INTER-INDIVIDUAL CALIBRATION OF FREQUENCY ESTIMATES

Albert F. Smith, Danny R. Hager, and Alison J. Palphreyman
State University of New York at Binghamton
Jared B. Jobe, National Center for Health Statistics
Albert F. Smith, Department of Psychology, Box 6000, SUNY Binghamton, NY 13902-6000

Key Words: Questionnaire design, cognitive methods

Survey respondents are often asked about the frequency of occurrence of events or behaviors during specified periods of time. A respondent might be asked how many artichokes he or she ate during the preceding month, how many times he or she visited a dentist during the preceding six months, or how many days of work he or she missed due to illness during the preceding year. All such questions presume that the respondent has in memory some relevant information that can be retrieved and--possibly with some transformation to the appropriate response format--reported (Smith, 1991).

Suppose that a respondent has been asked to indicate how often he or she has experienced each of some set of events. We can look at two principal aspects of reporting performance--the relative aspect of frequency judgments and the absolute aspect of frequency judgments. To the extent that larger numbers are assigned to events that occurred more frequently, relative frequency judgments are good. It is clear, however, that the relative aspect of frequency judgments may be superb even while the absolute aspect--the correspondence between the judgment and the true frequency--is poor.

Much psychological research on memory for frequency of occurrence has shown that the relative aspect of frequency judgments is quite good (e.g., Naveh-Benjamin and Jonides, 1986; Smith, 1991). However, for many purposes, good relative frequency judgments are not adequate--what is required is accurate absolute frequency judgments. This is the case when data is to be used to order respondents according to the frequency with which they engage in particular behaviors. Suppose, for example, that Martha eats baked potatoes more often than french fries but that the reverse is true for Naomi, who eats french fries more often than baked potatoes. Suppose, during the last month, that Martha ate 9 baked potatoes and 6 servings of french fries, and Naomi ate 9 servings of french fries and 6 baked potatoes. When asked for frequencies, suppose that Martha says 11 baked potatoes and 8 servings of french fries, and Naomi says 8 servings of french fries and 5 baked potatoes. Each of these respondents has reported numbers that preserve approximately her true baked potato to french fries ratio, but these numbers indicate that Martha and Naomi ate french fries equally often even though Naomi ate french fries half again as often as Martha. It is not difficult to imagine that Martha, who ate french fries less often than Naomi, would report a larger number for french fries than Naomi.

We hypothesize that individuals make numerical frequency judgments by mapping some internal representation of frequency onto a scale of numbers. The basic idea is that each encounter with an item establishes some kind of representation in memory, and then, when a frequency judgment is made, a sample is drawn from memory and, based on the amount of evidence for the target event in the sample, a numerical response is reported (Smith and Jobe, in press). All individuals would tend to assign larger numbers to events that occurred more frequently and smaller numbers to events that occurred less frequently--both Martha and Naomi did this. However, if different individuals use the numerical response scale in arbitrarily different ways, their absolute judgments will not interlock. Thus, a given numerical response may be used by different respondents to refer to different true frequencies. We saw this in the responses of Martha and Naomi. In some previous research on food frequency judgments, we have found that subjects are internally consistent, but when we look at judgments for particular food items over subjects, the relationships between actual and reported frequencies are substantially less clean (Smith, 1991).

If this is a reasonable account of frequency judgments--at least of frequency judgments for some classes of events--then we can ask two

questions: First, how good are relative frequency judgments? That is, to what extent can samples be discriminated from each other? And second, can we improve the way respondents map the memory evidence they collect into response scales? The studies described here examined these issues.

Study 1: Relative Judgments of Frequency

The first set of data reported here actually recapitulates Smith, Jobe, and Mingay (1989). Several reservations remained about that experiment's implications, but further analysis of that data along with new data presented here satisfy those concerns.

The question that motivated the collection of the original data was: How good are relative frequency judgments? We thought that if survey respondents have high fidelity representations of frequency information, then a paired-comparison procedure would allow them to express more detail about frequencies than was possible by using absolute judgments. Our idea was that if we found that paired-comparison judgments were consistent with a unidimensional ordering, this would support the notion that the responses were based on a high fidelity internal representation of frequency information.

Method

The data collection procedure involved two phases: Subjects first used an absolute scale to classify approximately 200 items according to the frequency with which they ate them. This scale ranged from 0--very rarely--to 9--very frequently. Then, from among the items assigned to each level of this scale, up to 8 were selected at random, and all possible pairs of these items were constructed. All of these pairs of items were arranged into a random order. In addition, pairs were constructed by taking items that had been classified into different absolute levels and these too were randomized into the list of pairs. For each pair, the subject was asked to indicate which of the two items he or she eats more frequently.

