Memory & Cognition
2003, 31 (4), 552-562

Estimating the frequency of events from unnatural categories

FREDERICK G. CONRAD
University of Michigan, Ann Arbor, Michigan

NORMAN R. BROWN
University of Alberta, Edmonton, Alberta, Canada

and

MONICA DASHEN
U.S. Bureau of Labor Statistics, Washington, D.C.

We report two experiments about how people estimate the frequency of event properties when they are explicitly (e.g., *spinach*–GREEN) and implicitly (e.g., *spinach*) presented. In Experiment 1, verbal reports indicated that, for explicitly presented properties, participants used several retrieval- and impression-based strategies and were relatively accurate. Implicitly presented properties led to *off-target retrieval*, which brought to mind more instances of nontarget than of target properties and degraded estimates. A third group estimated the frequency of taxonomic categories (e.g., *furniture*) much as the explicit property group did, suggesting that people can use properties to organize remembered events. In a second experiment, estimation time patterns underscored the results of Experiment 1 and eliminated reactive verbal reports as an explanation. Off-target retrieval was both ineffective and slow.

People are surprisingly skilled at determining how often events occur (e.g., Alba, Chromiak, Hasher, & Attig, 1980; Brown, 1995, 1997; Burton & Blair, 1991; Means & Loftus, 1991). However, their ability to estimate the frequency of events defined by properties (e.g., color, size, smell, or shape) is quite poor. Barsalou and Ross (1986, Experiment 1) found that participants were relatively accurate in reporting the number of instances presented from superordinate categories, such as *birds*, but that their estimates of property (e.g., RED) frequency did not vary as actual frequency varied. Freund and Hasher (1989, Experiment 1) reported a similar finding: Informing participants that they would be tested on the frequency of particular properties led to considerably more accurate estimation than did not informing them.

Why do people perform so poorly when it comes to estimating the frequency of properties? We propose that people's insensitivity to property frequency may result from the kinds of strategies they use as the basis of their estimates. In particular, if the way they understand and classify events does not correspond to the way they are later asked about those events, it becomes difficult to recall the relevant events when questioned about them.

We use the phrase *unnatural category* to describe groupings of events that do not correspond to the way people spontaneously classify those events. The idea is that, when people are asked about the frequency or size of an unnatural category, the mismatch between whatever categories they have spontaneously used to encode events and the one on which they are being tested interferes with the recall of relevant events and information about their frequency. Of course, the categories that people spontaneously use might vary in idiosyncratic ways, and people may think of an event or an object as being an instance of multiple categories at the same time (e.g., Ross & Murphy, 1999). The point is that a mismatch between encoding and test categories degrades recall, which in turn, limits the set of estimation strategies. The present study tested this by documenting the strategies used to estimate property frequency when properties vary in how explicitly they are encoded.

The Multiple Strategy Perspective on Frequency Estimation

People ordinarily evaluate event frequency by using one of many possible strategies. We have called this the *multiple strategy perspective* (Brown, 1995, 1997, 2002; Conrad, Brown, & Cashman, 1998). The basic idea is that, although many strategies are potentially available,

Some of the material contained in this article was presented at the 1999 conference of the American Association for Public Opinion Research. The work was carried out while the first author was employed at the U.S. Bureau of Labor Statistics. The views expressed here reflect the opinions of the authors and not those of the U.S. Bureau of Labor Statistics. We thank Jennifer Berger, Pam Douglas, Cathryn Dippo, Scott Fricker, Liz Shulman, and Clyde Tucker for assistance and support. In addition, we thank Alfred Smith and two anonymous reviewers for helpful comments. The experimental materials are available from the authors. Correspondence concerning this article should be addressed to F. G. Conrad, Institute for Social Research, University of Michigan, 426 Thompson Street, P. O. Box 1248, Ann Arbor, MI 48106-1248 (e-mail: fconrad@isr.umich.edu).

people are restricted in which strategies they can use by the kind of information they remember about the events in question. If people can recall specific events—most likely when the events are distinctive—they should be able to recall and count (enumerate) them. If events occur on a regular schedule, people are relatively likely to have encoded the rate of occurrence and base their estimate on this information. It is also possible that nonnumerical impressions, such as *a lot* or *rarely*, can serve as a proxy for a more quantitative estimate if the impression can be converted into a number. If none of these is available (specific events, rate information, or nonnumerical impressions), people can still derive a frequency estimate from information about the recall process (as opposed to what is actually recalled). We call this *memory assessment* and propose that it should be available by default even if very little of substance can be remembered. People using this type of strategy estimate frequency on the basis of how familiar the event seems (Whittlesea, 1993), how similar it is to encoded instances (Hintzman, 1988), or how available (easily brought to mind) relevant instances are (e.g., Tversky & Kahneman, 1973).

We (Brown, 1995, 1997; Conrad et al., 1998) have observed people using all of these strategies when the test categories were familiar (e.g., *cities*, *musical instruments*, and *trips to the grocery store*) and probably matched people's encoding of events (this match was ensured for laboratory participants by presenting events that consisted of an instance and its category, the same category on which participants were later tested). But on what basis do people respond when they have not encoded the events as instances of the test categories, presumably because the instances fit more naturally with other categories? For example, is it possible to form a qualitative impression of frequency, such as *that happened a lot*, for a category that one is not thinking of when potential instances occur? The estimation of property frequency provides an excellent test bed for this question, because people do not seem to encode events in terms of properties unless the properties are made particularly salient (Barsalou & Ross, 1986, Experiment 3; Freund & Hasher, 1989, Experiment 2).

