STRATEGIES FOR ESTIMATING CATEGORY FREQUENCY: EFFECTS OF ABSTRACTNESS AND DISTINCTIVENESS[1]

Frederick G. Conrad, Bureau of Labor Statistics, Norman R. Brown, University of Alberta
Frederick Conrad, BLS, 2 Massachusetts Ave. NE, Room 4930, Washington, DC 20002

Introduction

We all divide our worlds into categories. However, our mental categories may not exactly correspond to those in which researchers are interested. Under these circumstances, data quality seems likely to suffer. One point of this paper is to explore how different types of categories affect peoples' accuracy when they report frequencies for those categories.

Survey respondents are often asked to report their frequency of activity for particular categories of events or objects (*e.g.*, Blair & Burton, 1987; Burton & Blair, 1991). In order to answer "How many magazines did you purchase last month?" one must determine which publications qualify as magazines and report a number for all of those items but no others. If there is more than one way to answer such questions, that could affect the accuracy of respondents' estimates.

In recent years, it has been demonstrated that respondents produce frequencies by either counting retrieved memories -- an Enumeration strategy -- or applying rate of occurrence knowledge -- a Rate strategy (*e.g.*, Blair & Burton, 1987; Burton & Blair, 1991; Means & Loftus, 1991; Menon, 1993). More recently, it has been shown that they also rely on a non-numerical sense of magnitude (Brown, 1994; Conrad, Brown & Cashman, 1993). When instructed to verbalize their thinking, respondents and experimental subjects sometimes justify their estimates with statements such as "that happened a lot," or "that was very rare." Under certain conditions, respondents predominantly rely on such "general impressions." A second point of the current paper is to monitor the use and accuracy of strategies that rely on non-numerical information, especially for different types of categories.

The Experiment

Rationale. Categories in many survey questions differ in their level of abstraction. For example, a category such as *Poultry* is more abstract than a category such as *Chicken* which is itself more abstract than a category like *Chicken Parts*. To answer a question about poultry purchases, a respondent might need to consider their chicken and turkey purchases if that is how they have structured the relevant information in their memories (Felcher & Calder, 1992). However it is also conceivable they might need to further decompose those categories to retrieve frequency-relevant information.

One influential view of mental categories (*e.g.* Rosch, Mervis, Gray, Johnson & Boyes-Braem, 1976) holds that there is an optimal or "basic level" of abstraction, midway between the most abstract and most concrete categories, which people overwhelmingly prefer to use (for reviews, see Barsalou, 1991 and Lakoff, 1987). In the current study we compare frequency reports for basic level categories to those for higher level or superordinate categories. The effects on frequency judgments of certain non-basic level categories have been reported (Barsalou & Ross, 1986), but to our knowledge, this simple variation in level of abstractness has not been explored.

In the current study we look at the interplay between category abstractness and the performance of different response strategies. The many strategies used by respondents to answer frequency questions are applied under identifiable circumstances (Conrad, Brown & Cashman, 1993; Menon, 1993; Means & Loftus, 1991). Enumeration strategies are observed primarily when events are distinctive, occur on an irregular schedule and are low in actual frequency. Rate strategies require that rate information be available which is only likely when events take place on a regular schedule. General impressions seem to be used when events are not distinctive and happen irregularly. They are especially likely when these conditions are present and actual frequency is high. In this study, we control these conditions through an experimental technique (Brown, 1994). In particular, we vary the distinctiveness of a set of common, consumer products from both basic level and superordinate product categories, and present them on an irregular schedule.

Design. The subjects were presented with a sequence of product names, one at a time, to study for a later memory test. They were then required to report the frequency of study products in each of a number of product categories. So, for example, if the product

[1]We thank Erin Cashman for her help collecting data.

category on which the subject was being tested was *Newspapers* the subject would base a response on the number of times the study item had been a newspaper. The subjects were in one of four experimental conditions defined by two factors: level of abstractness and level of distinctiveness. The first of these factors, abstractness, concerned the type of product category presented in the test phase. Subjects were asked about either basic level (for example, *Newspapers*) or superordinate categories (for example *Reading Material*). The second factor, distinctiveness, was introduced through the presentation of the study items. To make category members relatively distinctive, different products were presented one time each for a particular category, for example, *Washington Post, New York Times, Baltimore Sun*. To make them similar to one another -- that is, not distinctive -- a single product per category was presented multiple times, for example *Washington Post, Washington Post, Washington Post*. The four conditions therefore are referred to as Basic-Same, Basic-Different, Superordinate-Same and Superordinate-Different.

