding, schools were not allocated a precise location (i.e., they were not given Cartesian coordinates) but were recorded by state and an identifying arbitrary code. Including schools as a separate tag might improve accuracy. However, if the school's state is used as a tag (so that all schools in the same state are identical for predictive purposes) this weakens the scores to 0.59, 0.87. Using the school's unique code as a tag improves the accuracy, but only by 1%, to 0.60, 0.90 for difficult and easy scores respectively. Hence inclusion of schools is of no real help in predicting choice.

The difficulty with interpreting these results stems almost entirely from the biased way the choice tags were obtained. It seems intuitively obvious that if one obtains some choice tags from the target for which that choice was made, then that choice is likely to be the one with the largest number of tags matching with that target. Clearly, we need to estimate how likely it is that we achieve the levels of accuracy observed in our data.

To calculate this we need three sets of probabilities. The first, α_r, $r = 0, 1, 2, \ldots, 5$, are the probabilities that the actual choice has r tags matching the target. These probabilities, from the data, are shown in Figure 12. The mean number of matchings tags is 1.56, s.d. 0.84. The other sets are β_n, $n = 11, 12, \ldots, 23$, the probabilities that a target has n tags (previously in Figure 8) and γ_m, $m = 0, 1, \ldots, 12$, the probabilities that a choice has m tags (given previously in Figure 9).

We assume all tags are of a similar type (retaining the different categories would involve awkward partitions of integers, without adding significantly to the results). Let there be N tags in total (there are 451 different target tags in all: 126 location, 84 occupation, 80 hobbies, 109 organizations, 37 ages, 2 sexes, and 13 religion tags). We might take N as 451, or perhaps only (126 + 84), depending on our interpretation of the number of tags.

Now the probability that a random choice tag matches a random target tag is clearly 1/N. If the target has n different tags, and the choice hs m different tags, then the probability of exactly t matching tags in

$$q_t = {}_m C_t \left(\frac{n}{N}\right)^t \left(1 - \frac{n}{N}\right)^{m-t}$$

by the binomial theorem (the n/N factor derives from n chances of 1/N, of course). Hence the probability of less than r matches, P_r, is given by

$$P_r = \sum_{s=0}^{r-1} q_s.$$

Thus, if the correct choice has r matches, the probability that another random choice achieves less than r matches, and is not optimal, is P_r, and the probability that a random choice achieves the same number of matches is q_r. The probability that another choice achieves more tag matches than the correct choice is $(1-P_{r+1})$.

These probabilities, so far,are conditional upon the values of r, m, and n. Summing over r, m, and n, and multiplying by $\alpha_r \beta_r \gamma_r$, yields the probability P that a choice has fewer tag matches than the correct choice, and Q that a choice has the same number of tag matches (and, therefore, $R \equiv 1-(P+Q)$ that a choice has more tag matches).

The expected value for the easy score E is then given, since there are on average 41 different choices, by

$$\mathcal{E}(E) = (P + Q)^{40} = \mathcal{E}(E^2)$$

Which is simply the probability that all the other choices score less than the correct choice. The variance is then given by

$$\sigma_E^2 = (P + Q)^{40} - [(P + Q)^{40}]^2.$$

The difficult score is slightly more awkward to evaluate numerically. The probability that any other choice achieves the same score as the correct choice is

$$\mu_a \equiv {}_{40} C_a \, Q^a P^{40-a},$$

giving expectancies for the difficult score D as

$$\mathcal{E}(D) = \sum_{a=1}^{40} \mu_a/a, \mathcal{E}(D^2) = \sum_{a=1}^{40} \mu_a/a^2$$

and variance

$$\sigma_D^2 = \mathcal{E}(D^2) - [\mathcal{E}(D)]^2.$$

We can now compare the observed scores with those expected. With N = 451 different tags,

$$P = 0.922, \quad Q = .078$$

$$\mathcal{E}(E) = 0.859, \ \sigma_E = 0.348$$

$$\mathcal{E}(D) = 0.271, \ \sigma_D = 0.204$$

Over 2,500 cases, the observed mean easy score was 0.89. Dividing σ_E by $(2,500)^{\frac{1}{2}} = 50$, we find that 0.89 is some 4 standard deviations above expected. Similarly, the observed difficult score of 0.60 is 80 standard deviations above expected. Thus both observed scores are significantly** higher than expected by chance, although it should be borne in mind that 0.86 is the expected easy score (i.e., the system is biased toward a high score).

Adjusting the effective number of tags only makes this conclusion firmer. Restricting attention to location and occupation, which, from Table 10, achieves difficult and easy scores of 0.52, 0.82 respectively, yields expected scores of 0.15 (s.d. 0.15), 0.65 (s.d. 0.48) respectively. Again, the observed scores are significantly** high.

It should be noted that we deliberately did <u>not</u> restrict the target's tags to what each informant knew about the target (i.e., we compared choice location tags with a target's places of travel, whether or not the informant had asked about the target's travel). This was to permit the other choices more chance of matching tags with the target. If we <u>do</u> restrict the target's tags to what each informant asked about, the mean difficult and easy scores rise still further to 0.78, 0.94 respectively. It is clear that this is too biased to be regarded as a fair test of the tag concept.

X. *Tags: A More Detailed Examination*

The tag concept discussed in the previous section is a simplistic one. A detailed ethnographic study ofhow informants make a selection between choices of apparently equal utility would be very valuable. However, the 89% success rate of the straightforward counting procedure, although biased, obviously accounts for a large amount of the decision process.

The number of tags differs strongly between targets; and the total number of choice tags differs between informants. We examined whether characteristics of either targets or informants could account for this variation. Multiple regression of the number of occupation tags for a given target showed that 45% of the variance could be accounted for by a linear combination of socioeconomic variables (as usual, we suppress all cases where less than 40% of the variance is explained). The largest contributors are targets' age and occupation level: the higher the target's age, and the lower his occupation level, the more occupation tags that target possesses. This is plausible: too low an occupation level forces informants to search for other occupations related to that target. The only other target characteristic which accounted for any variance in the data (apart from trivial connections like number of hobby tags with number of hobbies) was that targets who travel have significantly* more hobby tags than those who do not travel. We cannot account for this; it may simply reflect a bias in the construction of the target data.

Similarly, informant data accounts for little variation in the total number of tags (not necessarily distinct) which their choices possessed. Each informant had a total of 75 tags on average, of which 29 were location and 25 occupation. The only significant result is that female informants have more* occupation tags than male informants, by 28 to 21. Thus, little about targets or informants accounts for variation in tag density.

Another, more subtle, bias in our use of tags is the degree of utility of each tag. Because of the manner in which choice tags were deduced, the majority of them have, in some undefined sense, a high degree of utility for that informant. Thus, with hindsight, equal tag weighting is likely to yield the most accurate results.

Clearly, not all tags are really of equal utility: it seems plausible that a choice currently living in Chicago is more likely to be chosen for a Chicago target than, say, a choice whose father travels to Chicago. Given the limitations of our experiment, however, we could not test for this.

To extend the investigation, we conducted a follow-up interview with informant 15. He provided two new sets of data. The first was a count of the number of connections or tags between each of his 33 choices and each of the 50 targets, now given all target information, rather than the limited information he requested during the original experiment. Then, armed with all the information, he told us which choice he would now make for each target.

This mini-experiment was slightly flawed for two reasons. First, in order to reduce the many hours of the follow-up, we had collated all locations and occupations relevant to each target. As a result, if the target lived in a small town but was attending college in a neighboring big city, both the big city ("the town is near X") and the college ("attends X college") were available as location tags. This doubling-up of essentially the same information made interpretation of the data somewhat difficult. Second, the informant ignored the myriad of possible intervening choices, as we had requested. This automatically removed such reasons as "I choose Y, because she knows someone at Z oil company," which had been used in the original experiment. A more precise, informant-defined interview would be very valuable.

Informant 15 generated many tags between choices and targets. On average, between any choice and any target, there were 0.27 location matches, 0.05 occupation matches, 0.21 hobby matches, and 0.06 organization matches, or 0.58 matches in all. Corresponding s.d.s are 0.3, 0.06, 0.21, 0.06, and 0.41, respectively.

In the original experiment, the difficult and easy scores for the tag concept for informant 15 had been 0.50, 0.86 respectively. Repeating the calculation based on the more complete tag information reduces the accuracy noticeably to 0.44, 0.25 (although this is now unbiased, of course, so interpretation of the scores is somewhat altered). His final choices contained 17 alterations from the original set of choices; on three occasions he would now prefer to make a choice outside his initial 33. Rather surprisingly, the tag scores based on his final choices decreased slightly to 0.24, 0.42.