Estimation Error

When asked to estimate the frequency of properties, it seems likely that people will 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*. Conversely, when asked about such categories, people may search their memories haphazardly, retrieving instances that really do not qualify. This would lead to net *overreporting*. Presumably, overreporting is more prevalent for low-frequency events: A modest number of intrusions could more than compensate for forgotten legitimate events because, by definition, there are not many of the latter. And presumably, underreporting is more prevalent for high-frequency events: Forgetting could be large, since there are more legitimate events to forget and this could swamp intrusions. This pattern of error—low-end overestimation and high-end underestimation—could produce the flat estimation functions observed by Barsalou and Ross (1986) and Freund and Hasher (1989).

Despite the emphasis on accuracy in establishing the difficulty of estimating property frequency, accuracy raises problems as a measure for this task. The main problem is that instances are often characterized by more than one property—for example, a skillet may legitimately be considered to be both metal and round. Thus, individuals might disagree about what property is most associated with a particular instance. Without more straightforward associations between instances and properties, it is hard to establish the property's true frequency and, thus, determine estimation accuracy. This is much less of an issue with instances of conventional taxonomic categories, which are primarily associated with one category—for example, a chair is primarily a piece of furniture. Although some instances of such categories may be equally good exemplars of other categories (e.g., Ross & Murphy, 1999), this is not usually the case. In the present article, we focus on *how* people estimate property frequency—that is, the type of strategy they use—rather than on *how accurate* they are. Thus, we primarily discuss measures of strategy use—in particular, verbal protocols and reaction times (RTs).

We report two experiments in this paper in which we compare the processes people use to estimate frequency of categories that match and do not match the way they have represented the relevant events. Our intuition at the outset was that strategy differences could help explain the differences in performance for estimates of property frequency that others have observed. In the first experiment, verbal protocols are used to directly assess strategy use when the estimation task involves the frequency of implicitly encoded properties, explicitly encoded properties, and conventional taxonomic categories. The second experiment uses RTs to corroborate and extend the verbal protocol findings.

EXPERIMENT 1

We conducted the first experiment to investigate the range of strategies used to estimate event frequency when people vary in how well they have encoded the relevant aspects of the events. In particular, we compared the strategies used to estimate the frequency of properties that are only implicitly indicated when the events are encoded with the strategies used when properties are explicitly indicated. Two groups of 8 participants completed a study and a test phase of the experiment. Both the *implicit* and the *explicit property* groups studied a set of ordinary nouns (e.g., *tomato*, *fur*, *needle* . . .), but the explicit property group studied the nouns in the context of the properties on which they would later be tested (e.g., *tomato*–RED, *fur*–SOFT, *needle*–SHARP . . .). Because the test property was a salient part of each event for the ex-

plicit property group, it seemed plausible that these participants would encode it during the study phase and would group together events that shared a property.

There are both intuitive and empirical reasons to believe that the participants in the explicit property group would organize events by their properties. Consider everyday events that share a highly salient property, such as acute pain. It seems likely that people spontaneously form a coherent category for such events and should be able to answer questions about their frequency fairly accurately. In contrast, it seems less likely people will group together events that share a less salient property, such as minor pain; as a result, they should be less able to estimate the frequency of such events. In addition, blocking experimental items by their properties (Barsalou & Ross, 1986, Experiment 3) led to performance that closely resembled that for common taxonomic categories (*superordinates*, in that study). If the participants in our explicit property condition do, in fact, classify events by the properties on which they are later tested, they should be able to use strategies similar to those used to estimate the frequency of ordinary taxonomic categories, where category membership seems to be encoded as part of the event and is a powerful organizational criterion (Brown, 1995, 1997, 2002; Conrad et al., 1998). In order to test this, we asked a third group of participants, the *taxonomic* group, to study a similar set of nouns (e.g., *table*, *Boston*, *emerald*) and then tested them on the frequency of taxonomic categories to which the study items belong (e.g., FURNITURE, CITY, JEWEL . . .).

Method

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 6 sec each. For the explicit property group, a property appeared above each study word. The participants were told to "study the word carefully. This is because your memory for the words will be tested during the second phase of this experiment."

Twenty-four participants were recruited from an advertisement in the Washington Post and were paid $25 for the session. Their demographic characteristic s (age, education, race, and sex) were roughly balanced across the three groups.[1]

After completing the study phase, the participants were asked to estimate, as accurately as possible, how many instances of 19 properties or taxonomic categories they had just studied. The test items (properties or taxonomic categories) were presented individually on a computer screen, and the participants were not given any information about the upper bound of plausible responses. The participants were asked to think aloud while determining the frequency of the test item. When they had determined the frequency, they typed it into a response field to the right of the test item, using the numerical keypad on the computer keyboard, and pressed the Enter key when it was complete. The test item was then removed from the screen.

Materials. The words were chosen so that each was primarily associated with a single property or a single taxonomic category. As was indicated above, this is less straightforward in the case of properties, where many words may be good examples of more than one property, than in the case of taxonomic categories. The associations between instances and properties were based on a set of norms published by Underwood and Richardson (1956) for 213 ordinary nouns. They asked participants to provide a "sense impression" of each noun, where sense impression was defined as the response "one might use to describe an object upon seeing it for the first time" (p. 86). Underwood and Richardson picked words that they believed would elicit a single sense impression and excluded from the final list those words that elicited "a wide variety of responses." However, most words elicited more than one sense impression across all participants. We chose 109 words from their final list as our study items and picked the most frequently reported sense impression for each as its test property. This resulted in 16 test properties.

Some of these properties were also associated with other nouns presented in the study phase. For example, in the Underwood and Richardson (1956) norms, GREEN was most often elicited by words such as *spinach* and *ivy* but was also associated (second most often) with *lizard*; in fact, the sense impression most often associated with *lizard* was SLIMY, which was the test property we used. By selecting the most frequently associated sense impressions as the test properties in the present study, we intended to choose properties that, if people encode any properties when they study the nouns, it would be these. However, all we can say with certainty is that these associations between nouns and properties were the strongest associations in the norms.

