Journal of Experimental Psychology:
Learning, Memory, and Cognition
1995, Vol. 21, No. 6, 1539-1553

0278-7393/95/$3.00

Estimation Strategies and the Judgment of Event Frequency

Norman R. Brown
University of Alberta

Processes underlying judgments of absolute event frequency were investigated in 3 experiments. In all 3, word pairs consisting of a target (a category label, e.g., *CITY*) and context (a category exemplar, e.g., *London*) were presented in a different- or same-context study list. In the different-context condition, each target was paired with a new context on each presentation; in the same-context condition, a target always appeared with the same context. Verbal protocols (Experiment 1) and response times (Experiments 2 and 3) indicate that multiple estimation strategies were used and that strategy selection was related to memory contents. In particular, different-context participants often enumerated, and same-context participants did not. Also, because range information only affected same-context estimates (Experiment 3), it appears that a numerical conversion process was necessary when nonenumeration strategies were used.

Judgments of absolute frequency are collected in most experiments concerned with the encoding and representation of event frequency (for reviews see Hasher & Zacks, 1979, 1984; Hintzman, 1976, 1988; Howell, 1973). The research described in this article was aimed at understanding how these judgments are produced. In particular, I argue that participants use multiple strategies when estimating event frequency; that strategy selection is determined, in part, by the contents of memory; and that the magnitude and accuracy of participants' frequency estimates are related to the strategy they select.

The multiple-strategy perspective adopted in this article has numerous precedents. It is well established that people use multiple strategies to perform a wide variety of simple and not-so-simple cognitive tasks. Among these are recognition (e.g., Mandler, 1980), mental arithmetic (e.g., Siegler, 1987), mental rotation (e.g., Just & Carpenter, 1985), question-answering (e.g., Reder, 1987), reading (e.g., Aaronson & Ferres, 1986), real-world estimation (e.g., Brown, 1990; Brown, Rips, & Shevell, 1985), problem solving (e.g., Simon & Reed, 1976), and decision making (e.g., Payne, Bettman, & Johnson, 1992). Of more direct relevance, there is good evidence that survey respondents use multiple strategies to answer "behavioral frequency" questions (Blair & Burton, 1987; Burton & Blair, 1991; Conrad, Brown, & Cashman, 1993; Means & Loftus, 1991; Menon, 1993; Menon, Raghubir, & Schwarz, 1993; see also Bruce & Van Pelt, 1989). These questions require respondents to estimate the number of times they have engaged in a particular activity during a given reference period, for example, "How many times have you shopped for groceries during the last month?" In addition, many memory researchers have observed that there may be more than one way to generate a frequency judgment (Begg, Maxwell, Mitterer, & Harris, 1986; Bruce, Hockley, & Craik, 1991; Hintzman, 1976; Howell, 1973; Johnson, Raye, Wang, & Taylor, 1979; Jonides & Jones, 1992; Jonides & Naveh-Benjamin, 1987; Voss, Vereb, & Bisanz, 1975). Interestingly, the implications of this observation are just beginning to be investigated (e.g., Bruce et al., 1991; Marx, 1985). As a result, factors that lead participants to choose one strategy over others, and the consequences of these choices, are not well understood. This is unfortunate because strategy selection can have a strong effect on estimation performance (Burton & Blair, 1991). This means that performance differences found in frequency estimation tasks can be very difficult to interpret: They may reflect differences in the encoding and representation of frequency information, they may reflect differences in strategy use, or they may reflect both strategic and representational differences. One way to reduce this ambiguity is to develop a theory of frequency estimation that identifies people's estimation strategies, details performance characteristics of each strategy, and specifies conditions that promote the use of one strategy over others. The research reported in this article was intended to provide the empirical basis for such a theory.

Although prior research has rarely addressed the issue of strategy selection in frequency estimation, a number of distinct estimation processes have been proposed. Figure 1 presents a taxonomy of these processes. The basic division in this taxonomy is between *enumeration* and *nonenumeration* processes. Enumeration occurs when individual items or events are retrieved and counted and when the count arrived at serves as the basis for an estimate (Barsalou & Ross, 1986; Begg et al., 1986; Blair & Burton, 1987; Bruce et al., 1991; Burton & Blair, 1991; Conrad et al., 1993; Greene, 1989; Menon, 1993; Schmidt, 1978; Williams & Durso, 1986). It is likely that there are two types of enumeration strategies: *simple enumeration* and *enumeration and extrapolation.* When a simple enumeration strategy is applied, the value of the estimate is equal to the

This research was supported by an operating grant from the Natural Sciences and Engineering Research Council of Canada. I would like to thank Navin Arora, Justine Chun, Trish Dmytriw, Tom Harke, Shanny Hwang, Aaron Kucher, Barb Melhorn, Helen Moon, Zahida Rana, Stephanie Schorr, Karen Sumka, Yuko Tainaka, Lauren Weisler, and Gary Wong for their assistance. I would also like to thank Jeff Bisanz, Ann Bostrom, Fred Conrad, Alinda Friedman, Judy Goodman, Bob Heller, Bob Sinclair, and Fred Smith for their useful comments on a version of this article.