The study products were chosen to be members of both basic level and superordinate test categories. For example, *Washington Post* is an instance of both *Newspapers* and *Reading Material*. The basic level categories were chosen so that more than one could share a superordinate category. For example, *Newspapers*, *Magazines* and *Reference Books* all share the superordinate category *Reading Materials*. There were 36 Superordinate categories, each of which were associated with one, three or four basic level categories. Product frequency for the basic level categories was counterbalanced so that the product frequency for a superordinate was not related to the number of basic level categories with which it was associated. Overall frequency for the superordinate categories was either four, seven, nine or twelve. The stimuli are available from the authors.

The sequence in which the study products were presented was random with the constraints that products from the same basic level category be spread throughout the sequence in roughly even intervals and products from the same superordinate (but different basic level categories) be separated by at least one product from a different superordinate. items, in effect, occurred irregularly. Altogether, the unique study sequence was generated for each subject.

The test sequences were essentially random. Twenty per cent of the test categories were "catch trials," that is, no study items from those categories had been presented and therefore the frequency was zero. In the Basic-Same and Basic-Different conditions, the test sequence included 42 trials and a unique sequence was generated for each subject. In the Superordinate-Same and Superordinate-Different conditions, the test sequence consisted of 17 trials. In all conditions, the first two test trials were treated as practice.

Procedure. Each product in the study phase was presented on a computer screen for six seconds. In addition to the product, its basic level category was also presented above it on the screen to reduce ambiguity about category assignment. The subjects were instructed to study each product-category pair for a later memory test, but they were not told the nature of this test.

In the test phase, the subjects were instructed to report the number of times they had been presented products from each test category. They were instructed to report zero when no study items had been presented for a test category. They were encouraged to be as accurate as possible and take as much time as they needed. Each test category appeared on the computer screen until the subject typed in a frequency and pressed the Enter key. The subjects were instructed to verbalize their thinking while arriving at a frequency, that is, to provide concurrent, verbal protocols, and an experimenter was present to assure that they kept speaking (Ericsson & Simon, 1993). The subjects were instructed to complete their verbal report prior to entering a frequency response. Their protocols were tape recorded and their frequency responses were recorded by the experimental software.

Subjects. Thirty two subjects were recruited from an advertisement placed in the *Washington Post*. Eight were randomly assigned to each of the four conditions. All of the subjects (except three who were employed by U.S. government) were paid $25; government employees were not compensated and were assigned to different conditions.

Predictions. Since the study items were not presented in a regular sequence, Rate strategies are not possible and therefore we do not expect to observe them in the protocols. When individual episodes are distinctive in peoples' memories, it is possible for them to be retrieved and enumerated. The Basic-Different and Superordinate-Different conditions create such circumstances, and so we expect Enumeration to be prevalent in these conditions. While subjects might have other information available to them in this condition, such as general impressions, we expect them to enumerate whenever it is possible.

In the Basic-Same and Superordinate-Same conditions, it is difficult, if not impossible, for subjects to retrieve and count episodes. Each time they are presented a product from a test category, it is the same product. There is nothing to distinguish one

presentation from another. Under these circumstances subjects are likely to rely on non-numerical frequency information. In particular, we predict a high incidence of General Impression statements in the protocols and some kind of Memory Assessment -- judgments based on memory processes and states, rather than content. A well-publicized example of Memory Assessment is the availability heuristic (Tversky & Kahneman, 1973) in which subjects' estimates are based on the ease of retrieving category members. Unfortunately think aloud procedures are not sensitive to Memory Assessment strategies because estimating frequency on the basis of Memory Assessment does not require people to be aware of its use. Therefore, in the two Same conditions, we also expect a relatively high number of uninformative protocols.

Turning now to accuracy, we predict that subjects will underestimate actual frequency in the Basic-Different and Superordinate-Different conditions. This should occur because Enumeration will be prevalent and should lead only to errors of omission: It is much more likely that subjects will forget a pertinent study episode than will "recall" one that never actually occurred. We expect that in the Superordinate-Different condition subjects will decompose superordinate test categories into their component, basic level categories. This should lead to more underestimation in the Superordinate-Different than Basic-Different condition because in the former condition such error can arise when subjects forget entire basic level categories and when they forget products from within categories that they do recall (Tulving & Pearlstone, 1966).