The assignment of instances to taxonomic categories was based on a set of study and test items used by Brown (1995) in an experimental paradigm that closely resembled our present approach. Brown classified study items as members of the corresponding test categories on the basis of two sets of norms (Battig & Montague, 1969; McEvoy & Nelson, 1982). The generally high levels of accuracy in his study suggested that participants usually agreed with his classifications.

In order to create variation in the levels of *actual* frequency, we selected different numbers of study items for different test properties, ranging from 0 to 19. Many of the presentation frequencies in the present study were substantially higher than those used in the studies by Barsalou and Ross (1986) and Freund and Hasher (1989); presentation frequency ranged from 0 to 4 in the former study and from 0 to 8 in the latter study. The reason for the high presentation frequencies in the present study was to create the conditions for participants to use nonnumerical estimation strategies, which have been observed primarily when frequencies are high (e.g., Brown, 1995, 1997; Conrad et al., 1998). One constraint introduced by testing high-frequency items is that it limits our ability to vary the actual frequency of test properties and categories across participants, because the norms simply do not include large numbers of instances for all test items. In fact the particular frequency levels that we used were determined by the number of instances in the norms for particular properties (e.g., there were exactly 19 instances of WHITE as the primary associate; next most frequent, there were 15 words for which SMELLY was primarily associated, and so on). The frequency of the taxonomic categories was matched to the frequency of the properties.

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. The test sequence included 19 items; 16 category terms, from which at least two members were presented in the study phase, and three catch trials—properties or taxonomic categories from which no members were studied and whose frequency was, therefore, zero. There were 10 levels of actual frequency for the 19 test items.

Results and Discussion

We first will examine the range of estimation strategies used by the participants in the three groups and then

will examine how well they performed using different strategies. We will argue that differences in accuracy between the groups resulted from differences in the strategies used.

Verbal protocols. On the basis of the think-aloud protocols, we coded the strategies that the participants used to produce their estimates. Two coders, who were unaware of the hypotheses or experimental conditions to which the participants were assigned, worked together and then resolved discrepancies through discussion. They were able to code a strategy 78% of the time for the participants in the implicit property group, 85% of the time for those in the explicit property group, and 95% of the time for those in the taxonomic group. For these codable responses, three strategies accounted for the majority of the estimates: two types of enumeration and one nonnumerical strategy. Broadly defined, enumeration involves summing retrieved instances—for example, "I remember milk, snow, and sugar, so I'll say three things were white." When the number of items listed in such protocols equaled the number entered into the computer, the strategy was coded as *simple enumeration.* When the number of enumerated items differed from the number entered into the computer, this kind of report was coded as *adjusted enumeration.* In the following example, three items are listed, but the adjustment leads the participant to enter six: "Maple, elm, oak . . . and there were maybe three more trees."

The strategy was coded as *general impressions* when the protocol included qualitative statements of frequency (e.g., "There were a lot of those" or "I saw a few fish"). When estimates were not accompanied by a verbal protocol, they were coded as *unjustified*. Although we cannot definitively attribute unjustified estimates to any one strategy, memory assessment strategies are unlikely to be reportable, because the processes involved bypassing working memory, a requirement for producing verbal reports (Conrad et al., 1998; Ericsson & Simon, 1993). Some unjustified reports may, therefore, involve memory assessment.

The average percentages of the strategies used by each participant in the three groups are displayed in Figure 1. The patterns of use differed considerably between the groups [group × strategy^2 interaction, $F(6,63) = 6.34$, $MS_e = 0.039$, $p < .001$], although the patterns were generally similar for the explicit property and the taxonomic category groups and different for the implicit property group. The participants in the implicit property group based 59% of their estimates on simple enumeration, although far fewer used this strategy in the explicit property group (13%) and the taxonomic group (18%). This pattern is reversed for the adjusted enumeration and general impressions strategies. The participants in the implicit property group based many fewer estimates on adjusted enumeration (13%) and general impressions (4%) than did the explicit property group (34% and 33%, respectively) or the taxonomic group (37% and 23%, respectively). The implicit and explicit property groups provided relatively few unjustified responses (2% and 5%, respectively), in contrast to the taxonomic group, for which many more responses were unjustified (17%).

The similarity of strategy use for the explicit property and the taxonomic category groups suggests that properties can be explicitly encoded as part of events, and

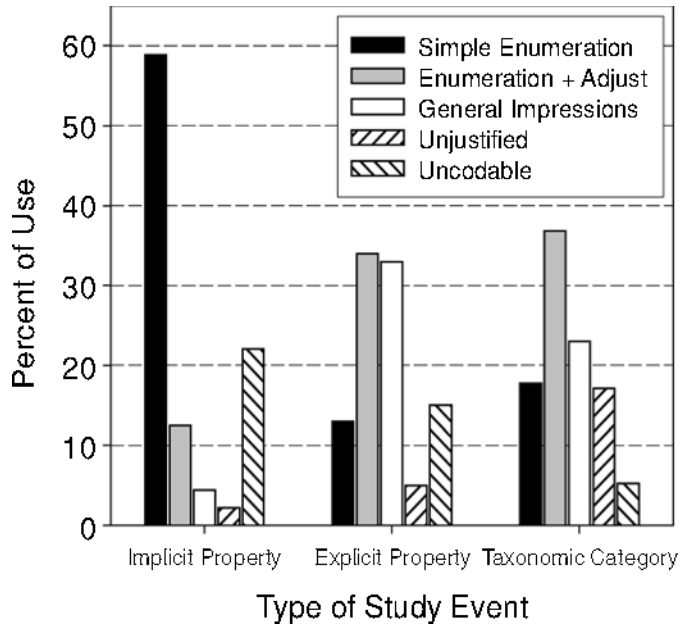

Figure 1. Percentages of estimates based on particular strategies for the implicit property group, the explicit property group, and the taxonomic group.

when they are, their frequency is estimated much the way that the frequency of instances from taxonomic categories is estimated.[3] Moreover, these patterns of strategy use are consistent with what has been observed for estimating the frequency of autobiographical events, such as grocery shopping (Conrad et al., 1998), indicating that the particular mix of strategies used by these groups in the present experiment is robust and general. In contrast, the different distribution of strategies used by the implicit property group—especially their reliance on simple enumeration—suggests that, whatever people remember about the properties of these events, it does not support the range of conventional estimation strategies.