We examined the cases where simple tag-counting yielded the wrong choice. Almost invariably it was a matter of "how strong" a tag was: a choice living in the target's location (1 tag) being preferred, for example, to a choice who was born in that location and whose family still lives there (2 tags). Thus the original tag experiment had, as we suspected, automatically scaled the utility of most tags, and chosen the most useful ones. The follow-up interview, however, did not contain any utilities, and this probably accounts for the reduction in accuracy.

This suggests a variety of informant-defined experiments both to find out what weighting of tags and tag types is necessary to yield accurate prediction of choices, and to find what other qualities of informants and/or targets are important in determining why some targets have stronger ties with some informants than with others.

FOOTNOTES

1. In RSW, 25% of the targets were housewives; the uniformity of response by informants to these targets led up to reduce the number of housewives for the present experiment. With hindsight, three housewives out of 50 targets are too few to produce reliable statistics.

2. "Cities" are defined (except for Morgantown) as places we felt informants anywhere in the U.S. would recognize.

3. The same conclusion holds if attention is restricted to those portions of a string up to when the most useful question was asked.

4. The conditional probability enables rarely-asked questions to be added to a group.

5. This is rather unusual for the social sciences, perhaps. However, a scatter diagram of a regression accounting for 1½% of the variance shows very little in the way of a signal; hence our suppresion of low variance.

6. We examined only the 25 most frequently asked questions, which account for 95% of all questions ever asked.

7. Of course, we had a great deal of data about the choices in the one or two sentence explanations given by informants. However, these data were not collected with the idea of systematic comparison, and we simply did not know how to code the data for such comparison. It is obvious how to collect comparable data about choices; but this would have increased the time required for interviews by so much that we were forced to abandon this part of our original design.

References Cited

Bernard, H. Russell and P. D. Killworth
1979 "A review of the small-world literature." Sociological Symposium (in press).

Duncan, O. D.
1961 "Socioeconomic index scores." In A. J. Reiss, Jr. (ed.), Occupations and Social Status. New York: Free Press.

Killworth, P. D. and H. R. Bernard
1978 "The reverse small-world experiment." Social Networks, 1:159-192.

1979 "A pseudomodel of the small-world problem." Social Forces (in press).

Lin, N., P. Dayton, and P. Greenwald
1978 "Analyzing the instrumental use of relations in the context of social structure." Sociological Methods and Research (in press).

Milgram, S.
1967 "The small-world problem." Psychology Today 1: 61-67.

de Sola Pool, I., and M. Kochen
1978 "Contacts and influence." Social Networks 1:1-48.

	Mean	Standard Deviation	Median	Mode	Range
Age (years)	44.8	14.4	45.2	45	50
Occupation level (Duncan scale)	45.5	28.8	44.5	19	93
Income ($1000/year)	19.9	11.8	16.1	14	65
Distance from Morgantown (km)	575	668	231	0	2470
Number of hobbies	2.6	0.8	2.5	2	3
Number of organizations	2.5	0.8	2.2	3	4
Number of children	3.2	1.4	3.2	4	6

TABLE 1. Some socioeconomic data about the 50 targets; income was not provided a priori.

	Mean	Standard Deviation	Median	Mode	Range
Age (years)	41.9	15.1	42.5	30	59
Occupation level (Duncan scale)	42.9	24.5	43.5	19	75
Other occupation level	34.0	24.6	19.2	19	61
Spouse's occupation level	49.8	26.4	51.0	19	94
Previous occupation level (first of several)	43.9	23.0	44.3	39	82
Income ($1000/year)	17.4	8.6	16.2	13	25
Number of hobbies	2.8	1.0	2.8	3	4
Number of organizations	2.4	1.3	2.4	3	4
Number of children	2.6	1.6	2.2	2	7

TABLE 2. Some socioeconomic data about the 50 informants.

*	NAME:	Benjamin S. Clay
*	LOCATION:	Charleston, West Virginia (South Charleston)
	BIRTHPLACE:	Columbus, Ohio (July 12)
*	OCCUPATION:	Papermill laborer
*	PLACE OF EMPLOYMENT:	Wilhelm Paper Co. (worked there four years)
	INCOME:	$15,000
*	AGE:	32
*	RACE:	White
*	RELIGION	Catholic (not an active member)
*	EDUCATION	Graduated from Richwood High School
*	HOBBIES:	Pistol-shooting (competition), hunting
*	ORGANIZATIONS:	National Rifle Association, Committee for Handgun Control, United Paperworkers International Union
	SERVICE RECORD:	Served in the Navy for four years; stationed in San Diego, Calif.
*	MARITAL STATUS:	Divorced
*	SPOUSE'S OCCUPATION:	Was a housewife
*	CHILDREN:	Two sons, ages 10 and 12 (both live with spouse in Frostburg, Maryland)
	FAMILY:	Came from small family
	TRAVEL:	Kentucky, North Carolina (Smokey Mountains)
	PREVIOUS OCCUPATION:	Gas station attendant in Columbus, Ohio.

Table 3. A typical target dossier, after all 50 informants have asked questions. Asterisks denote information allocated at the beginning of the experiment.

TABLE 4

(See legend at end of table.)

1) What is the target's name?

2) What is the target's age?

3) What is the target's occupation?

4) Where does the target work (i.e., name of company)?

5) Does the target have any other occupation?

6) What is the occupation of the target's spouse?

7) Does the target have any previous occupation?

8) What is the target's income?

9) What is the target's religion?

10) Is the target active in religion?

11) Where does the target attend church?

12) What is the race of the target?

13) What is the target's ethnic background)

14) Where does the target live?

* 15) Is the target a resident or a commuter?

16) Where is the target's location near? That is, what is the closest well-known urban area?

17) Has the target ever liced anywhere else?

18) How long ago did the target move?

19) How long has the target lived in the present location?

20) Where was the target born and raised?

21) What are the target's hobbies?

22) How far has the target gone in school?

23) Where did/does the target go to school?

24) What is/was the target's field of study in school?

25) Where did/does the target's spouse go to school?

26) What organizations does the target belong to?

* 27) Is the target active in his/her organizations?

* 28) What organizations does the target's spouse belong to?

29) Is the target male or female?

30) What is the target's marital status?

31) How long has the target been married?

* 32) Is the target's spouse alive?

33) How old is the target's spouse?

34) Has the target been divorced? [=30]

35) Does the target have a family (mother, father, etc.)?

36) Where does the target's family live?

37) Does the target travel?

38) Where does the target travel?

39) Does the target have children?

40) How many children does the target have?

41) What are the ages of the target's children?

42) Where do/did the target's children go to school?

43) Has the target served in the armed services?

44) What branch of the armed forces did the target serve in?

* 45) What is the name of the target's spouse?

46) Does the target play an instrument? [=21]

47) Does the target's spouse go to school? [=6]

48) How long has the target worked as his/her present occupation?

49) What is the rank of the target in college?

50) Does the target have a boy/girlfriend?

51) Does the target go to school?

52) Exactly where in the city does the target live (i.e., neighborhood)?

53) Where do the target's children live?

54) What are the occupations of the target's children?

55) What is the place of employment of the target's children?

56) How many boys and girls does the target have?

57) Is the target retired?

58) Has the target had any previous occupation? [=7]

59) Is the target's spouse a native of the state where they live?

60) Is the target's business large or small?

61) What is the age of the target's parents?

62) Has the target's spouse lived anywhere else? [=68]

63) Where does the target's spouse work?

64) What is the divorced spouse's location?

65) What is the medical specialty of the target?

66) Does the target have any grandchildren?

67) Has the target lived anywhere else? [=17]

68) Where was the target's spouse born and raised?

69) What is the exact birthdate of the target?

70) What is the target's spouse's age? [=33]

71) How long has the target been married? [=31]

72) How far has the target's spouse gone in school?

73) What is the income of the target? [=8]

74) Does the target live in a house, apartment, or condominium?

75) What is the target's physical condition?

76) What is/are the occupations of the target's parents?

77) What is the SES of the target's parents?

78) Does the target have a boy/girl friend? [=50]

79) What is the occupation of the target's boy/girl friend?

80) Where does the target's boy/girl friend work?

81) What is the previous location, if any, of the target's spouse? [=68]

82) How many years has the target's spouse worked at the present occupation?

83) What is/was the target's children's field of study?

84) Is the target's spouse active in religion?

85) When did the target graduate or stop going to school?

86) Where do/did the target's parents work?

87) Is the target's family large or small (i.e., mother, father, brothers, etc.)?

88) What is the target's SES?

89) How many people work for the target?

90) What are the occupations of the members of the target's family, excluding the target's mother and father?

91) Where was the target stationed in the armed service?

92) How many years was the target in the service?

93) How long ago was the target divorced?

94) Has the target ever published anything?

95) Who supports the target's child/children?

96) Is the target's child/children enrolled in a day care center?

97) Does the target live alone?

98) Is the target paying alimony?

99) Other questions are lumped in this category (six one-offs).