We expect a different pattern of results for the Basic-Same and Superordinate-Same conditions because we expect these subjects to rely predominantly on strategies other than Enumeration. Our view of non-numerical strategies is that they first involve retrieving an impression or forming one via Memory Assessment. Subjects must then convert this sense of magnitude into a number. If people lack a metric for this conversion, then in general, they can be quite inaccurate: Subjects can either underestimate or overestimate but the overestimates should be larger, leading to net overestimation. This is because underestimates cannot be smaller than zero so the largest underestimate possible is 100%; overestimates, are essentially unbounded and can, in principle, be many times larger than the actual frequency[2].

[2]Subjects are told the number of study items which may provide a practical upper bound on their estimates.

On the basis of this reasoning, we expect large overestimation in the Basic-Same condition. In the Superordinate-Same condition, the ability to enumerate individual, basic level categories could reduce overestimation. In particular, subjects can forget entire, basic level categories or use the number of categories that they do recall as a kind of anchor. In fact, it is hard to know how the tendency to overestimate (due to the lack of metric information) and the tendency to underestimate (due to forgetting or anchoring) will play out in the accuracy scores, but at minimum, we can expect reported frequencies in the Superordinate-Same condition to be larger (even if they are not overestimates) than in the Superordinate-Different condition where there should be only underestimation.

Results and Discussion

Strategy Use. The verbal protocols for each trial were classified into several strategy categories by two coders. A sample of 25% of the protocols were classified by both coders and the correlation was $r = .99$, $p < .01$. The major classes of strategy were Enumeration, General Impression, and Unjustified Response (uninformative protocols). If subjects reported that a frequency was zero that was in fact non-zero, the protocol was not coded. Details of the coding criteria are available from the authors.

Figure 1 shows the proportions of responses based on Enumeration and General Impression strategies as well as Unjustified responses. Clearly, the Same-Different manipulation had the predicted effect on strategy use. Enumeration strategies dominated the two Different conditions, 86% and 94% in the Basic-Different and Superordinate-Different conditions, respectively. Presumably people prefer to enumerate when they can, even if they have other options. In the Basic-Different condition, six per cent of the responses were based on general impression statements, primarily at the two higher frequencies (9 and 12). This is consistent with the idea that Enumeration is most likely when frequencies are low (Burton & Blair, 1987; Blair & Burton, 1991); when frequencies are high, people are more likely to rely on general impressions (Conrad, Brown & Cashman, 1993). One explanation for this is that the more often an event occurs, the more opportunities one has to form an impression of its frequency.

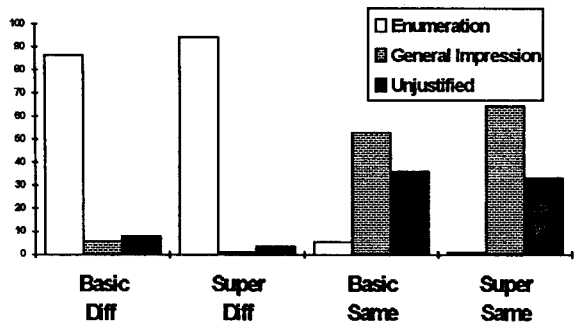

Figure 1. Proportion of responses based on Enumeration and General Impression strategies, and Unjustified Responses. Number of observations are 213, 89, 236 and 93 for Basic-Different, Superordinate-Different, Basic-Same, and Superordinate-Same.

In the two Same conditions, the subjects used a General Impression strategy more than any other approach, on 53% of the trials in the Basic-Same condition and 65% in the Superordinate-Same condition. As we expected, the number of Unjustified responses was large in these conditions as well, 36% and 33% respectively. This suggests that subjects do not have access to discriminable memories for individual study presentations.

Accuracy. In all conditions of the experiment, subjects were extremely sensitive to whether or not products from a test category had appeared in the study phase. Subjects correctly reported zero as the frequency for 97% of the catch trials. This strongly suggests that the subjects were paying attention in both phases and striving to be accurate in the test phase.

In order to compare estimates in the two Basic conditions to those in the two Superordinate conditions we have summed the estimates for the basic level categories within each superordinate. For example, the estimates for *Newspapers, Magazines,* and *Reference Books* are combined and then compared to estimates for *Reading Materials.* To evaluate our predictions, we use a signed, proportional error measure: (estimated frequency - actual frequency) / actual frequency. A score less than zero indicates underestimation; a score greater than zero indicates overestimation. These scores are presented in Table 1 for the four experimental conditions.