Results and Discussion

To reiterate, our concern is with whether the pattern of responses to the pairs is consistent with a unidimensional ordering of the items in each sublist. Our unit of analysis is the item-triple, and for each triple, we ask whether the judgments are consistent with an ordering among the items. For example, for the triple including items A, B, and C, three pairs were judged by the subject, A-B, A-C, and B-C. Responses are consistent with an order only if one item was chosen as the more frequent twice, one once, and one not at all (see Nelson & Narens, 1980).

The results were compatible with the notion that there is a memory representation that supports internally consistent judgments of relative frequency. For the pairs generated by taking items that had originally been classified differently in absolute frequency, the proportion of triples in which intransitivities occurred was only .06, and the average correlation between the frequency scale derived from the paired comparison judgments and original judgments was .82. The proportion of triples from within absolute frequency levels in which there was intransitive responding was only .17. Chance responding would have resulted in a value of approximately .72.

These data suggested that individuals respond to this task on the basis of high fidelity representations of frequency--that people have access to and use such representations. Here, two additional pieces of support for this conclusion will be presented.

The first is the pattern of discrepancies from perfect linear orders. For each subset of items, we can order the items according to how often they were chosen as the more frequent: For a set of eight items, if responses over the set are consistent with a linear order, one item will be chosen as more frequent 7 times, one 6 times, and so on down to one item that will never have been chosen.

If we arrange the paired comparison choices in a matrix, ordering the rows by overall number of choices, we can look at the pattern of violations of the overall scheme. If discrepancies from an order are random, then violations would not conform to any particular pattern. For example, the overall least frequent item in a set might be chosen as more frequent than the overall most frequent. However, if violations result from discrimination failures between ordered items, then violations would be confined to between adjacent items.

To evaluate the pattern of violations in the frequency judgment task, we cumulated, over subjects and categories, the optimized dominance matrices. Figure 1 shows, as a function of steps in

the presumed order, the average number of violations per cell per matrix. Each 8 x 8 matrix contains 7 one-step cells, 6 two-step cells, and so forth. From the 18 subjects, there were 142 sublists of eight items that contributed to the within-category data; each of the 18 subjects contributed a matrix to the between-category data. Thus, the values in Figure 1 were obtained by cumulating all of the violations for a particular step-size and dividing by the number of cells that contributed. The simulation was insensitive to the distance between items being compared, but was programmed to violate the order at the same rate that the subjects violated the order within categories. For both sets of judgments, the violations tend to be for items that are near to each other in the order rather than far apart, and the simulation shows that this is not an artifact of optimizing the matrix according to the order. The pattern of results observed here is believed by several researchers to indicate the presence of some ordered array of items in memory (e.g., Potts, 1974; Trabasso and Riley, 1975).

Given that subjects were responding according to some ordered relationship, the second concern was this: Suppose that subjects in this experiment were not responding according to *frequency* at all, but rather were using arbitrary rules or clever strategies (e.g., food preference) to make the paired-comparison judgments. This seems implausible, but it is a possibility. To determine the viability of this account, we conducted Study 2 in which subjects were instructed to respond to pairs of arbitrary items as if they were in an order. This is in contrast to the procedure of the food experiment in which we asked whether the subjects' judgments implied an order among the items--presumably by frequency.

Study 2: Construction of Ad Hoc Orderings

Method

The basic design was as follows: 64 words, representing food items, were partitioned into 8 sublists of 8 items each, and the 28 possible pairs of words were generated from each sublist. Each of 20 introductory psychology subjects were told that they would see pairs of words to which they should respond as if the words were in a line. They were instructed that for each pair, they should indicate which word was closer to the front of the line. (We motivated subjects to perform this task by telling them that a large prize would be paid for each order they generated that matched a previously generated order.)

Data were collected in two different conditions: In the *blocked* condition, all 28 pairs generated from each sublist were presented before those from the next list were presented. In the *mixed* condition, the pairs from the eight sublists were intermixed--this is the analog of the situation that we presented to subjects in the food frequency experiment.

Results and Discussion

Figure 2 shows performance by list; the abscissa shows the order in which the list was encountered by the subject. (This is meaningful for the blocked subjects, but not really so for the mixed subjects. For the mixed subjects, order of presentation was determined by the sequence in which the first pair of words from each sublist was presented). The difference between the *blocked* and *mixed* conditions was statistically significant ($F(1,18) = 6.992$, $p < .05$). When the pairs for a list were blocked, subjects were quite capable of constructing arbitrary orders--that is, responding to pairs as if the items were in an order--but when pairs from the different lists were intermixed, performance suffered. Given that performance in Smith, Jobe, and Mingay (1989) was quite good despite intermixing of pairs from different sublists, we draw the conclusion from the present experiment that subjects in Study 1 did not construct arbitrary orders. Although we know from other experiments that it is possible for subjects to apply strategies that lead to nearly equivalent performance in the mixed and blocked conditions, we think that this was unlikely to have occurred in the food experiment.