Table 4. All questions ever asked by all informants. There is no connotation as to order. Asterisks imply questions asked in pretesting only. Brackets imply question equivalent to an earlier one.

Group	Contents	Comments
1	3	occupation
2	14	location
3	18	how long ago did target move? (rarely asked)
4	77	what is the socioeconomic status of the target's parents? (very rarely asked)
5	82	how many years has the target's spouse worked at the present occupation? (very rarely asked
6	2,21,26,90,97	age, hobbies, organizations and rare questio
7	29,30,39,48,98	sex, marital status, children, and rare questions
8	all other questions	

TABLE 5. Groups of questions found by clustering.

Topic	Splitting of Informants
No. of times question 1 asked	have children (11) vs. no children (2)
No. of times question 38 asked	travel (1.9) vs. non-travel (9)
No. of times question 2 most useful	males (0.05) vs. females (0.8)
No. of times question 38 most useful	travel (0.4) vs. non-travel (2.3)
No. of times question 6 second most useful	urban background (.06) vs. rural backg. (0.06)
No. of times question 16 second most useful	urban background (1) vs. rural backg. (0.02)
No. of times question 3 at all useful	males (22) vs. females (32)
No. of times question 38 at all useful	travel (0.9) vs. non-travel (4)
No. of times question 1 was not useful	have children (9.2) vs. no children (1.5)
No. of times question 4 was not useful	have children (4.8) vs. no children (2.5)
No. of times question 38 was not useful	travel (1) vs. non-travel (5)
No. of times question 14 was most useful and asked last in a string	males (0.6) vs. females (3.3)
No. of times question 38 was most useful and asked last in a string	travel (0.3) vs. non-travel (1.7)

TABLE 6. The mean differences between question usage, significant at the 1% level, between informants. 5% significances have been removed for reasons given in the text.

Topic	Splitting of Targets
No. of times question 29 asked	males (17) vs. females (19)
No. of times question 63 asked	males (0.15) vs. females (2)
No. of times question 3 was most useful	males (18.5) vs. females (13.5)
No. of times question 6 was most useful	males (0.03) vs. females (1.5)
No. of times question 3 was most useful	urban (13.5) vs. rural (18.5)
No. of times question 14 was most useful	urban (20.5) vs. rural (12)
No. of times question 21 was second most useful	urban background (1.4) vs. rural backg. (0.4)
No. of times question 6 was useful	have children (2.3) vs. no children (0.2)
No. of times question 6 was not useful	have children (4.5) vs. no children (0.2)
No. of times question 40 was not useful	travel (1.2) vs. non-travel (0.4)

TABLE 7. The mean differences between question usage, significant at the 1% level, between targets. Obviously significant splits (e.g., question 16 split by urban and rural targets) are suppressed.

TOPIC	ORDINAL VARIABLES							DICHOTOMOUS VARIABLES (opposite sign if dichotomy negated)				% OF VARIANCE ACCOUNTED FOR
	INCOME	OCCUPATION LEVEL	AGE	NO. OF HOSPITALS	NO. OF CHILDREN	EDUCATION LEVEL	DISTANCE	POPULATION SIZE	MALE	URBAN	CHILDREN	ACTIVE RELIGION	
no. of questions asked		−	+	+	+	−	+		−	+	−		44
no. of times quest. 9 asked		−	+	−	−	+	+	−	−	+		−	46
no. of times " 17 "		+	−	−	−	+	+	−	−	+		−	56
no. of " " 21 "		−	+	+	+	+	+	−	−	+		+	62
no " " " 23 "		−	−	−	+	−	+	+	+	−		−	52
no. " " " 26 "		−	+	−	+	+	+	+		−		+	58
" " " " 37 "			+	−	+	+	+	+		+		+	56
" " " " 38 "		+	+	+	+	+	+	+	−	−		+	40
" " " " 84 "		−	+	−	+	−	+	−	+	+		+	50
no. times quest. 3 most useful		+	−	−	+	+	+		+	−			55
no. " " 6 " "		−	+	+	−	−	−		−	−	+	+	51
" " " 16 " "		−	+	+	+	+	+		−	−	+	+	43
" " " 41 " "		−	+	−		−	−		+	+	+		45
" " " 84 " "		−	+		−	−	−		+	−	+	+	41
" " " 63 " "		−	+	+	−	−	−		−	−	+	+	45
probability quest. 1 was most useful (given that it was asked)	−	−	+		−	−	−	+		+		+	62
" " 3 " "	−	+	−	−	−		+		+	−		−	67
" " 14 " "		+	−	+	+	+	+	−	−	+		+	48
" " 21 " "	+	−	+	+	+	−	+	+	+	−			58
" " 26 " "	+		+	−	+	+	+	−	−	+		+	52
" " 29 " "	−	+	−	+	+	+	+	−	−	+		−	46
probability quest. 14 was second most useful		+	−		+	−	−		+	+	−	−	71
" " " 16 " "			+	+	+	+	−		−	−	+	+	51
" " " 23 " "		+	−	−	+	+	+		+	+	+	+	46
probability quest. 3 not useful	−	−	+	+	+	−	−	+	−			+	52
" " 14 " "	+	−	+	+	−	−	+		+	−		−	72
" " 17 " "	+	+	+	−		+	+	−	−	+			47
" " 20 " "	+	−	+	−		+	+	−	−	+		+	46
" " 21 " "		−	+	−	−	+	+	−	−	+		+	54
" " 22 " "	+	−	+	+	−	+	+	−	−	+		+	59
" " 23 " "	−		−	−	+	−	+	+	+	−		−	58
" " 26 " "	−	−	+		−	+	+	+	+	−		+	50
" " 30 " "	−	+	+	+	+	+	+	+	−	+		+	41
" " 37 " "	+	−	+	−	+	+	+	−	−	+		+	53
" " 84 " "	−	+	+	+	+	+	+	−	+	+		+	40

Table 8. Target characteristics related to questions asked. Sign is direction of relationship.

Cutoff	444 km	222 km	111 km	56 km	28 km	14 km	7 km	0 km
Difficult score	0.31	0.37	0.48	0.54	0.57	0.59	0.59	0.60
Easy score	0.53	0.63	0.77	0.85	0.87	0.89	0.89	0.89

TABLE 9. Average accuracy scores obtained by predicting that informant will make the 'optimal' choice.

	Weighting	Difficult Score	Easy Score
1.	Only location tags counted	0.34	0.59
2.	Only occupation tags counted	0.33	0.49
3.	Only location and occupation tags counted	0.52	0.82
4.	Only location tags counted; tags which are direct hits are given double weighting	0.32	0.46
5.	As 4, but for occupation	0.33	0.49
6.	As 4, but for location and occupation	0.47	0.70
7.	Only non-(location or occupation) tags counted	0.19	0.31
8.	All tags counted; location and occupation weighted as in 4.	0.55	0.77

TABLE 10. Average easy and difficult scores for various weightings of tags.