Correspondence concerning this article should be addressed to Norman R. Brown, Department of Psychology, University of Alberta, Edmonton, Alberta, Canada T6G 2E9. Electronic mail may be sent via Internet to nbrown@psych.ualberta.ca.

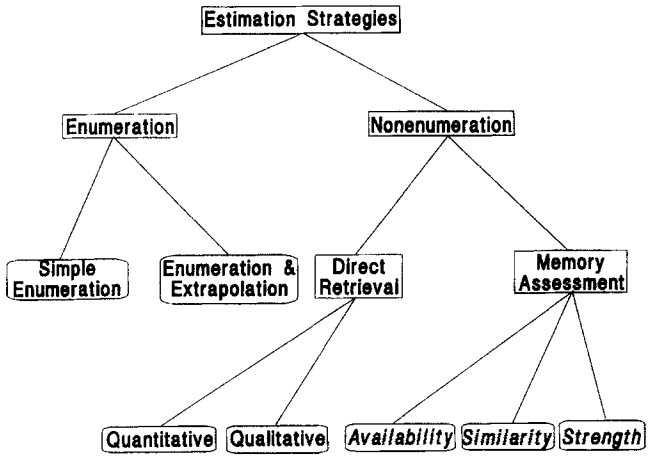

Figure 1. A taxonomy of frequency estimation strategies.

number of relevant episodes retrieved; when an enumeration-and-extrapolation strategy is applied, the value of the estimate is greater than the number of episodes retrieved.

The nonenumeration processes can be divided into *direct retrieval* and *memory assessment* strategies. Direct retrieval models assume that facts about event or item frequency are directly stored in memory. These facts may be explicitly *quantitative* (e.g., "*cup* appeared on the study list 6 times"; Jonides & Jones, 1992; Underwood, 1969) or they may express frequency in a more *qualitative* manner (e.g., "*cup* appeared on the study list several times"; Alba, Chromiak, Hasher, & Attig, 1980; Brooks, 1985; Watkins & LeCompte, 1991). Either way, direct retrieval positions assert that these facts are retrieved from memory during the estimation process and that the estimated value is determined by the contents of the retrieved fact.

Memory assessment approaches assume that some aspect of memory performance is evaluated during the estimation process and that the outcome of this evaluation serves as the basis for a frequency judgment. For example, Tversky and Kahneman (1973) have proposed that *availability*—the ease with which relevant information is retrieved—plays an important role in frequency estimation. A person using availability to judge event frequency would provide high estimates when relevant information is easy to retrieve and low estimates when it is not.[1] Others have proposed that people evaluate the *strength* of a unitary memory representation (Hintzman, 1969; Morton, 1968) and that they assign high values to items judged to have strong representations. Finally, a number of recent models have assumed that the *similarity* between a probe item and the contents of episodic memory is used as an index of item frequency (Hintzman, 1988; Jones & Heit, 1993; Nosofsky, 1988). Here, frequency estimates are relatively high when the target item closely resembles the items stored in memory and low when it does not.

It should be noted that with the exception of quantitative direct retrieval, nonenumeration strategies produce a relative or qualitative evaluation of event frequency. In principle, these relative values must be converted to numerical ones before participants can respond with a judgment of absolute frequency. Prior research suggests that this conversion process is independent of the assessment processes that feed into it (Brown & Siegler, 1993) and that the judgments it produces can strongly be affected by the way that stimulus and response scales are defined (Anderson, 1982; Poulton, 1982; Stevens, 1975). Because enumeration-based strategies deliver *numerical* information in the form of counts, participants should not have to engage a conversion process. As a result, participants should be far less susceptible to scale effects when they enumerate than when they do not. This prediction is directly tested in Experiment 3.

The taxonomy presented in Figure 1 is useful because it organizes existing theoretical positions and illustrates the variety of processes capable of generating frequency estimates. In addition, there is evidence that people use at least some of these strategies to answer behavioral frequency questions. In particular, a number of researchers have found that survey respondents often enumerate when event instances are judged to be distinctive and that they rely on nonenumeration strategies when they are not (Conrad et al., 1993; Means & Loftus, 1991; Menon, 1993). This finding suggests that distinctive event instances produce memory traces that are readily retrieved and easily distinguished and that the presence of such traces fosters the use of enumeration-based estimation strategies. In contrast, it appears that event instances that are very similar to one another blend in memory, making it difficult, if not impossible, to estimate frequencies by retrieving individual traces.