One explanation for this pattern of results is that the participants in the explicit property and taxonomic groups were able to form qualitative impressions but that those in the property group were less able to do so. Such impressions can serve as the primary basis of the response (the general impressions strategy) and as the basis of an increment in the adjusted enumeration strategy. Presumably, impressions of quantity accrue for the categories and properties of events that people spontaneously encode as they are exposed to events. When people do not spontaneously encode properties—as seems to be the case for the implicit property group, as well as for Barsalou and Ross's (1986) property group and Freund and Hasher's (1989) uninformed group—they are unlikely to have impressions of property frequency available when tested. In a similar vein, when people do not encode property information, they should be less likely to use memory assessment strategies: If recall is poor for events with particular properties, there is little by which to judge familiarity, similarity, or ease of retrieval. This may be reflected by the difference in the number of unjustified responses between the implicit property and the taxonomic groups. If this were the case, it would indicate that the implicit property group was unable to take advantage of information (familiarity, similarity, or ease of retrieval) that is typically available.

Estimation performance. Different strategies have been shown to lead to different levels and directions of error (Burton & Blair, 1991; Brown, 1995, 1997, 2002; Menon, 1993). Thus, the different proportions of strategy use across the groups could produce associated differences in overall estimation performance. In particular, enumeration strategies are associated with underestimation at all levels of frequency, and nonnumerical strategies, such as those involving general impressions, are associated with overestimation, especially at higher levels of frequency. Because the implicit property group relied predominantly on simple enumeration, we would expect them to underestimate, relative to normative values (Underwood & Richardson, 1956), across all levels of frequency. However, Barsalou and Ross (1986) and Freund and Hasher (1989) observed pronounced regression effects—low-end overestimation and high-end underestimation—among participants whose task resembled that of our implicit property group.[4]

The average size of the estimates for the 10 levels of frequency, relative to actual frequency, did in fact differ between the three groups [interaction of group × frequency, $F(18,189) = 3.66$, $MS_e = 7.671$, $p < .001$]. Figure 2 displays mean estimated frequency plotted against actual presentation frequency. For the implicit property and taxonomic category groups, instances were studied in isolation, so we need to infer actual presentation frequency for the properties and categories. To do this, we treat the property or category normatively associated with each instance as having been presented each time an associated instance was presented. The main thing to notice in the figure is that all three groups tended to underestimate actual frequency but the implicit property group did so to a greater extent than the other two. At the same time, the implicit property group appears to have overestimated actual frequency when it was low.

To explore this in more detail, we measured estimation performance in three ways for each participant. The three measures were (1) absolute difference (the absolute value of the difference between the actual and the estimated frequency), (2) signed difference (the difference between the actual and the estimated frequency), and (3) the rank order correlation between the actual and the estimated frequency. Absolute difference aggregates all deviations from the actual frequency. Signed error indicates whether responses tend to be larger or smaller than the actual frequency. Rank order correlations indicate the extent to which the participants were sensitive to the relative frequency—that is, estimates increase as normative frequency increases—although the estimates do not need to be close in size to actual frequencies. These three performance measures are presented in Table 1.

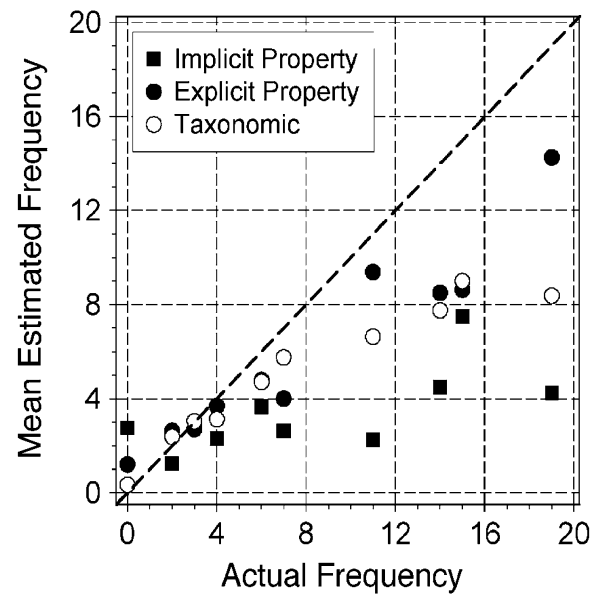

Figure 2. Mean estimated frequency at each level of actual frequency for the implicit property group, the explicit property group, and the taxonomic group.

By all of the performance measures, the implicit property group scored poorly, as compared with the explicit property and the taxonomic groups. Mean absolute error was higher for the implicit property group (5.65) than for the explicit property (4.21) and the taxonomic (3.59) groups [$F(2,21) = 6.49, MS_e = 16.739, p = .006$; contrast of implicit to explicit property groups, $p = .031$; contrast of explicit property to taxonomic groups, n.s.]. By the signed error measure, all the groups underestimated, but the degree of underestimation was more extreme for the implicit property group (−4.739) than for the explicit property (−2.12) and the taxonomic (−2.95) groups [$F(1,21) = 2.96, MS_e = 23.055, p = .074$; contrast of implicit and explicit property groups, $p = .001$; contrast of explicit property to taxonomic groups, n.s.]. The greater underestimation for the implicit property group than for the explicit property and taxonomic groups can be seen in Figure 2. Finally, the implicit property group was less sensitive to relative size of actual frequency ($r = .31$) than was the explicit property ($r = .83$) and the taxonomic ($r = .75$) groups [$F(1,21) = 11.67, MS_e = 0.053, p < .001$; contrast of implicit and explicit property groups, $p < .001$; contrast of explicit property to taxonomic category groups, n.s.].