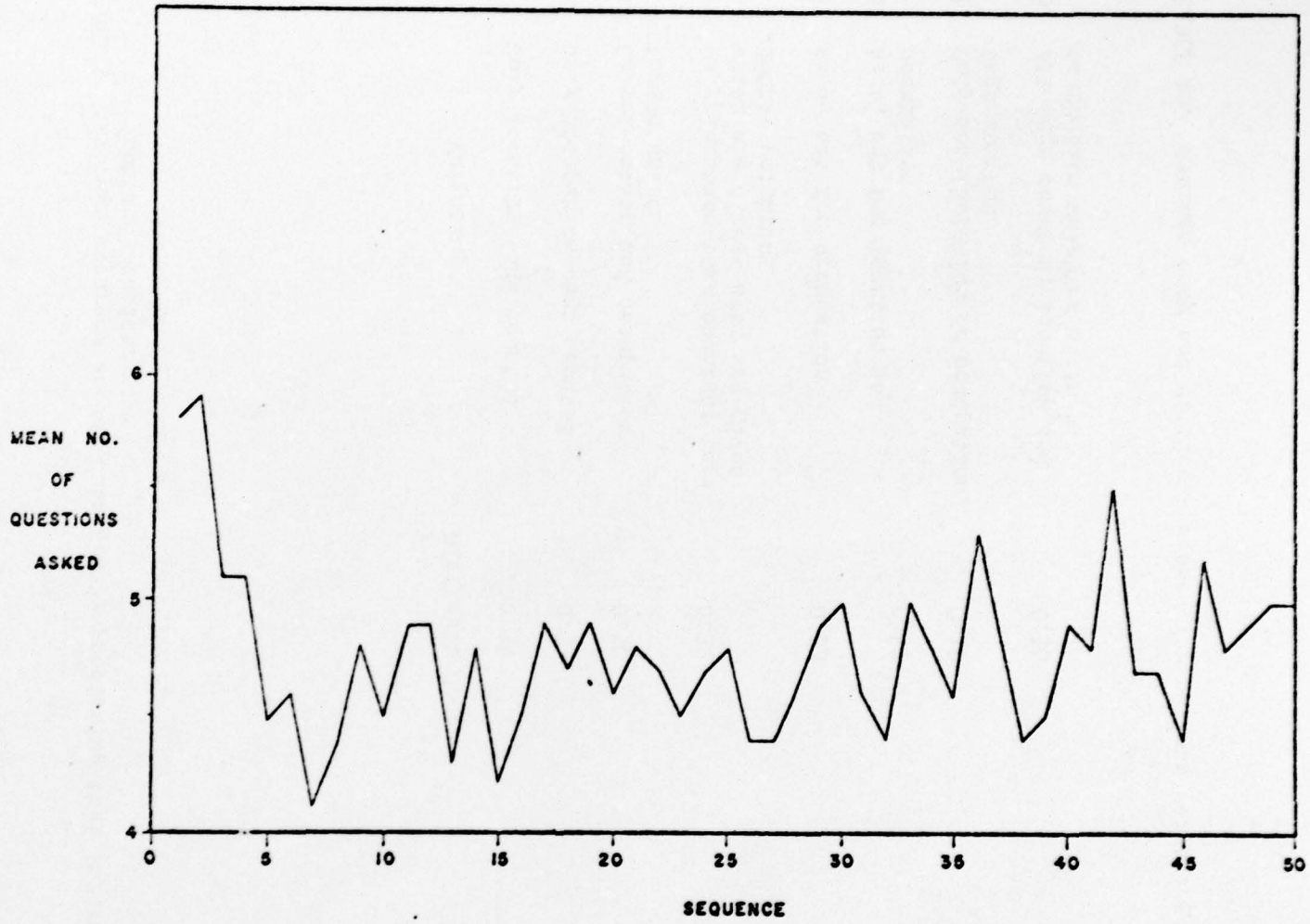


Figure 1. Mean number of questions asked about the target presented ith in sequence, for varying i.

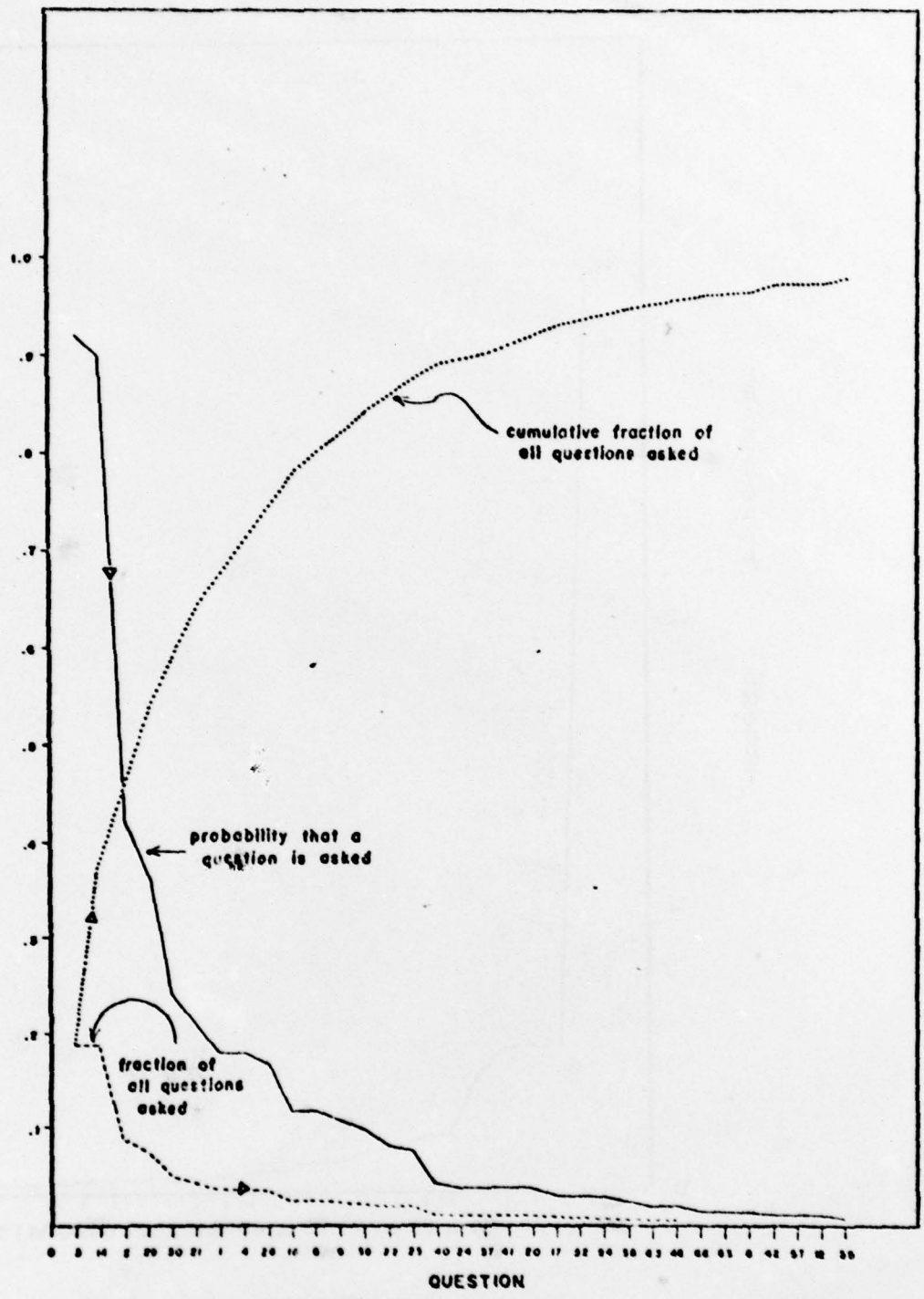


Figure 2. Question usage, by questions.

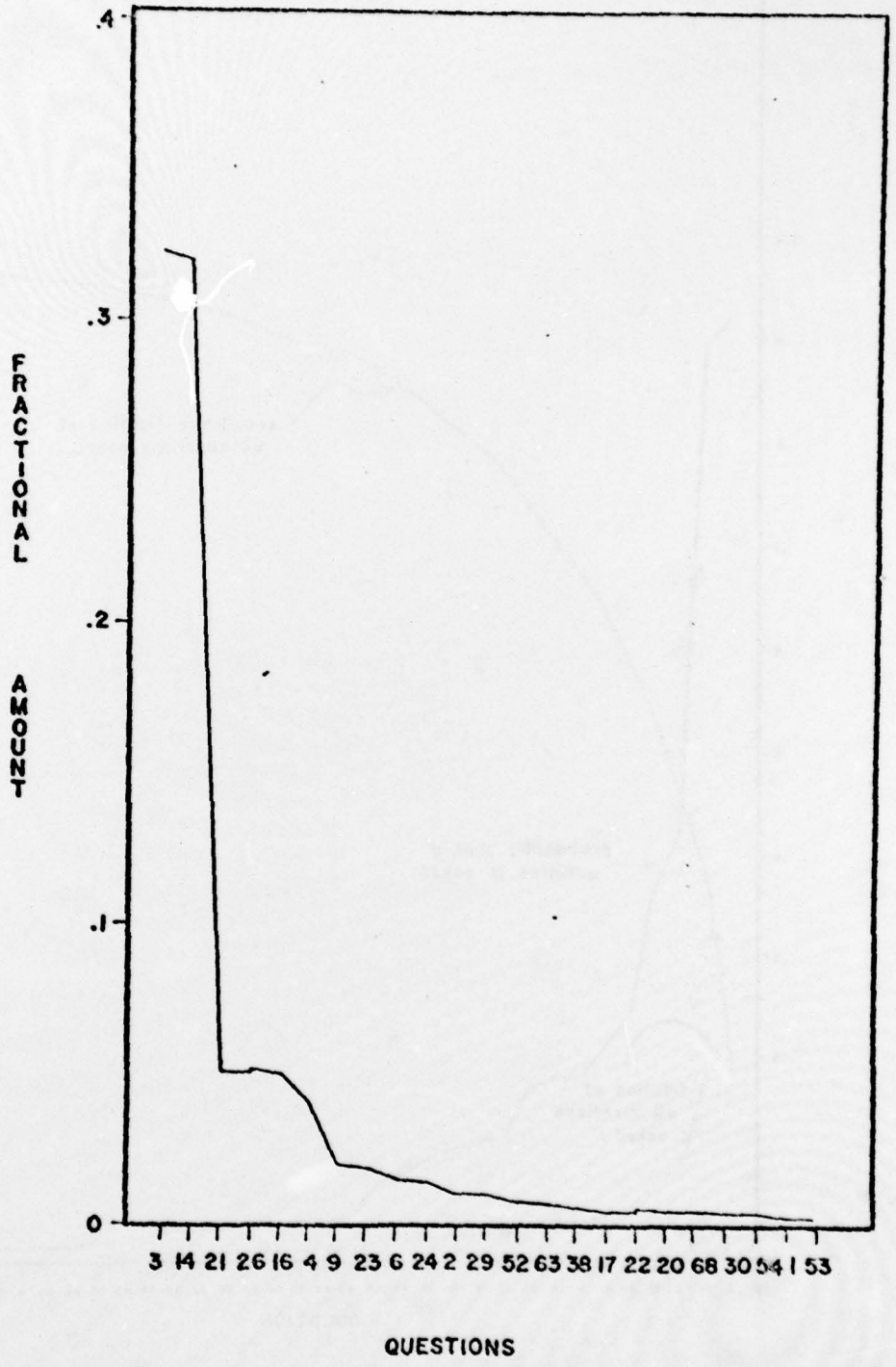


Figure 3a. Fractional amount of questions deemed "most useful" by informants.

