ESTIMATING THE FREQUENCY OF EVENTS FROM UNNATURAL CATEGORIES[1]

Frederick G. Conrad, Bureau of Labor Statistics
Norman R. Brown, University of Alberta
Monica Dashen, Bureau of Labor Statistics
Frederick G. Conrad, Bureau of Labor Statistics Room 4915,
2 Massachusetts Ave., N.E., Washington, DC 20212

Key Words: behavioral frequency, categorization, classification, measurement error.

INTRODUCTION

Consider the following question from the National Health Interview Survey (NHIS: AHB.110):

How often do you do light or moderate activities for at least 10 minutes that cause only light sweating or a slight to moderate increase in breathing or heart rate?

This is a wordy question, but we believe there is something else that makes it difficult to answer. The problem, as we see it, involves the event category ("light or moderate activities …") about whose frequency respondents are asked. This category seems to be at odds with the way most respondents think about the events that the researchers intend to count. When a respondent rides her bike to work she seems more likely to think of it as "bicycling" or "commuting" or "stuff I do most mornings" than as a "light to moderate activity." It seems more natural to mentally group events according to the activities involved (bicycling, commuting) than according to the properties of the events (light to moderate).

There is little relevant experimental evidence about this. The one study that we are aware of demonstrates that people are less accurate when asked to estimate how many instances they have seen from categories organized around properties than from more conventional categories. Barsalou and Ross (1986) found that people were relatively insensitive to actual frequency when estimating the frequency of properties (e.g. sour); their estimates were about the same (between 2 and 3) when actual frequency varied from 0 to 4. They were more sensitive when estimating the frequency of what Barsalou and Ross called superordinates (e.g. toys); the participants' estimates for these categories increased as actual frequency increased. If this finding applies to answering the survey question about light to moderate activities, respondents might be inaccurate, reporting only some of the relevant events and, possibly, misreporting events from other categories as members of the one in question.

Unnatural Categories. Because this type of event category seems to differ from those that respondents use spontaneously, we refer to them as unnatural categories. It's as if they cut across more natural event categories and, so, do not bring to the respondent's mind the kinds of experiences the researchers are interested in. For example, respondents in the NHIS are asked "Do you now have any health problem that requires you to use special equipment, such as a cane, a wheelchair, a special bed, or a special telephone?" (NHIS AHS.070). It is hard to think of other kinds of special equipment that might be eligible for inclusion in answering the question. We propose that this retrieval is difficult because the structures into which people classify their own experiences are, essentially, orthogonal to the structure about which they have been asked. Wheelchairs, beds and telephones involve distinct mental categories for most people, but respondents are asked to consider events about all of them as well as other unnamed equipment that is somehow similar.

What is critical in predicting the likelihood of retrieval for such tasks is the way respondents have encoded or classified events at the time they experience them. Respondents are unlikely to recall the kinds of events the survey designers are interested in if they have encoded those events as members of other – presumably more natural – categories.

This is not to suggest that there exists a canonical, *natural* scheme that all respondents use to classify their experiences; the categories which people spontaneously use might vary in idiosyncratic ways and from one situation to the next. And people may think

[1] We thank Elizabeth Brooks, Pam Douglas, Scott Fricker, Cathy Dippo, Liz Shulman, Michael Schober and Clyde Tucker for help with various aspects of this project. The opinions expressed are those of the authors and do not necessarily reflect the views of the Bureau of Labor Statistics.

of an event or object as being an instance of multiple categories at the same time (e.g. Ross and Murphy, 1999). But whatever an individual's preferred classification scheme, it seems more likely to be organized around actions than properties or other attributes of events.

Researchers often have sound reasons for collecting information about unnatural event categories, though their reasons usually have little to do with the respondents' conception of events. For example, respondents in the Point of Purchase Survey (conducted by the Bureau of Labor Statistics) are asked if, over the last week, they made any purchases or had any expenses for "fats, oils, peanut butter, salad dressings, or dairy substitutes." It would be unlikely for most respondents to spontaneously group these purchases together; the survey authors group these products together because they have similar price change characteristics; the products are legitimately related from an econometric perspective but most respondents do not interpret their experiences from this perspective.

Measurement error. The cost for researchers of directly asking about categories that respondents have not previously used is that this may compromise the quality of the information that is collected. When asked to estimate the frequency of events from such categories, respondents will likely omit events from their totals that they should actually include because the events just do not come to mind. This would lead to net underreporting. As a thought experiment, think of how many products you have purchased in the last three years that contain Velcro. The chances are good that you will find it hard to think of relevant purchases, presumably because we don't usually organize products on the basis of their attributes like Velcro. Yet the chances are also good that you have purchased products with Velcro that you cannot recall.