In sum, the new analysis and this complementing experiment lead us to conclude that subjects have high fidelity representations of relative food frequency that were used in the paired-comparison task.

Study 3: Experimental Control of the Response Function

The next question concerns how the information in memory is used when the individual is asked to map the representation into the scale of real

numbers. The following experiment, we believe, supports the notion that the mapping operation is separate from the gathering from memory of evidence concerning frequency.

The underlying rationale of this experiment is as follows: If we can show that we have experimental control over how subjects map memorial evidence concerning frequency into numbers, then perhaps the best procedure for eliciting high quality judgments from people in surveys would be to have them anchor their frequency judgments on events about whose frequencies they have a high degree of certainty. Specifically, we investigated whether providing anchoring information to a subject about the frequencies of a small number of events would influence the subject's entire response function.

Method

In the acquisition phase of the experiment, each of 16 introductory psychology subjects were exposed to a long series of events--words presented on a computer screen. To insure subjects' attention to these words, we required that they recite each word aloud, and we monitored them throughout this task. These long lists included target words, the frequencies of which are controlled, intermixed with irrelevant words.

In the test phase, the subjects were asked to indicate how often each target word occurred during the acquisition phase. Prior to this frequency test, however, we provided anchoring information--information about the frequency of occurrence of two of the target words--one word that was presented relatively infrequently, and one word that was presented relatively frequently. This information remained on the computer screen during the entire frequency test. In this experiment, the anchoring information was false, so we could determine whether the false anchoring information would influence subjects' frequency estimates.

In the acquisition phase of the present experiment, subjects saw 720 words. Each of 6 target words occurred with frequencies of 4, 8, 12, 16, and 20. Thus, 360 presentations of target words were intermixed with 360 presentations of non-target words, each of which occurred at least once. Therefore, the subjects encountered equal numbers of target and non-target words and the average frequency of occurrence of the target and non-target words was identical.

The anchors defined the experimental conditions: We call these the *slope-increasing* and *slope-decreasing* conditions according to the intended effects of the anchors on the subjects' performance. In the slope-increasing condition, we provided false low frequency information for a target item that occurred with relatively low frequency and false high information for an item that occurred with relatively high frequency. In the slope-decreasing condition, we provided false high information for an item that occurred with relatively low frequency and false low information for an item that occurred with relatively high frequency.

Results and Discussion

The average judgments of subjects in the two experiments are shown in Figure 3. The false anchoring information was effective in influencing the mean slope. The difference between the slopes was significant ($t(14) = 2.393$, $p < .05$). Figure 3 shows that the slopes of subjects in the slope-increasing condition are significantly higher than those of subjects in the slope-decreasing condition. We could be considerably more analytical about the impact of the anchors on the processes subjects used to make their judgments, but for now it is sufficient to note that these anchors influenced judgments about the *other* items. Note that judgments about anchor items are excluded from these data.

By providing anchors, we have influenced the mapping of frequency representations to numerical responses. In the next phase of our research program, we will evaluate whether subjects who are exposed to the same events and presented with the same anchors provide frequency judgments that are more consistent with each other than do subjects given no anchoring information. This would show that anchoring information can bring the response scales of different individuals into congruence with each other.

General Discussion

The problem on which we are working is this: Is there a way to get different people to use the scale of real numbers in the same way when they estimate how often they have engaged in various behaviors or experienced various events? We have tried first to increase our confidence that some data that we reported previously really do indicate that people have high fidelity representations of

frequency information. We then showed that we can control experimentally the way in which frequency information is attached to real numbers: Although considerably more could be said about how this is done, we are satisfied that providing anchors does influence the subject's set of judgments. The key demonstration is that the anchoring information influences judgments about targets for which no anchoring information is given. The next step will be to show that by providing the same anchors to different people, their judgments about the frequencies of a set of events can be brought into better congruence than when such information is not provided.

References

Naveh-Benjamin, M., & Jonides, J. (1986). On the automaticity of frequency encoding: Effects of competing task load, encoding strategy, and intention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 378-386.

Nelson, T. O., & Narens, L. (1980). A new technique for investigating the feeling of knowing. *Acta Psychologica, 46,* 69-80.