Figure 3b. As in Fig. 3a, but for questions deemed "at all useful."

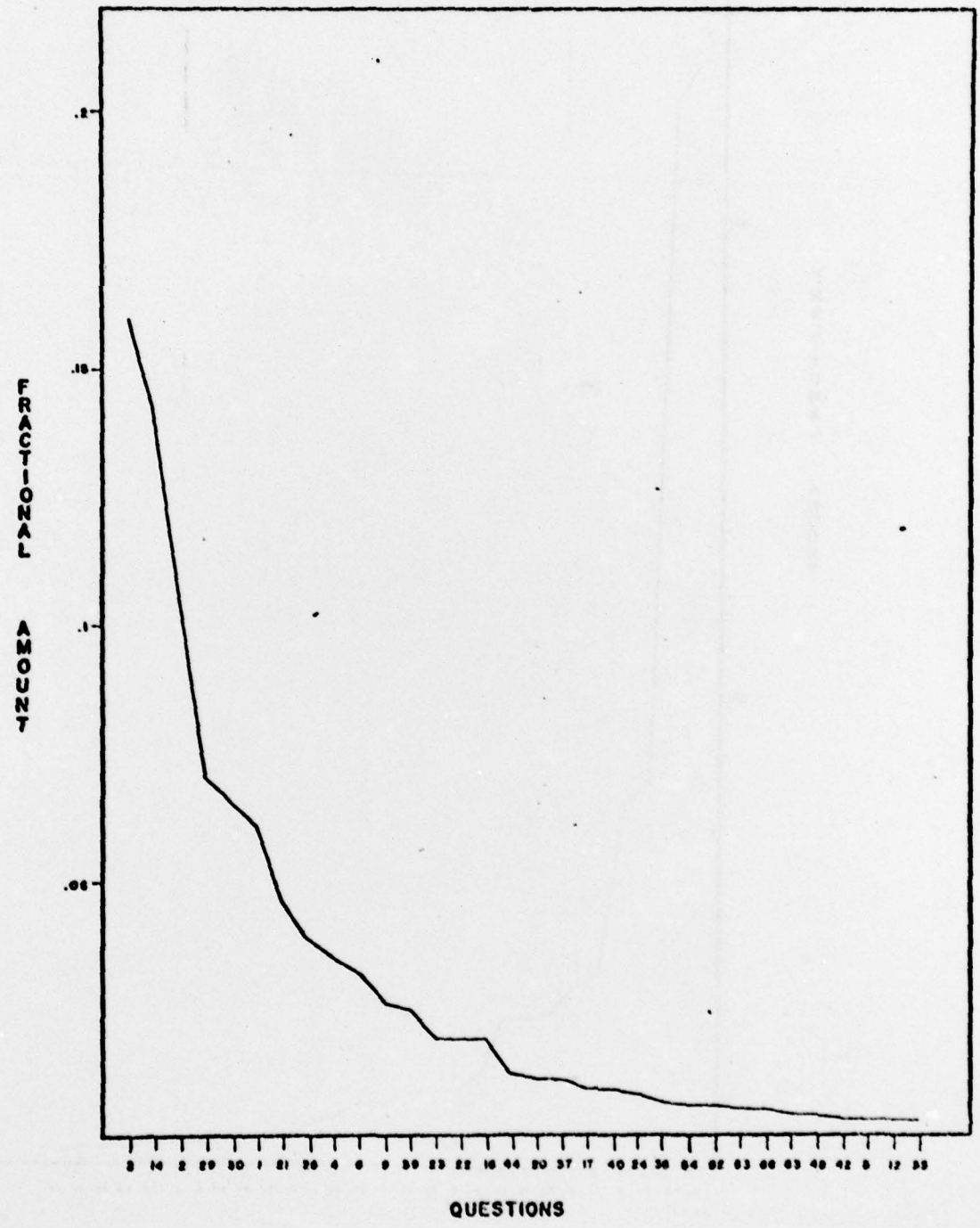


Figure 3c. As Figure 3a, but for questions deemed "not useful."

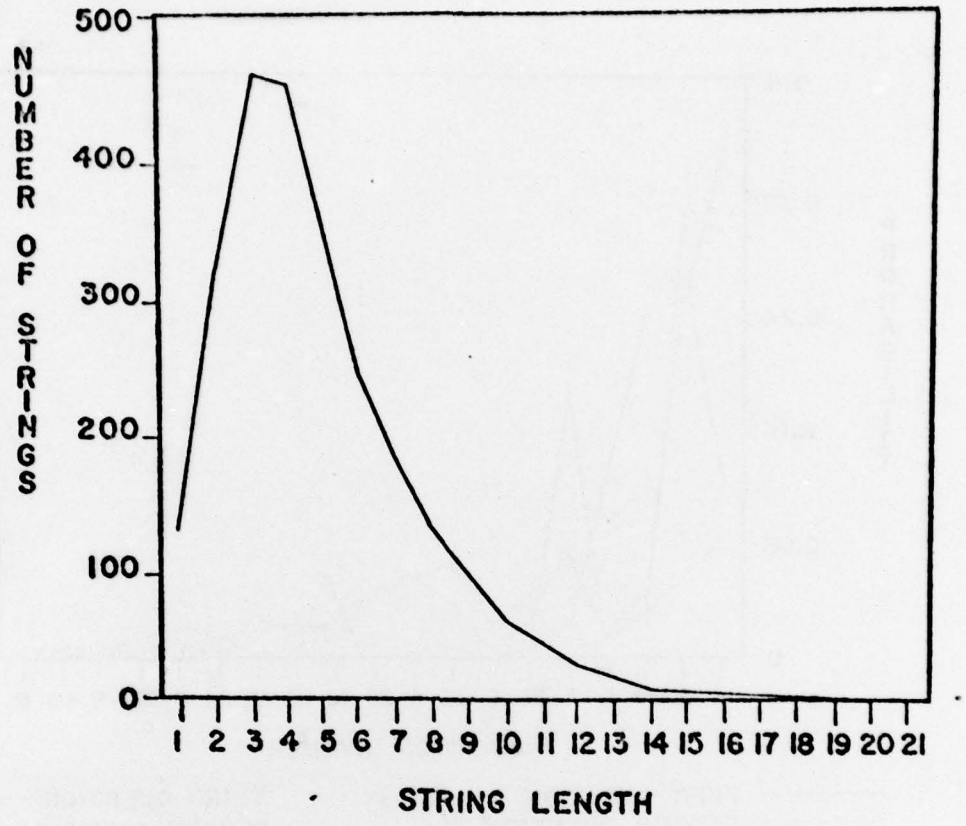


Figure 4. Distribution of lengths of question strings.