Conversely, such categories may bring to mind instances that really should be excluded but are counted nonetheless. This could happen because of the poor alignment between the category in the question and respondents' mental categories. If it is not possible to retrieve instances stored as members of the test category, people may search their memories haphazardly, retrieving instances that the survey authors would not want to count. For example, one might reason "perhaps that camera case that I bought contains Velcro?" when it really does not. This would lead to net overreporting.

Data about such categories may be further compromised because answering such questions may be more laborious than many respondents will tolerate. Informally, we have found that people can continue to recall products with Velcro after more than five minutes of trying. This is too hard for most respondents. We have found that if respondents find it difficult to answer behavioral frequency questions, they are likely to truncate the retrieval process and adjust their total to account for unretrieved information (Brown, 1995; 1997; Conrad, Brown and Cashman, 1998). Such adjustment is usually inadequate.

Behavioral frequency questions (During the last month how many times did you …?) are very common in surveys, and they have been widely studied by survey methodologists and psychologists (e.g. Blair and Burton, 1987; Brown, 1995, 1997; Brown & Sinclair, 1999; Conrad, Brown and Cashman, 1998; Menon, 1993). One of the major themes of this literature is that people use multiple strategies to answer these questions and particular strategies affect (1) the size and direction of error, and (2) the amount of effort required to produce an answer. It is usually assumed that the categories in the questions correspond to the categories in respondents' heads. But as we have noted, there are empirical and intuitive reasons to suspect this might not always be the case.

We conducted two experiments to explore people's ability to answer questions about the frequency of events from unnatural categories. In particular, we asked if people are less accurate at estimating the frequency of events from unnatural than from more natural categories. If so, are they biased, that is do they consistently overestimate or underestimate? Do people respond quickly or slowly? How do they come up with their estimates? Are some types of categories inherently unnatural (e.g. properties) or can they sometimes be used naturally?

OVERVIEW OF EXPERIMENTAL METHOD

Both experiments followed the same basic procedure. The experimental sessions consisted of a study phase and a test phase. During the study phase, 109 common words (all of which were nouns) appeared one at a time on a computer screen in front of the participant for six seconds each. The words were members of 16 different categories. Participants were instructed to study each of the words in order to answer some questions about them later. This was designed to simulate experiencing everyday events about which one might be questioned in a survey.

After completing the study phase, participants were asked to estimate the number of instances they had studied from each of the 16 categories. The categories were either natural or unnatural, and were presented individually on a computer screen until the participant entered a numerical response. The "correct" assignment of instances was based on published norms of frequently generated members of categories (Battig and Montague, 1969; McEvoy and Nelson, 1982; Underwood and Richardson, 1956), not on our intuitions. The actual frequency for each of the 16 test categories, that is the number of study items that were members of each, ranged from 0 – 19. The test phase was intended to simulate the task of answering behavioral frequency questions in a survey.

Each participant was exposed to the study items in a different random order, with the constraints that items from the same category appeared in roughly even intervals and two items from the same category were separated by at least one item from a different category. The test items were also presented to each participant in a different random order.

EXPERIMENT 1

We conducted the first experiment to investigate (1) the patterns of response accuracy when people estimate the frequency of instances from natural and unnatural categories, and (2) the range of strategies they use to produce these estimates. Two groups of 8 participants completed the study and test phases of the experiment. Both studied similar types words (e.g. Dog, Chicago, Guitar …) but differed in the kinds of categories on which they were tested. One group was tested on the number of instances from common taxonomic categories (e.g. TREE, FISH, FURNITURE, TOOL …). We chose these to correspond to the categories that most participants would naturally use, at least some of the time. The other group was asked to report the number of words with particular properties (e.g. SMELLY, YELLOW, FUZZY, ROUND). These served as our unnatural categories. All respondents were asked to think aloud as they estimated the number of instances for each category in the test phase.

Results. One measure of overall accuracy is the correlation between each estimate and the actual frequency. By this measure the taxonomic group (*r* =.73) was almost three times as accurate as the property group (*r* = .28). Actual frequency, therefore, was a substantially better predictor of estimated frequency for the taxonomic group than for property group. Yet it was certainly not a perfect predictor for the taxonomic group. Most of the error for both groups was due to underestimation. Average estimated frequency (the grand mean at each level of frequency) is plotted against actual frequency in Figure 1. If the estimates were perfectly accurate they would be plotted on the diagonal axis running from 0 to 19. Instead, almost all of the points in the figure fall below this axis. The slopes of the regression lines are less than 1 for both groups, .44 (r^2= 92) and .22 (r^2=.46) for the taxonomic and property groups respectively. The underestimation bias is more extreme for the property group, F(1,127)=5.68, p<.05.