Potts, G. R. (1974). Storing and retrieving information about ordered relationships. *Journal of Experimental Psychology, 103,* 431-439.

Smith, A. F. (1991). Cognitive processes in long-term dietary recall. *Vital and Health Statistics,* Series 6, No. 4 (DHHS Publication No. 92-1079). Washington, DC: U. S. Government Printing Office.

Smith, A. F., Jobe, J. B., & Mingay, D. J. (1989). A cognitive investigation of responses to dietary surveys. *Proceedings of the Section on Survey Methods Research, American Statistical Association,* 407-412.

Smith, A. F., & Jobe, J. B. (in press). Reports of long-term dietary memories: Data and a model. In N. Schwarz & S. Sudman (Eds.), *Autobiographical memory and the validity of retrospective report.* New York: Springer-Verlag.

Trabasso, T., & Riley, C. A. (1975). The construction and use of representations involving linear order. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola symposium.* Hillsdale, N.J.: Erlbaum.

Figure 1. Pattern of Violations of Linear Order of Food Frequency Relative Judgements

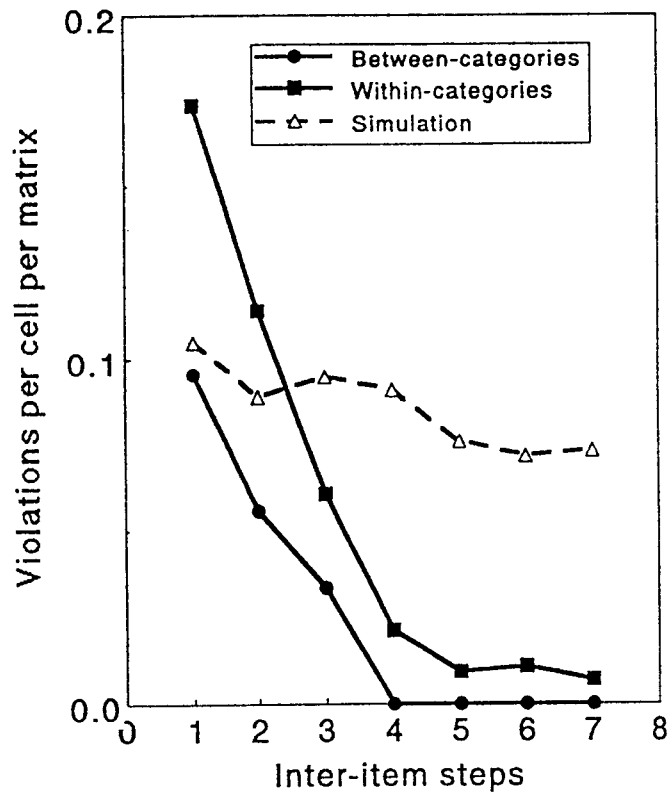

Figure 2. Proportion of Intransitivities by Sequence

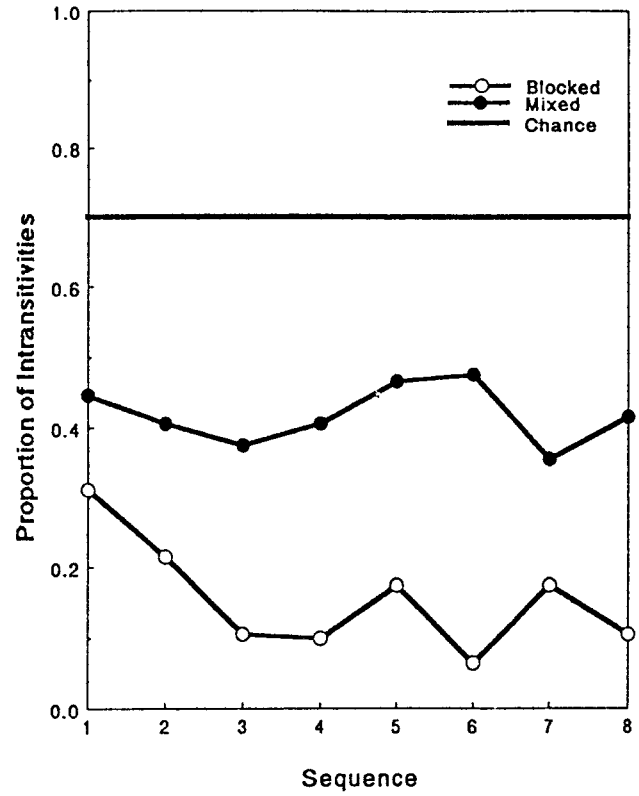

Figure 3. Mean Frequency Judgements