The relative insensitivity of implicit property participants to actual frequency led to low-end overestimation, despite their overall underestimation bias. In particular, when the true frequency was zero, these participants reported having studied, on average, 2.75 items with the test properties; in contrast, the explicit property and the taxonomic groups reported studying only 1.21 and 0.34 instances of the same properties, respectively. In conjunction with high-end underestimation, the implicit property group's low-end overestimation essentially reproduces the flat frequency function observed by Barsalou and Ross (1986) and Freund and Hasher (1989). This contrasts with moderate, across the board underestimation by the explicit property and taxonomic groups and is reflected by the lower rank order correlation between the actual and the estimated frequency for the implicit property group than for the other two groups.

The relatively flat estimation function for the implicit property group is at odds with what is typically observed with enumeration strategies (estimated frequency typically increases with actual frequency; e.g., Brown, 1995, 1997), yet this group used some type of enumeration (either simple or adjusted) for 72% of their estimates. To explore this further, we classified the instances that the participants specifically mentioned in their verbal reports as *hits* or *intrusions*. A hit was an instance that was primarily assigned to the test category or property by the participants in the normative studies; an intrusion was assigned primarily to one of the other test categories or properties or, in a few cases, did not appear in the norms at all. The ratio of hits to hits plus intrusions indicates the degree to which the participants retrieved and counted events that were primarily associated with the target property or category. A ratio close to one indicates that most of what was counted was primarily associated with the target property or category, and a ratio closer to zero indicates that most of what was recalled and counted came primarily from other properties or categories. The average ratio for the implicit property group (.38) was lower than those for the explicit property (.83) and the taxonomic category (.96) groups [$F(2,21) = 57.28, MS_e = 0.013, p < .001$].

Apparently, when test properties are not encoded as part of an event, they bring to mind instances other than those that are primarily associated with the property. In our previous studies involving more conventional categories (e.g., Brown, 1995, 1997; Conrad et al., 1998), participants almost never enumerated instances from categories other than the target category. We refer to the strategy used by the implicit property participants in the present study as *off-target* enumeration, to capture the inaccuracy of the underlying retrieval.

It is possible that the low hit rate for the implicit property group actually involved recall of some study items that the participants encoded as instances of the test property but that, on the basis of the Underwood and Richardson (1956) norms, we did not classify this way. Nonetheless, the lack of impression-based strategies in this group suggests that their performance is not just the result of disagreement about the pairing of instances and properties. The RT data in Experiment 2 corroborate our sense that what we have dubbed *off-target* in the present experiment does not involve ordinary retrieval processes.

In summary, Experiment 1 indicated that the processes are qualitatively different for estimating the frequency of events when the test attribute is well encoded (explicit property or taxonomic category) than when it is not (implicit property). In addition to simple enumeration, the participants in the explicit property and taxonomic groups used strategies that involved mapping nonnumerical information onto numbers (adjusted enumeration, general impressions, and possibly, memory as-

Table 1
Mean Estimation Performance in Experiment 1

	Absolute Error		Signed Error		*r* (Actual and Estimated)		Proportion of Hits to All Enumerated Items	
Group	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Implicit property	5.65	0.518	−4.73	0.617	.31	.104	.38	.054
Explicit property	4.21	0.410	−2.12	0.575	.83	.026	.83	.038
Taxonomic	3.59	0.421	−2.95	0.477	.75	.092	.96	.024

sessment). Overall, these strategies produced relatively high levels of performance for these groups. The implicit property group, in contrast, used primarily a simple enumeration strategy, which by our definition, relies entirely on counting and so does not involve converting nonnumerical information into numbers. Off-target enumeration seems to bring to mind items that are not the best exemplars of the test property and may better illustrate other properties. This off-target enumeration led to relatively inaccurate estimates by several measures.

EXPERIMENT 2

The conclusions from Experiment 1 concerning strategy use are based on respondents' verbal reports, produced by thinking aloud as they estimated frequencies. However, the process of thinking aloud has been shown to interfere with and, in rare cases, facilitate the task about which people are reporting (e.g., Russo, Johnson, & Stephens, 1989). To test the possibility that the process of thinking aloud in Experiment 1 somehow led to the off-target retrieval underlying performance by the property group—a so-called reactive effect—we assessed strategy use through RT measures in the present experiment. If the implicit and explicit property groups in the present experiment differed in their strategy use, as they did in Experiment 1, we can rule out the think-aloud task as the explanation.

A secondary issue that RT measures allowed us to investigate was whether implicit property participants can retrieve more *on-target* items if they persevere. With autobiographical information, retrieval of additional facts has been observed after nine sessions of 1 h each (Williams & Hollan, 1981); perhaps more time would also lead to continued recall of events with particular properties. Alternatively, the mismatch between encoding and testing may be insurmountable despite any amount of effort. We addressed this in Experiment 2 by comparing the estimation times for implicit and explicit property groups. If off-target responding were to be both inaccurate (i.e., departs from the normative frequencies) and slow, this would suggest that these participants were at a serious disadvantage and that more time alone would not have overcome the retrieval difficulty.

Three groups of participants estimated property frequency. The implicit and explicit property groups followed exactly the same procedure as their counterparts in Experiment 1. Assuming that strategy use in Experiment 1 was not affected by thinking aloud, we expected these groups to provide evidence of off-target and on-target enumeration, respectively, in the present experiment. To provide a further contrast, a third group was tested under conditions likely to promote the use of nonnumerical frequency information, such as general impressions. The participants in the third group studied just property names (e.g., RED, SOFT, SHARP . . .) and were then tested on the frequency with which these properties had been presented. Because all study events with a particular property were identical—that is, they consisted of just the name of the property—it would be difficult for these participants to differentiate remembered events in order to recall and count them. That would leave them with little option but to use qualitative frequency information, such as "that appeared pretty often." We refer to this third group as the *property-only* group.