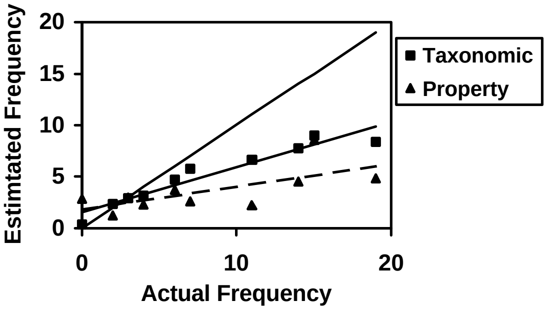

Figure 1. Actual versus estimated frequency

Despite this underestimation bias, participants in the property group actually overestimated the true frequency when it was low, interaction of group x actual frequency, F(9,127)=6.01, p<.001. More specifically, when true frequency was zero, these participants reported having studied 2.9 words with those properties. This sort of overestimation, if it happens in real survey contexts, would be especially troubling when it comes to answering questions like "During the past week did you have purchases or expenditures for fats, oils, peanut butter, salad dressings, or dairy substitutes?" For such questions, this type of overestimation qualitatively changes the response from "no" (the correct answer) to "yes."

What might cause this overestimation? Because the participants thought out loud during the test phase, we often could tell which of the study instances they were considering in their estimates. We classified every instance that participants specifically mentioned as either correct or incorrect, based on the published norms. A correct instance was one that appeared in the norms for the test category or property; an incorrect instance appeared in the norms for one of the other test categories or properties, or, in a few cases did not appear in the norms at all. Almost all of the items specifically enumerated by the taxonomic participants were correct, but a high percentage of the items enumerated by the property participants were incorrect. On average the taxonomic participants mentioned 2.04 correct instances and .05 that were incorrect; in contrast property participants based their estimates on only .93 correct instances but 2.02 that were incorrect. Apparently when properties serve as stimuli (in the laboratory and, presumably, in survey questions) they bring to mind instances that often better exemplify other properties than the one being tested. We call this *off-target enumeration*. This is the first time we have observed this strategy.

On the basis of the think aloud protocols we coded the strategies that participants used to produce their estimates. Participants used one of three strategies for almost all of their estimates: enumeration, adjusted enumeration and general impressions. Broadly defined, enumeration involves summing retrieved instances (e.g. "I remember milk, snow and sugar so I'll say three things were white."). *Enumeration* was coded when the number of items listed in the protocol equaled the number entered into the computer. *Adjusted enumeration* was coded when the number of enumerated items differed from the response entered into the computer. The use of *general impressions* was indicated by qualitative statements of frequency (e.g. "There were a lot of those." or "I saw a few fish.").

The proportions of the codable strategies used by the two groups were quite different. The property group relied overwhelmingly on enumeration (80% of their responses were based on this strategy). The taxonomic group made more balanced use of the three strategies (20%, 53% and 24% of their responses were based on enumeration, adjusted enumeration and general impressions, respectively). One explanation for this pattern is that the participants in both groups did not encode the study instances in terms of their properties and so when those in the property group were tested, they did not have any pre-existing impressions of property frequency on which to base their estimates or adjust their tallies of enumerated instances. In contrast, taxonomic participants were able to supplement their memory of the study instances with impressions formed during the study phase. We propose that in order to form impressions of frequency for a particular category, people need to use that category when they engage in and encode the relevant activities into their memories. This was more likely the case with taxonomic than property categories.

The performance of property participants is particularly troubling because they consistently enumerated – even though other strategies could potentially have improved their estimates and even though more than two of every three items they enumerated were incorrect.

Summary. Experiment 1 indicated that people are quite inaccurate at estimating the frequency of one type of unnatural category, properties. They show a strong underestimation bias – much greater than their counterparts estimating the frequency of instances from taxonomic categories. And they overestimate at the low end of the frequency scale, in particular, when actual frequency is 0. We attribute this low end overestimation to off-target enumeration – counting misclassified instances of exemplars of the test property. Survey data based on such estimates would be of inherently poor quality and potentially misleading.

EXPERIMENT 2

While it is clear that people were inaccurate in estimating the frequency of properties in Experiment 1, it is not clear if this is because properties are an inherently poor way to organize events or because they are just not noticed as often as other event attributes, like the actions involved. Is it possible that if several events share a salient property, people will think of these as a group? If this is the case, then there might be occasions when survey researchers can collect high quality data about property-based categories – if they can determine that people pay attention to the property at the time the event is experienced. For example, it could be that people classify episodes of acute pain (a property) as members of a painful episode category, even if the individual events are otherwise quite different.