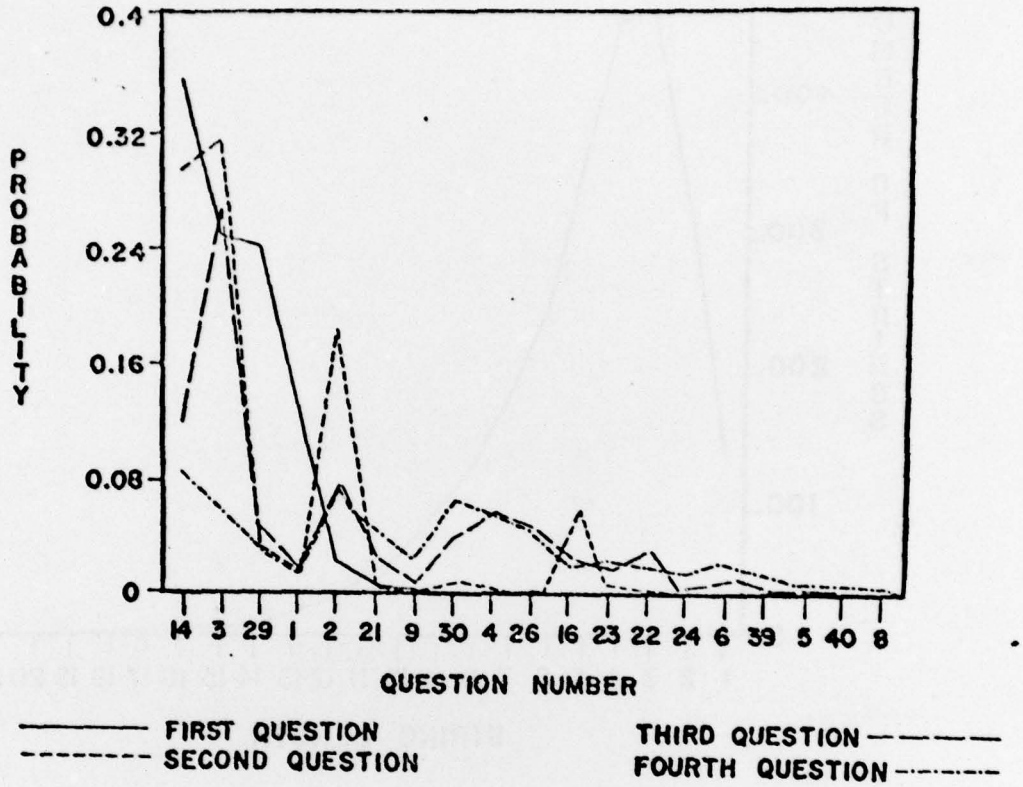


Figure 5. Probability of various questions being asked first, second, third or fourth in a string.

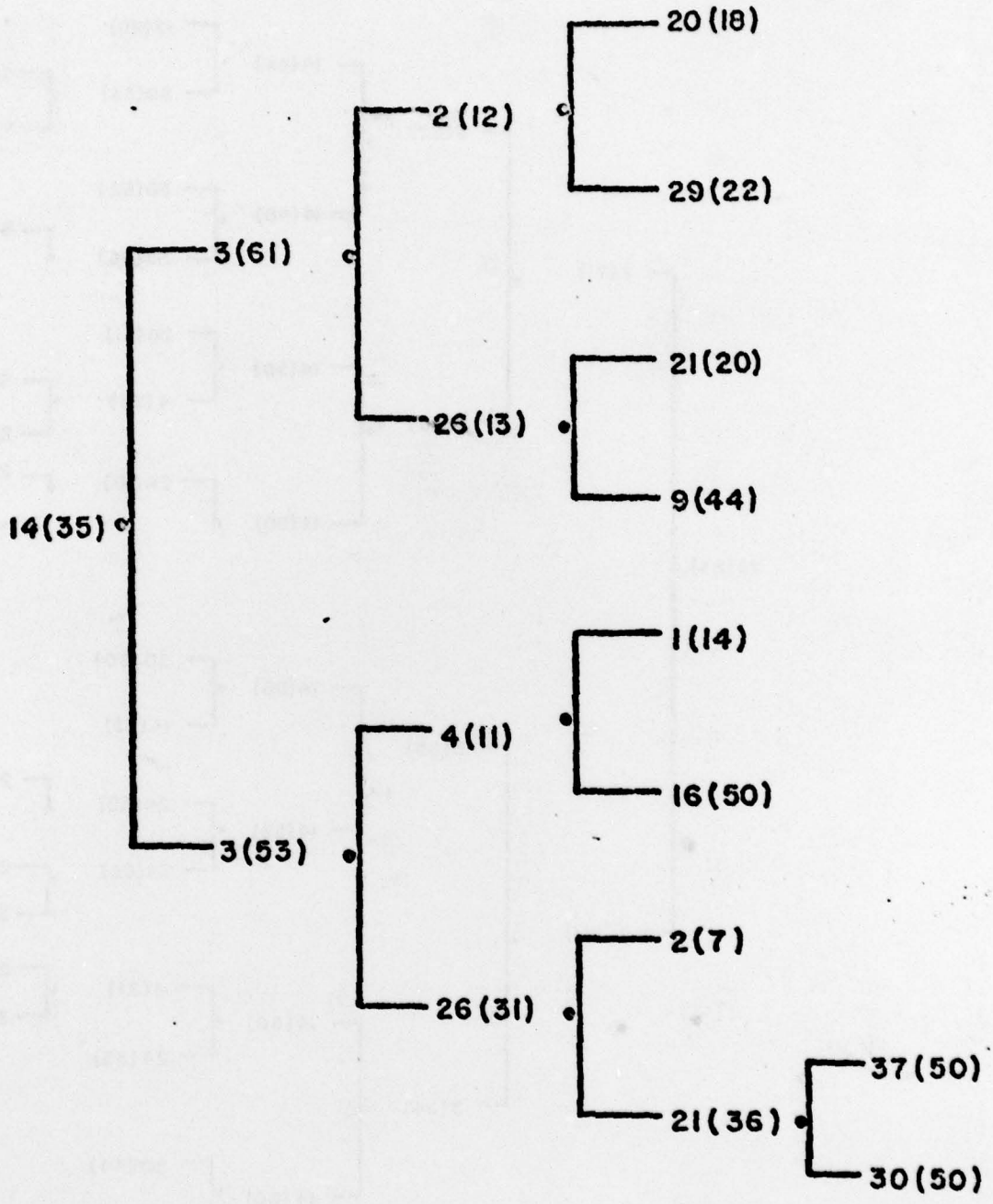


Figure 6a. Most likely sequence of questions for strings beginning with question 14. Figures in parentheses are percentage probabilities of questions being asked, given that their predecessors were asked and were either useful or non-useful (a useful response branches upwards; a non-useful downwards). Sequences end when all potential strings are exhausted, or after 6 questions.

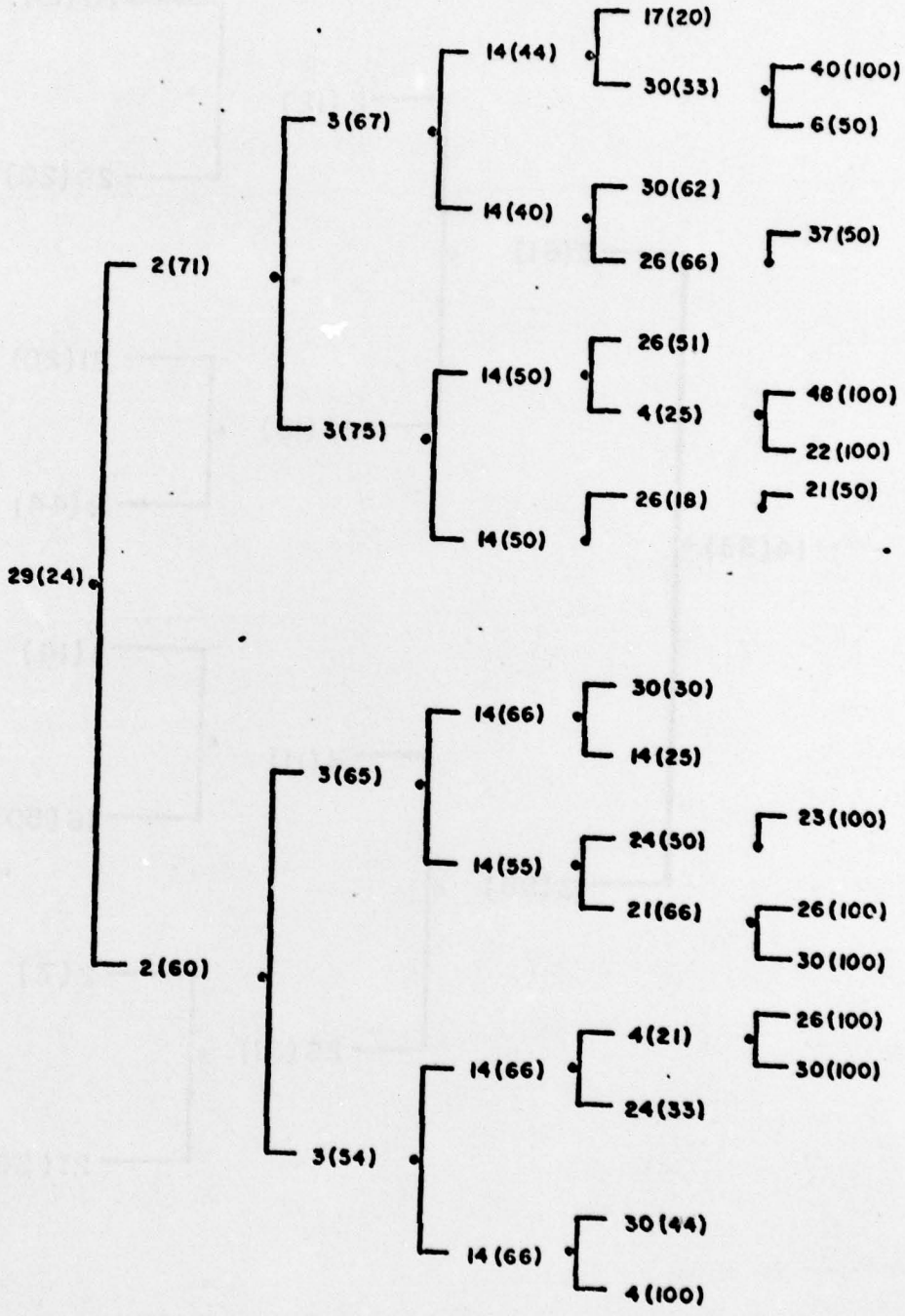


Figure 6b. As for Figure 6a, but for strings beginning with question 29.