We can assess strategy use with RT because characteristic temporal patterns are produced when people enumerate (on-target) items and use general impressions. The majority of the responses provided by the explicit property group in Experiment 1 were based on either simple or adjusted enumeration, so we expected RTs in the present experiment to increase with actual frequency. This is because the number of operations increases as more events are retrieved (and counted) and each such operation takes time (Brown, 1995, 1997; Conrad et al., 1998). In contrast, we expected the property-only group to produce a fast, flat RT function, because converting a qualitative sense of frequency into a quantity involves the same number of retrieval operations—probably, just one—irrespective of the size of the impression (Brown, 1995, 1997; Conrad et al., 1998).

If RT increases with the size of the estimate for the implicit property group, it would indicate that off-target retrieval involves more operations as frequency estimates increase, even though the retrieval is inherently inaccurate. A flat function seemed unlikely for this group, because of their inability to form qualitative impressions in Experiment 1 and, thus, use nonnumerical strategies, although it is conceivable that they could retrieve and count the same number of events at all levels of presentation frequency, which would produce a flat RT function. It seemed more likely that, without impressions of frequency, they would not know when to persevere and when to cut their losses in trying to retrieve instances. This could produce RTs that vary across frequency level but are unrelated to their estimates.

Method

Experiment 2 followed the same procedure as that in Experiment 1, with a few exceptions. In the study phase, implicit property and explicit property group participants were presented with the same items and were tested on the same properties as were their counterparts in Experiment 1. Property-only participants were presented just these properties in the study phase. In the test phase, the participants' estimation time was recorded from the presentation of the test property until they pressed the space bar, indicating they had finished estimating the frequency. They then typed the estimated frequency into the computer. Estimation times were also measured from the presentation of the test property until the participant had completed typing the response. Because these two measures of RT were highly correlated ($r = .997$, $p < .001$), only the former are reported.

Ninety participants were recruited from an advertisement in the Washington Post and were assigned at random to one of three conditions. Their demographic characteristics were roughly balanced across the three groups. Each participant was paid $25 for the session.

Results and Discussion

We turn first to the RT results and then to estimation performance, in order to corroborate our interpretation of the RTs. The major RT results appear in the first four

Table 2
Response Time (RT, in Milliseconds) and Estimation Performance in Experiment 2

	RT		Slope of RT Function		Absolute Error		Signed Error		*r* (Actual and and Estimated)	
Group	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Implicit property	15,185	1,747	54	76.10	6.59	0.323	−3.58	0.449	.22	.029
Explicit property	6,478	1,058	313	103.57	4.05	0.226	−2.25	0.296	.78	.029
Property only	5,245	1,411	−23	41.26	6.56	0.809	1.86	0.886	.75	.034

columns of Table 2, and the major estimation performance results appear in the remaining six columns of the table.

Response time. The implicit property group was substantially slower, overall, than the explicit property and property-only groups [average RT = 15,185, 6,478, and 5,245 msec, respectively; $F(2,87) = 11.97$, MS_e = 650,893,960, $p < .001$; contrast of implicit property to other two groups, $p < .001$]. Clearly, the implicit property participants invested considerable effort in the task—over 15 sec, on average, per trial. In addition to the large variation in overall RT, the patterns for the different groups have fundamentally different forms, as is displayed in Figure 3. These differences are statistically reliable [interaction of group and presentation frequency for average RT, $F(18,783) = 4.02$, $MS_e = 35,384,138.879$, $p < .001$].

The pattern for the implicit property group provides further evidence that these participants struggled with the estimation task. The RT function has a relatively flat slope (average slope = 54 msec per item), but the times vary widely, seemingly at random, for different levels of actual frequency.[5] This could easily reflect the off-target retrieval observed in Experiment 1: Participants are likely to have little sense of whether or not they have recalled all instances of a test property (a type of nonnumerical impression), and so the temporal cutoff, after which they will no longer try to retrieve additional instances, not only will be long but also will vary irrespective of actual frequency.

In contrast, the RT function for the explicit property group increased as actual frequency increased (average slope = 313 msec per actual item), most likely signaling the use of (on-target) enumeration. The trend appears to reverse at the highest levels of actual frequency, since the RTs for a frequency of 19 (average RT = 7,886 msec) are faster than those for a frequency of 15 (average RT = 10,465 msec). However, this difference is not reliable [paired $t(29) = 1.122$, $p = .271$].

Finally, the estimates for the property-only group were relatively fast and unrelated to the size of actual frequency (average slope = −23 msec per item), consistent with the pattern we have previously observed for the use of nonnumerical strategies (Brown, 1995, 1997; Conrad et al., 1998). The property-only participants were apparently doing about the same amount of work across all levels of actual frequency. This is consistent with the idea that they were retrieving or deriving a single impression of frequency and converting it to a numerical estimate.

Estimation performance. Although these RT patterns are sensible in light of what was observed in Experiment 1, our confidence that they represent the use of the strategies we believe they represent is strengthened by also examining estimation performance. As in Experiment 1, we computed three measures of estimation performance for each participant: absolute error, signed error, and the rank order correlation between actual and estimated frequency (see Table 2).