Looking first at Column 1 (the Different conditions) it is clear that unique products within the test categories lead to underestimation, as predicted. The amount of underestimation is marginally greater in the Superordinate-Different condition than in the Basic-Different condition, F (1, 378) = 3.53, $p < .10$. We predicted a difference between these means on the grounds that subjects in the Superordinate-Same condition could forget entire basic level categories as well as products from within categories. One piece of evidence that subjects forgot entire categories is the fact that only 21% of the protocols in this condition mentioned all of the basic level categories that were presented. The observation that people are most likley to enumerate a relatively small number of items (Blair & Burton, 1987; Burton & Blair, 1991) seems to apply regardless of whether the items are members of the same of different categories.

	Different	Same
Basic	-.22	+.45
Superordinate	-.40	-.17

Table 1. Mean signed proportional error for the four experimental conditions.

When it is unlikely that people will enumerate (Column 2) they overestimate the frequency of products in basic level categories by 45%, an effect that was also predicted. In contrast, people underestimate the frequency of products in Superordinate categories by 17%. This amount of underestimation falls within the expected range, namely smaller than the estimates in the Basic-Same condition and larger than those in the Superordinate-Different condition. The difference between the means in the two Same conditions is significant, F(1, 378) = 43.86, $p < .01$.

The top row in Table 1 replicates a finding by Brown (1994). That study was carried out in a different laboratory than the current study with different subjects and different experimental stimuli. The fact that this effect generalizes for basic level categories, makes it that much more curious that it should take a different form for superordinate categories (the bottom row). Note that the estimates in the Superordinate-Same condition are relatively accurate even though, by our analysis, they result from competing sources of error: overestimation within categories and underestimation due to forgetting and anchoring.

This pattern of results leads to an interaction of Category Type (Basic versus Superordinate) x Distinctiveness (Same versus Different), F (1,378) = 11.24, $p < .01$. The overestimation observed in the Basic-Same condition is expected because there is no quantitative reference for subjects to map their impressions to a number. The mean signed error in this condition differs from the mean signed error in the other three conditions, F(1, 378) = 88.71, $p < .01$. This difference seems to be responsible for the main effects of Category Type, F(1, 378) = 36.07, $p < .01$, and Distinctiveness, F(1, 378) = 46.23, $p < .01$.

Another accuracy measure produces a different view of the reliability of the strategies used in the four conditions. Absolute error is the absolute difference between estimated and actual frequency on a given trial. It is insensitive to direction and simply accumulates all deviations from the true frequency. Average absolute error rates are presented in Table 2. The units are number of products reported so, for example, in the Basic-Different condition, estimates deviated from actual frequencies by 2.4 products, on average. The corresponding proportional measures (absolute error / average presentation frequency) are presented in parentheses, though our analyses are confined to the absolute measures. The Basic-Different condition is most accurate, leading to deviations of 2.4 products from the actual frequency versus errors of about 4 to 5 reported products for the other conditions. There is a Category Type x Distinctiveness interaction, $F(1, 366) = 11.90$, $p < .01$ which appears to be driven by the greater accuracy of the Basic-Different condition than in the other three conditions, $F(1,366) = 15.77$, $p < .01$. This suggests that when people cannot use numerical information or the test categories are abstract, that the quality of their estimates is likely to suffer.

The inaccurate estimates for the Superordinate-Same condition contrast with the relatively small signed error in that condition (-.17). Apparently subjects are exhibiting both overestimates and underestimates, which when aggregated, reflects what is, at best, an inconsistent strategy. The main effect of Distinctiveness, $F(1,366) = 6.87$, $p < .01$, underscores the point that using qualitative information generally produces inaccurate estimates.

	Different	Same
Basic	2.40 (.30)	5.16 (.65)
Superordinate	4.32 (.54)	3.94 (.49)

Table 2. Mean absolute error for the four experimental conditions.

The picture that is emerging indicates that estimates are rarely perfect. However, it is possible to lack pinpoint accuracy but still correctly order the frequency of a set of items: One could recognize that more fruit juices were presented than magazines without knowing how many. A numerical strategy such as Enumeration should lead to good relative accuracy because numbers are inherently ordered. However, it is unclear if non-numerical strategies will lead to reliably ordered estimates. If they do, then the impressions on which they based must accurately convey ordinal information.

	Different	Same
Basic	.78*	.83*
Superordinate	.75*	.73*

*$p < .01$

Table 3. Rank order correlations for the four experimental conditions.

Rank order correlations of estimated and actual frequencies measure subjects' ability to order their estimates (Brown & Siegler, 1993). Because the statistic is not sensitive to the precision of the subjects' reports, it allows us to partially disentangle these two types of accuracy.