We examined this in Experiment 2 by asking a group of 15 participants to study the same items as the property group in Experiment 1, only this time each study word was presented along with the relevant property (e.g. Corn – YELLOW, Ammonia – SMELLY, Chocolate – BROWN, Garbage – SMELLY …). We refer to these participants as the *instance + property* group. The idea was to make properties salient by presenting them specifically. In the test phase, this group estimated the frequency of the properties that appeared in the study phase. If they were to encode each instance presented in the study phase in terms of the property it was presented with, then the properties would serve as natural categories in the sense mentioned earlier. We would expect the instance + property group to perform in the test phase much as the taxonomic group performed in Experiment 1. On the other hand, it is possible that people will not use even very noticeable properties to organize events. In this case we would expect the instance + property group to perform like the property group in the first experiment.

To help evaluate the performance of this group, we also asked another group of 15 participants to study the instances without explicit properties, just as the property group did in the first experiment; we refer to them as the *instance-only* group. They were then tested on the same properties as the instance + property group. We expected the instance-only group to perform just like the property group because their study and test items were identical.

If the two groups respond like their counterparts in the first experiment, we might attribute the accuracy differences in the both experiments to more diligent performance by the more accurate group, perhaps because they searched for instances longer. If so, lower accuracy should be associated with quicker responses than would higher accuracy. On the other hand, it could be that inaccurate estimation of property frequency is due to the inherent difficulty of retrieving exemplars when the instances were not encoded in terms of the test property. In this case, lower accuracy would be associated with slower response times than higher accuracy. To test this we measured response time from the presentation of each test item until the participant entered a response (pressed the enter key). In order to measures estimation time as cleanly as possible, we did not ask the participants to think out loud.

Results. The instance + property group was relatively accurate overall. The correlation between their estimates and actual frequency was *r*=.76, virtually the same as what we observed for the taxonomic group in Experiment 1. The instance-only group was quite inaccurate. The correlation between their estimates and actual frequency was *r*=.16, even less accurate than the property group in Experiment 1. Again, most of the error was due to underestimation for both groups. The slopes of the regression lines were less than 1 for both groups, .61 (r^2 = .93) for the instance + property group and .05 (r^2 = .04) for the property group. As in Experiment 1, the underestimation bias was more extreme for the group that studied instances in isolation (the instance-only group in the current experiment, the property group in the first experiment), F (1,328)= 3.97, p<.05.

Despite the severe underestimation bias displayed by the instance-only group, these participants overestimated at the low end of the actual frequency range to a greater degree. The difference between actual and estimated frequency is displayed in Figure 2. Perfect estimation would lead to a difference of 0. Instance-only respondents substantially overestimated actual frequencies between 0 and 4 while underestimating actual frequencies between 7 and 19. In contrast the instance + property group was quite accurate across the low end of the frequency range, while they showed only moderate underestimation for higher values, interaction of group and actual frequency F(9,252) = 8.02, p<.001.

The poor performance of the instance-only group was not the result of rushing their responses. In fact they took three times as long (12.4 seconds per response) as the instance + property group (4.2 seconds) to produce their estimates (F [1,28]=26.99, p<.001) and these estimates were substantially less accurate. Participants apparently find it quite difficult to retrieve instances with the test properties even though they work hard (or at least long) at the task.

In addition to characterizing overall speed, the response times supplement the Experiment 1 think aloud data in helping to explain how the different groups produced their estimates. A response time function that increases with actual frequency implicates enumeration. The reasoning is that it takes a fixed amount of time to retrieve an instance and add it to the total; the more of these that are retrieved, the longer the response time (Brown, 1995, 1996; Conrad, Brown & Cashman, 1998). This is also the case for adjusted enumeration, though the function is not as steep as for pure enumeration. In contrast, a fast, flat response time function may indicate the use of general impressions because it takes as long to retrieve an impression of “very rarely” as an impression of “all the time.”

We expect the instance-only group to use the same off-target enumeration strategy that the property group did in the first experiment. However, we do not know what the response time function will look like. At the very least there is no reason to expect response times to change systematically with actual frequency.