When measured by absolute error, the explicit property group's performance was superior (4.05) to that of the property-only group (6.56) or the implicit property group (6.59). The overall effect of group was not significant [$F(2,87) = 1.41$, $MS_e = 452.815$, $p = .25$], but the contrast between the explicit property group and the other two groups was significant ($p = .027$). The use of conventional enumeration strategies by explicit property participants should certainly lead to more accurate responses than the off-target enumeration that was, presumably, used by the implicit property group. The high absolute error exhibited by the property-only participants (as compared with the explicit property participants) is

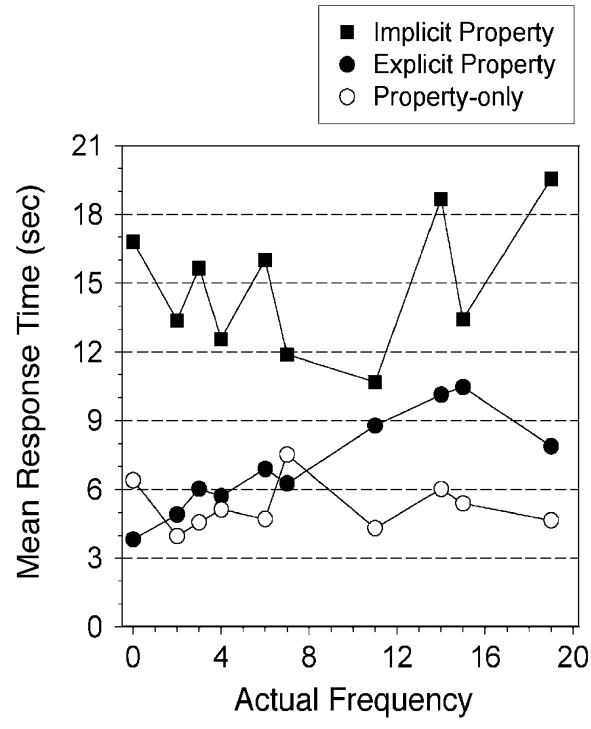

Figure 3. Mean response time at each level of actual frequency for the implicit property group, the explicit property group, and the property-only group.

consistent with the use of nonnumerical strategies to the extent that relative accuracy (rank order correlation) is also high. This is because qualitative descriptions of frequencies, such as *rarely* and *a lot*, capture relative, but not absolute, magnitude.

Turning to the rank order correlations, the explicit property group (.78) and the property-only group (.75) exhibited far greater sensitivity to the relative size of actual frequency than did the implicit property group (.22) [$F(2,87) = 102.32$, $MS_e = 0.029$, $p < .001$; contrast between implicit property and the other groups, $p < .001$]. By counting recalled episodes and, thus, being aware of the actual numbers, participants who use simple or adjusted enumeration (explicit property group) should be sensitive to absolute levels of frequency, from which sensitivity to relative frequency follows. By using nonnumerical strategies (property-only group), participants should be sensitive to relative frequency, because of the characteristics of qualitative descriptions such as *rarely* and *a lot*, just mentioned. In contrast, off-target enumeration (implicit property group) should lead to poor relative accuracy because, presumably, this occurs when people have neither explicit numerical nor qualitative, impression-based information available.

This interpretation is further corroborated by larger underestimation (negative signed error) for the implicit property group (−3.58) than for the explicit property group (−2.25), as in Experiment 1, and by overestimation for the property-only group (1.86) [$F(2,87) = 3.92$, $MS_e = 615.565$, $p < .05$]. The overestimation displayed by the property-only group is consistent with the use of nonnumerical strategies, in that the conversion of an impression to a number is bounded by zero at the low end and is unbounded at the high end, leading to net overestimation (Brown, 1995, 1997; Conrad et al., 1998). The patterns of under- and overestimation can be seen in Figure 4.

It is evident in the figure that estimation functions for the three groups have different shapes: The function for the implicit property group is virtually flat, despite a small peak at a frequency of 4, whereas the estimates increase with actual frequency for the explicit property group and the property-only group. These differences are reflected in the interaction of group and actual frequency for the average estimated frequency [$F(18,783) = 5.84$, $MS_e = 36.954$, $p < .001$].

Because the implicit and the explicit property groups produced patterns for all three measures that closely resembled the patterns observed in Experiment 1, we can be reasonably sure that we have correctly interpreted strategy use in the present experiment without the use of verbal protocols. Thus, it seems unlikely that the process of thinking aloud somehow created the earlier results.

Thus when off-target enumeration is most likely (implicit property group), people are slow, and their estimates depart from actual frequency, underscoring the inherent difficulty of their task. Their severe fluctuations in RT across the levels of actual frequency indicate that they lack the kind of nonnumerical information that could guide them in how much retrieval to attempt. This is further indicated by the difference in RT pattern between these participants and those in the property-only group, who seemed to base their estimates on nonnumerical information. The latter participants, who explicitly encoded property information but little else, were fast and stable across levels of actual frequency. The explicit property group estimated quickly overall but required more time as the actual frequency and, presumably, the amount of enumeration increased.

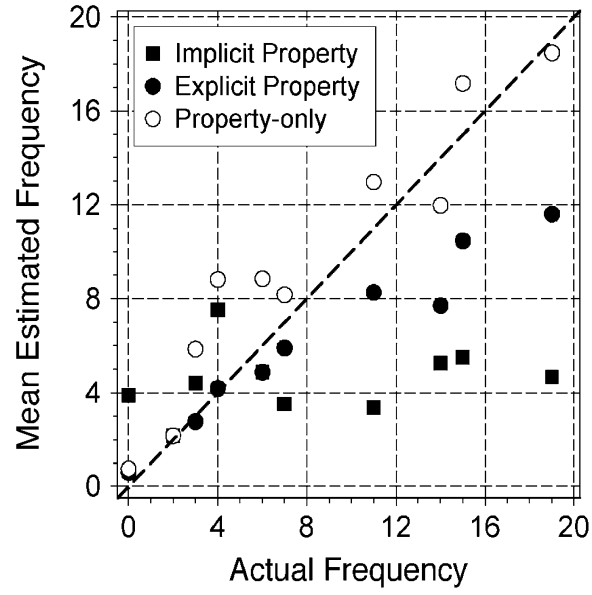

Figure 4. Mean estimated frequency at each level of actual frequency for the implicit property group, the explicit property group, and the property-only group.