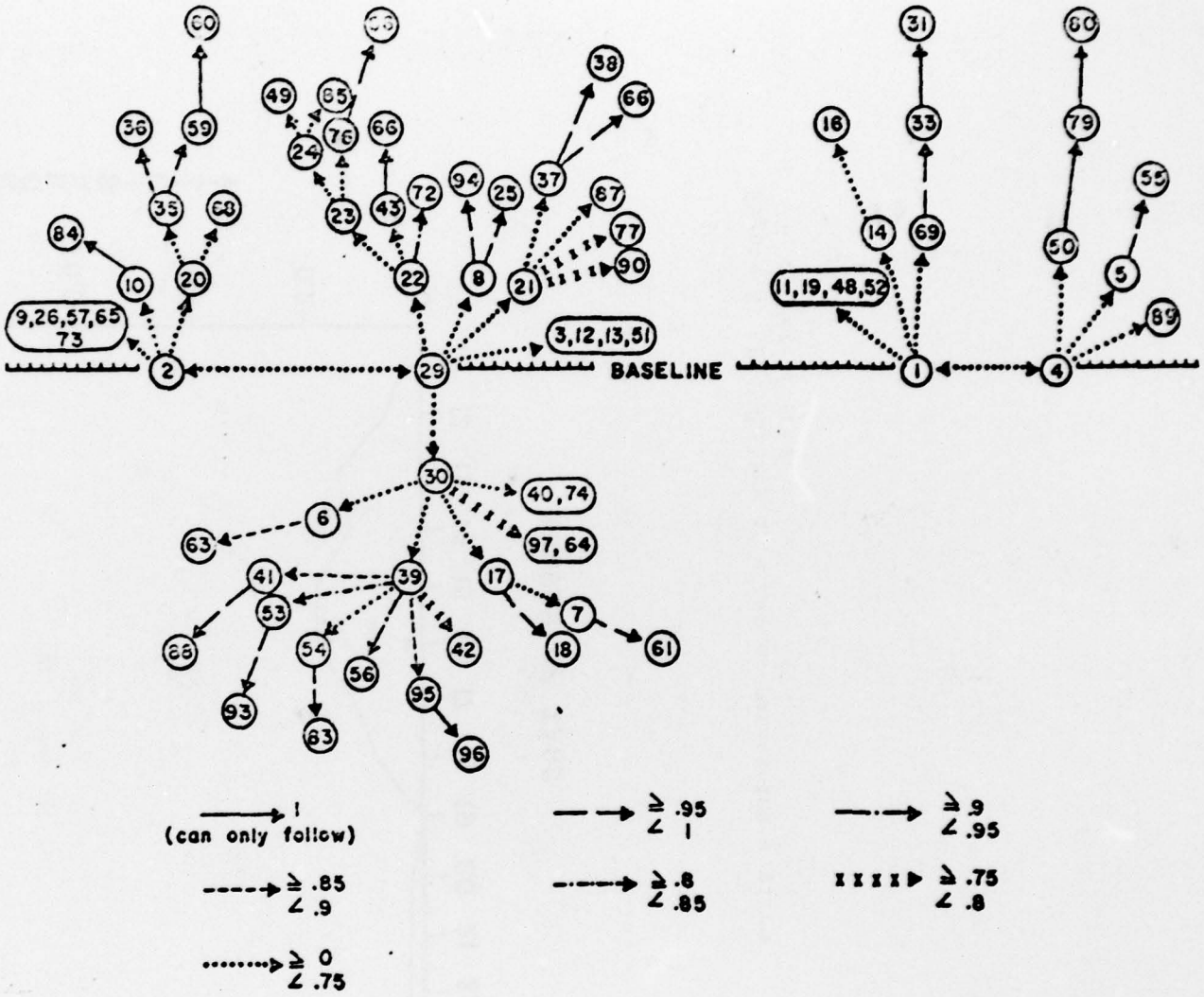


Figure 7. Chains of causality for questions. The lower half should ideally be inverted and proceed upward from 29, but for clarity it moves downwards. Question j is related to i at level p if the probability that i immediately precedes j, given j was asked, multiplied by the probability that j is not asked, given that i was not, exceeds p. Most of the 0 - 0.75 links are very weak (about 0.02).

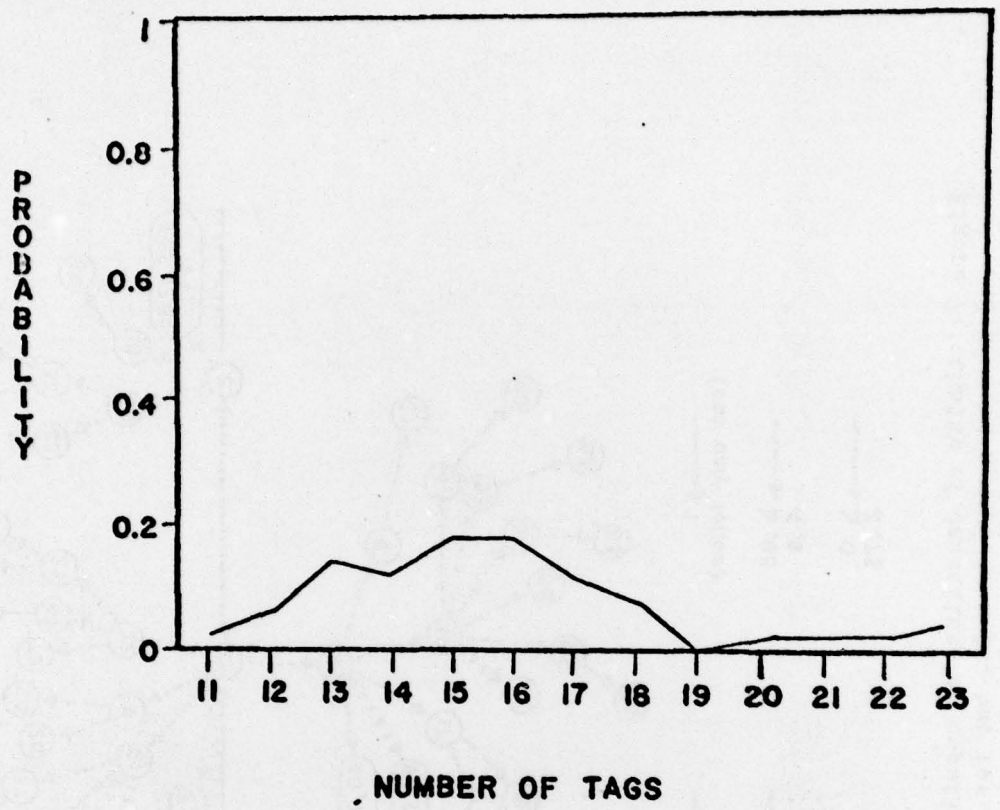


Figure 8. Probability of a target possessing a given number of tags.