Rank order correlations for each of the four conditions are presented in Table 3. These are computed over all subjects in a condition, though alternative methods of computing the statistics produce comparable results. The correlations are relatively high and all are significant beyond the .01 level. They are also of roughly equal magnitude. This implies that an impression such as "all the time" would have been consistently assigned a larger number than one such as "some of the time."

General Discussion

When people are asked about categories which diverge from the way they naturally structure their world (superordinates), the accuracy of their frequency estimates will suffer (by at least some measures). One implication would be for authors of questionnaires to replace superordinate categories with their basic level components under the assumption that the resulting questions will refer to more natural categories. While such a decomposition might improve the communication between researcher and respondent it will not necessarily improve the accuracy of frequency reports. If respondents are not able or willing to recall specific instances of the category in question, then they are likely to rely on their impressions of frequency. On the basis of our experimental results, using such a strategy on each of several, basic level categories could lead to overestimation in each, which when taken together, would radically inflate the estimate for the superordinate category in which the researcher is ultimately interested.

To improve the accuracy of frequencies based on non-numerical information one might ask respondents for judgments of relative frequency, or at least analyze their responses as ordinal judgments (Smith, Hager, Palphreyman & Jobe, 1992). This may not yield data that is as precise as researchers would like, but it may

be the only kind of frequency report in which researchers should have confidence for both numerical and non-numerical strategies.

At least one message resonates clearly from the current study: A task which is as common place as estimating frequencies is deceptively complex. Experimental studies continue to be a powerful tool for bringing this complexity to light. Ultimately this should help researchers craft better questions and more confidently interpret the estimates provided by respondents.

References

Barsalou, Lawrence. 1992. *Cognitive Psychology: An Overview for Cognitive Scientists.* Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Barsalou, Lawrence and Brian Ross. 1986. "The Roles of Automatic and Strategic Processing in Sensitivity to Superordinate and Property Frequency." *Journal of Experimental Psychology: Learning, Memory and Cognition*, 12: 116 - 134.

Burton, Scott and Edward Blair. 1991. "Task Conditions, Response Formulation Processes, and Response Accuracy for Behavioral Frequency Questions in Surveys. *Public Opinion Quarterly,* 55: 50 -79.

Blair, Edward and Scott Burton. 1987. "Cognitive Processes Used by Survey Respondents to Answer Behavioral Frequency Questions." *Journal of Consumer Research,* 14:280-288.

Brown, Norman. (1994). "Estimation Strategies and the Judgment of Event Frequency." Manuscript submitted for publication.

Brown, Norman and Robert Siegler. 1993. "Metrics and Mappings: A Framework for Understanding Real-World Quantitative Estimation." *Psychological Review*, 100: 511-534.

Conrad, Frederick, Norman Brown & Erin Cashman. 1993. "How the Memorability of Events Affects Frequency Judgments." *American Statistical Association, Proceedings of the Section on Survey Methods Research.*

Ericsson, K. Anders and Herbert A. Simon. 1993. *Protocol Analysis: Verbal Reports as Data, Revised Edition.* Cambridge, MA: MIT Press.

Felcher, E. Marla and Bobby J. Calder. May, 1991. "Answering Survey Questions: The Case of Behavioral Frequency Questions." Paper presented at the 46th Annual Conference of the American Association for Public Opinion Research, Scottsdale, AZ.

Lakoff, George. 1987. *Women, Fire and Dangerous Things: What Categories Reveal about the Mind.* Chicago: University of Chicago Press.

Means, Barbara and Elizabeth Loftus. 1991. "When Personal History Repeats Itself: Decomposing Memories for Recurring Events." *Applied Cognitive Psychology*, 5:297-318.

Menon, Geeta. 1993. "The Effects of Accessibility of Information in Memory on Judgments of Behavioral Frequencies." *Journal of Consumer Research*, 20:431-440.

Rosch, Eleanor, Carolyn Mervis, Wayne Gray, David Johnson & Penny Boyes-Braem. 1976. "Basic Objects in Natural Categories." *Cognitive Psychology*, 8:382-439.

Smith, Albert, Danny Hager, Alison Palphreyman and Jared Jobe. 1992. "Inter-individual Calibration of Frequency Estimates." *American Statistical Association, Proceedings of the Section on Survey Methods Research.*

Tulving, E. and Pearlstone Z. 1966. "Availability versus Accessibility of Information in Memory for Words." *Journal of Verbal Learning and Verbal Behavior*, 5: 381-391.

Tversky, Amos and Kahneman, Daniel. 1972. "Availability: A Heuristic for Judging Frequency and Probability." *Cognitive Psychology*, 4:207-232.