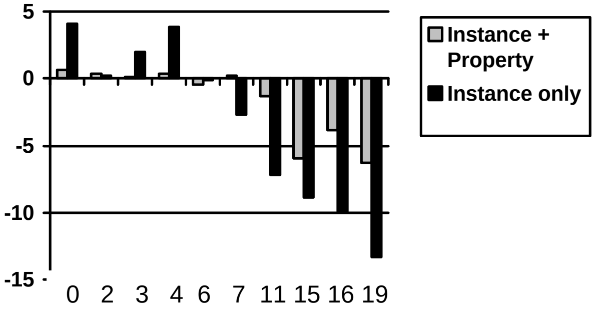

Figure 2. Difference between estimated and actual frequency, Experiment 2.

Median response time is plotted against actual frequency in Figure 3. The functions are quite different (interaction of actual frequency and context F[9,252] = 3.54, p < .001) suggesting that the two groups used fundamentally different strategies. The instance + property participants enumerated in well known ways much of the time, as did the taxonomic group in Experiment 1. The slope is 0.44 (r^2 =.42). The main point is that the pattern is familiar and consistent.

The instance-only group, in contrast, shows no clear evidence of any strategy with which we are familiar. Their average estimates fluctuate widely and the slope of the response time function is slightly negative, -.24 (r^2 =.04). Off-target enumeration appears to be a noisy process, suggesting participants are grasping for information.

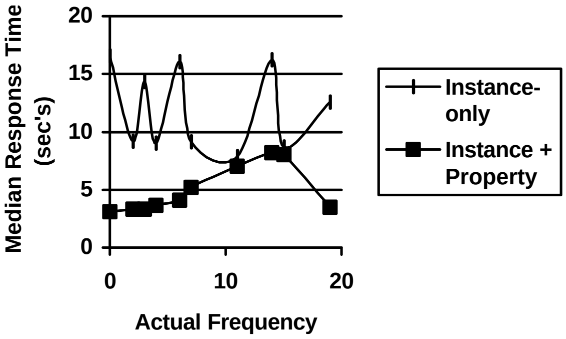

Figure 3. Median response times, Experiment 2

Summary. If several events share a salient property, participants can think of the events as members of the same category. Instance + property participants performed much like the taxonomic group in Experiment 1. They were relatively accurate, though they underestimated overall, and they seemed to base many of their estimates on the same kind of enumeration strategies. Apparently properties can serve as natural categories under some circumstances.

Instance-only participants were both inaccurate and slow, reflecting the inherent difficulty of the task. They overestimated at the low end of the range and underestimated at the high end.

CONCLUSIONS

Experimental participants perform quite poorly when estimating the frequency of events from one kind of unnatural category, properties. It seems likely that survey respondents do as well. What can practitioners do about this type of measurement error? One guideline that question authors can follow is to avoid asking for the frequency of events defined by *adjectives* (e.g. *light* to *moderate*, *special* equipment). Of course, the analytical goals of a survey may still involve estimates of properties. In this case, authors might decompose unnatural categories into their natural parts. Instead of asking about light to moderate activities, they might ask about walking, bicycling, cleaning the house, mowing the lawn, etc. This leads to many questions instead of one, which generally lengthens the interview. But our response time data indicate that the time to answer one question about an unnatural category may be greater than the time to answer several questions about natural component categories, and the results will be more accurate.

REFERENCES

Barsalou, L.W. & Ross, B.H. (1986). The roles of automatic and strategic processing in sensitivity to superordinate and property frequency.. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 11, 211-227.

Battig, W. P. & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology Monograph*, 80 (3, Part 2), 1-46.

Blair, E. & Burton, S. (1987). Cognitive processes used by survey respondents to answer behavioral frequency questions. *Journal of Consumer Research*, 14, 280-288

Brown, N. R. (1997). Context memory and the selection of frequency estimation strategies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23, 898-914.

Brown, N. R. (1995). Estimation strategies and the judgment of event frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 1539-1553.

Brown, N. R. & Sinclair, R. C. (1999). Estimating number of lifetime sexual partners: Men and women do it differently. *Journal of Sex Research*, 36, 292-297.

Conrad, F.G., Brown, N. R. and Cashman, E. R. (1998). Strategies for estimating behavioural frequency in survey interviews. *Memory*, 6, 339-366 .

McEvoy, C. L. & Nelson, D. L. (1982). Category names and instance norms for 106 categories of various sizes. *American Journal of Psychology*, 95, 581-634.

Menon, G. (1993). The effects of accessibility of information in memory on judgments of behavioral frequencies. *Journal of Consumer Research*, 20, 431-440.

Ross, B. H. & Murphy, G. L. (1999). Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology*, 39, 495-553.

Underwood, B. J. & Richardson, J. (1956). Some verbal materials for the study of concept formation. *Psychological Bulletin*, 53, 84-95.