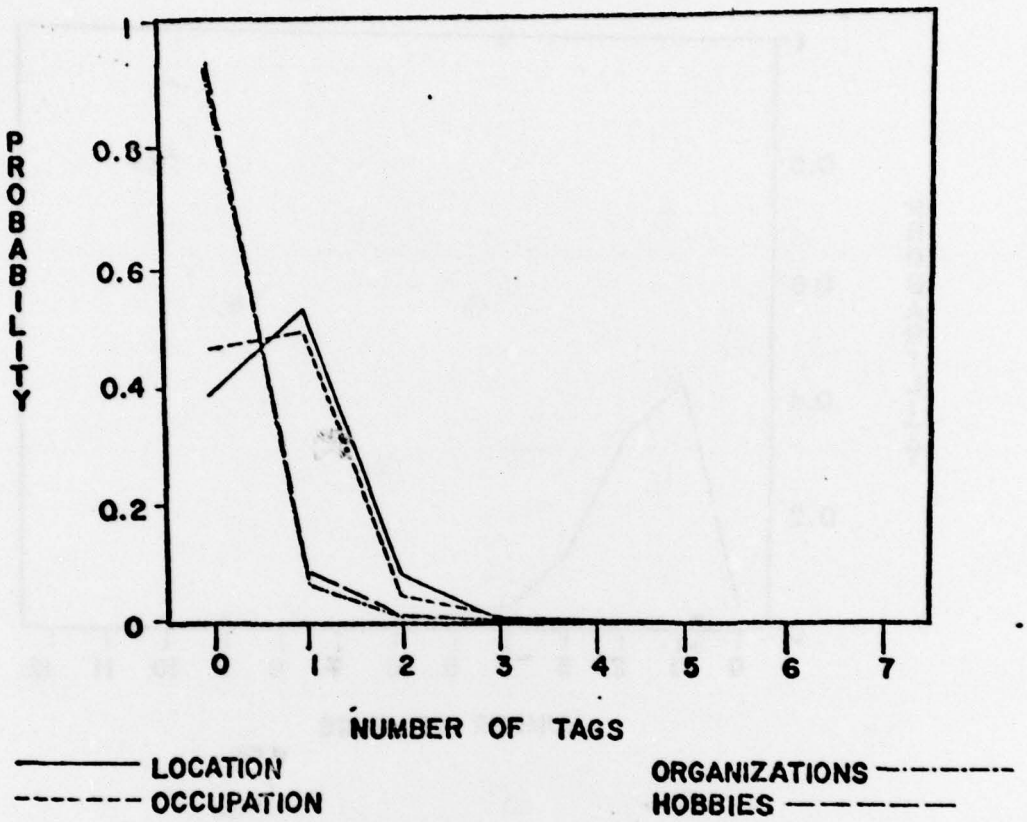


Figure 9. As in Figure 8, but distributed by different category of tag.

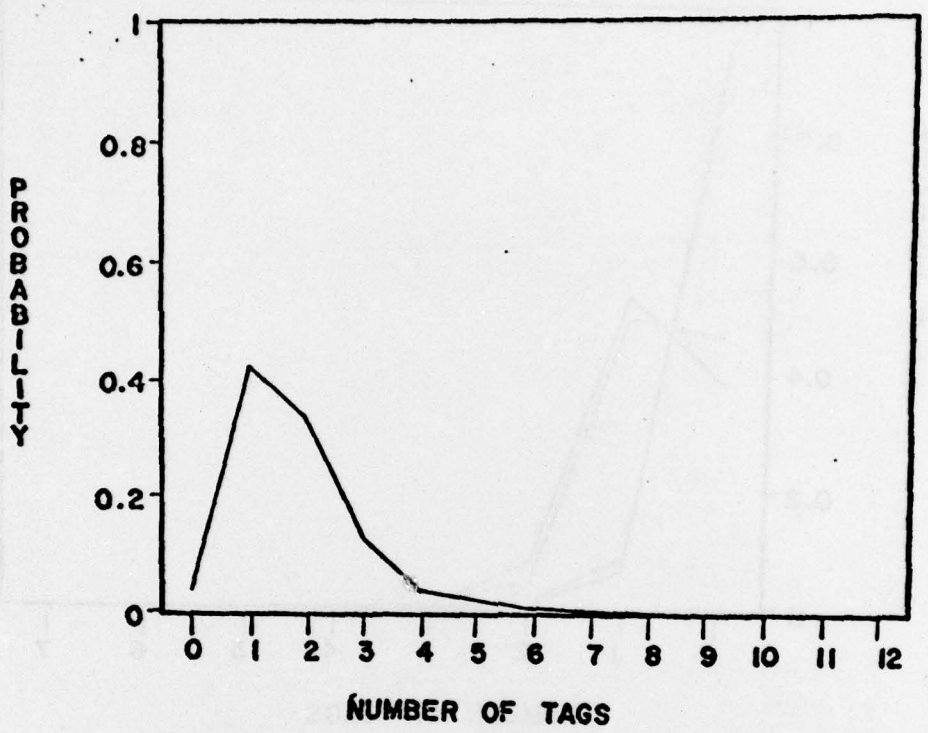


Figure 10. Probability of a choice possessing a given number of tags.

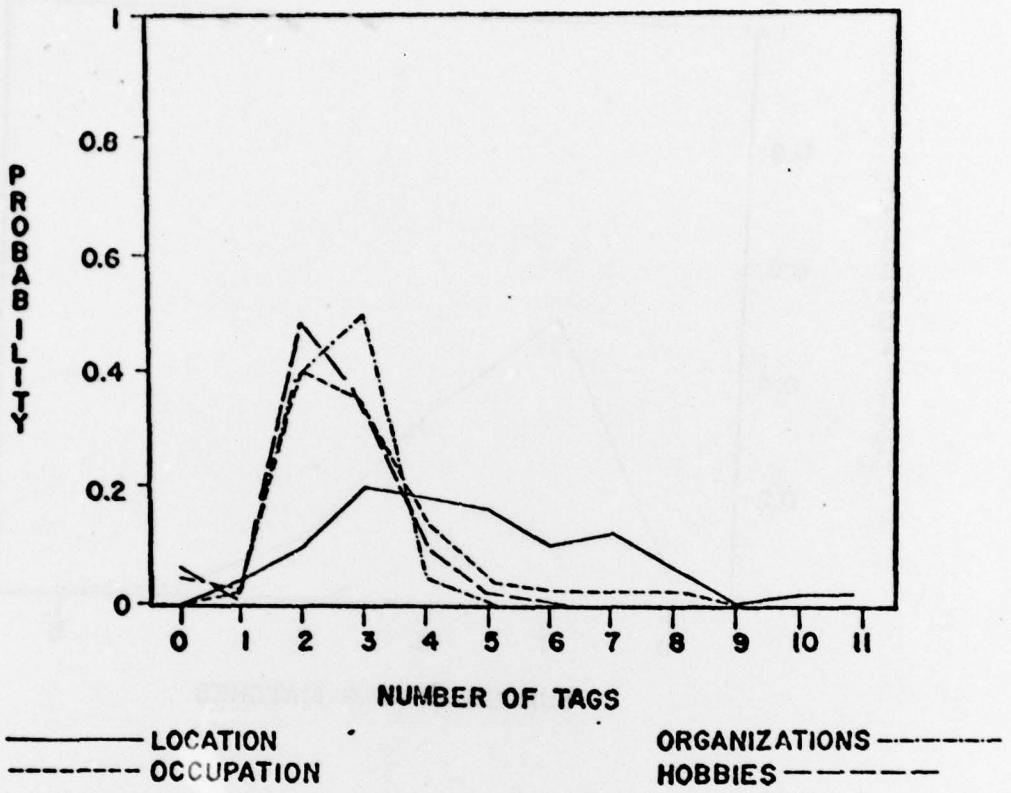


Figure 11. As for Figure 10, but distributed by different category of tag.

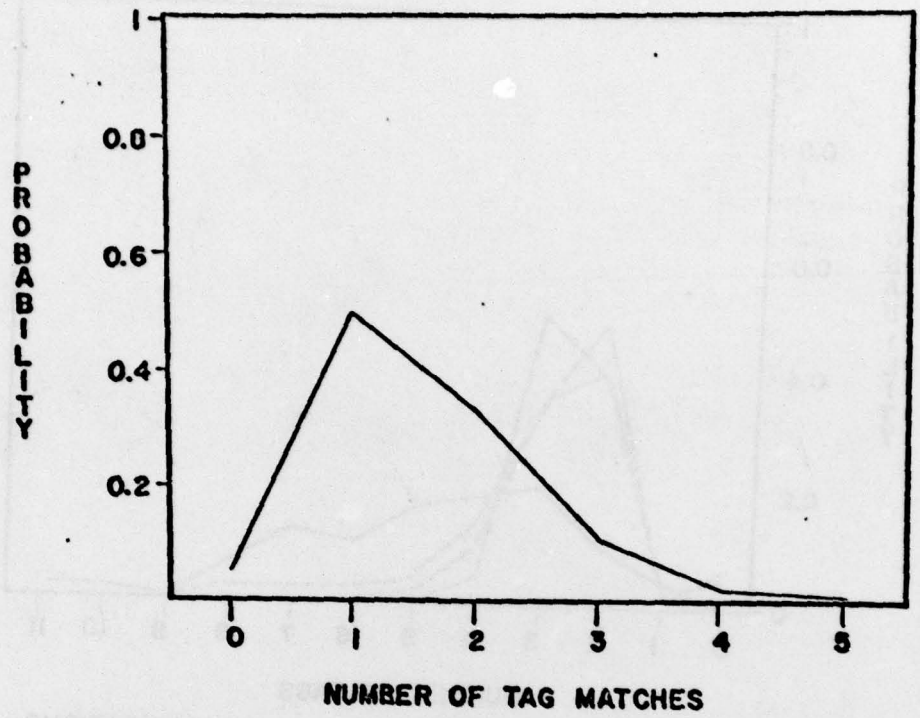


Figure 12. Probability of the correct choice having a given number of tags in common with the target.