GENERAL DISCUSSION

Experimental participants perform quite poorly when estimating the frequency of events from one kind of unnatural category—namely, properties. Participants appear to be denied access to the rich array of strategies typically available when people estimate frequency—that is, for natural categories. Instead, they struggle to retrieve instances of the test property and are, ultimately, more likely to retrieve instances that lack the property than instances that possess it. They do not build up qualitative impressions of frequency and seem unable to fall back on memory assessment. We have not previously observed this off-target retrieval and so must extend the multiple strategy perspective on frequency estimation to include it (see Brown, 2002). We propose that off-target retrieval results from the mismatch between the mental categories into which people assign events and the categories (properties in this case) on which they are later tested: Unless properties are salient when events are en-

coded, it is extremely difficult to later reorganize those events in terms of their properties.

This differs from so-called ad hoc categories (Barsalou, 1983, 1991), which are not preexisting but can be derived as needed. For example, *things to sell at a garage sale* is presumably organized around the goal of eliminating unwanted possessions. Although its instances are most likely not classified in this way before the category is needed, the instances presumably have preexisting features that are relevant to the goal (e.g., *unwanted possessions*). Our results concern situations in which there really is no prior mental structure or association to constrain retrieval.

It is possible that event properties are not routinely encoded because they are simply not as effective in distinguishing events as are taxonomic relations. Presumably, retrieving an event from a taxonomic structure involves searching increasingly smaller numbers of targets the deeper one proceeds into the hierarchy. However, it is hard to imagine a structure based on properties that promotes the same kind of efficiency. Moreover, it is hard to converse about events on the basis of their properties if different people represent the same event according to different properties—for example, what is fast to one person may be loud to another. (The lack of consensus by Underwood & Richardson's, 1956, participants suggests this may be the case.) This too would be a disincentive to encode event properties.

The inherently ambiguous relationship between events and properties is likely to pose practical problems for people interacting with artifacts designed without this issue in mind. Consider the following survey question:[6] "How often do you do light or moderate activities for at least 10 min that cause only light sweating or a slight to moderate increase in breathing or heart rate?" Introspection suggests the event category in the question is at odds with natural classification of the target events, leading, most likely, to inaccurate answers.

Our point is that experimental participants' insensitivity to property frequency is a symptom of a bigger problem. People cannot easily reclassify events they have experienced. When people try to recall events, using a criterion irrelevant to their original classification, the process is laborious and, in the end, largely fruitless. Although relatively inflexible in how they retrieve events, people are actually quite flexible in how they organize events: Properties can be an effective organizing criterion if they are salient. In the end, in order to understand estimation processes, we need to consider the underlying classification processes.

REFERENCES

Alba, J. W., Chromiak, W., Hasher, L., & Attig, M. S. (1980). Automatic encoding of category size information. *Journal of Experimental Psychology: Human Learning & Memory*, **6**, 370-378.

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition*, **11**, 211-227.

Barsalou, L. W. (1991). Deriving categories to achieve goals. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 27, pp. 1-24). New York: Academic Press.

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, & Cognition*, **12**, 116-134.

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 Monographs*, **80**(3, Pt. 2), 1-46.

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

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

Brown, N. R. (2002). Encoding, representing, and estimating event frequencies: A multiple strategy perspective. In P. Sedlmeier & T. Betsch (Eds.), *Frequency processing and cognition* (pp. 37-53). Oxford: Oxford University Press.

Burton, S., & Blair, E. (1991). Task conditions, response formulation processes, and response accuracy for behavioral frequency questions in surveys. *Public Opinion Quarterly*, **55**, 50-79.

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

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (rev. ed.). Cambridge, MA: MIT Press.

Freund, S. J., & Hasher, L. (1989). Judgments of category size: Now you have them, now you don't. *American Journal of Psychology*, **102**, 333-353.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple trace memory model. *Psychological Review*, **95**, 528-551.

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.

Means, B., & Loftus, E. F. (1991). When personal history repeats itself: Decomposing memories for recurring events. *Applied Cognitive Psychology*, **5**, 297-318.

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.

Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory & Cognition*, **17**, 759-769.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, **4**, 207-232.

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

Whittlesea, B. W. A. (1993). Illusions of familiarity. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **19**, 1235-1253.

Williams, M. D., & Hollan, J. D. (1981). The process of retrieval from very long-term memory. *Cognitive Science*, **5**, 87-119.

NOTES

1. The three groups of 8 participants were recruited and tested in sequence—that is, one group was run before the next. In a strict sense, then, the participants were not randomly assigned to groups. However, because the participants across the groups were selected from the same diverse participant pool and their demographic characteristics were comparable, there is no reason to believe that any of the results are related to the assignment procedure.

2. Percentage of use was analyzed for four strategies: simple enumeration, adjusted enumeration, general impressions, and unjustified responses. So that percentage of use would not sum to 100, consuming

all degrees of freedom for the strategy factor, uncodable responses were excluded from the analysis.

3. Although it is possible that the participants in the explicit property group treated the study word and the property label as paired associates, rather than as an instance of a property, there is good reason to believe they did encode the property information. When instances and category labels are arbitrarily paired (Brown, 1997), the range of strategies is severely restricted, in contrast to the rich set of strategies we observed in the explicit property group.

4. In both of these studies, weaker regression effects were also observed under conditions that resembled our taxonomic and explicit property conditions. However, these estimates were correlated with actual frequency, whereas there was no such relation for the groups that resembled our implicit property group.

5. As in Experiment 1, actual presentation frequency in the implicit property condition was determined by treating each presentation of an instance as a presentation of the normatively associated property.

6. National Health Interview Survey conducted by the National Center for Health Statistics (NHIS: AHB.110).

(Manuscript received August 1, 2001;
revision accepted for publication February 28, 2003.)