The three experiments reported in this article were designed, in part, to determine whether distinctiveness and strategy selection are related in the laboratory in the same way that they are in the real world. In all three, participants studied a list of word pairs and then provided judgments of absolute frequency. Each pair consisted of a target word and a context word. In all cases, the target word was a category label, and the context word was a category exemplar. In one condition, the *different-context* condition, a target word was paired with a different context word each time it appeared on the study list (e.g., *CITY–Boston, CITY–Cleveland, CITY–London*). In a second condition, the *same-context* condition, a target word was paired with the same context word each time it appeared (e.g., *CITY–London, CITY–London, CITY–London*). The context manipulation was expected to influence the representation of the target words in memory, with only the different-context condition producing distinctive memory traces for the various presentations of a given target word. If enumeration is related to distinctiveness, as the behavioral frequency literature suggests, and if the different-context condition produces distinctive memory traces and the same-context condition does not, then different-context participants should rely on enumeration strategies and same-context participants should rely on nonenumeration strategies.

Context manipulations similar to the ones used here have been used in studies of encoding variability (Begg et al., 1986;

[1] Tversky and Kahneman (1973) used *availability* to refer to two distinct estimation strategies: enumeration and extrapolation and assessment of retrieval difficulty. In this article, the term is used to refer only to the latter.

Hintzman & Stern, 1978; Johnson et al., 1979; Jonides & Naveh-Benjamin, 1987; Rose, 1980; Rowe, 1973; Voss et al., 1975) and categorical frequency estimation (Alba et al., 1980; Barsalou & Ross, 1986; Begg et al., 1986; Brooks, 1985; Bruce et al., 1991; Greene, 1989; Hanson & Hirst, 1988; Watkins & LeCompte, 1991; Williams & Durso, 1986). The current study differed from these prior studies in one important respect: It used two on-line methods, concurrent verbal protocols and response times, to investigate the frequency estimation process. Indeed, this is only the second study in the experimental literature to report response times collected from participants as they estimate event frequencies (Voss et al., 1975);[2] it is the first to make use of concurrent verbal protocols (but see Marx, 1985); and it is the only one to use response times and verbal reports in conjunction. Specifically, three frequency estimation experiments are reported. In Experiment 1, participants thought aloud as they generated frequency judgments; in Experiments 2 and 3, participants were timed as they performed the same task. The protocols collected during Experiment 1 were used to identify the participants' estimation strategies, to relate strategy selection to event context and presentation frequency, and to determinate how strategy selection affects frequency judgments. Response times and frequency estimates collected in Experiments 2 and 3 provided converging evidence for the relations identified in Experiment 1.

Experiment 1

In this experiment, one group of participants was presented with a different-context study list, and a second group with a same-context study list. All participants were then presented with the same target words and asked to think aloud as they estimated how frequently each had appeared in the study list. Presentation context was predicted to affect both strategy selection (and hence the contents of the verbal protocols; Ericsson & Simon, 1984) and estimation performance. For reasons stated above, it seemed likely that different-context participants would enumerate and that same-context participants would not. In addition, different-context participants were expected to provide smaller estimates than same-context participants. This pattern of performance has previously been observed (Hintzman & Stern, 1978; Rose, 1980; Rowe, 1973; but see Begg et al., 1986; Jonides & Naveh-Benjamin, 1987) and is readily interpreted in terms of the multiple-strategy position (see below).

Method

Design and materials. Participants studied a list of 260 word pairs. Each pair was composed of a target word and a context word. In all cases, the target word was a one-word category label, and the context word was a category exemplar, one or two syllables in length (e.g., *FISH–trout, COUNTRY–Greece, COLOR–red*). Presentation frequency was varied within subject, and context between subjects. Specifically, six target items were presented at each of the following five levels of presentation frequency: 2, 4, 8, 12, and 16. In the same-context condition, the target word (category label) was paired with the same context word (category exemplar) on each appearance; in the different-context condition, the target word was paired with a different context word on each appearance.

Target and context words were drawn from category norms published by Battig and Montague (1969) and McEvoy and Nelson (1982). Categories were selected to meet two criteria. First, each had to be clearly identified by a single noun (e.g., *mammal, sport, occupation*). These one-word category names served as target items. Second, each category had to include a reasonable number of frequently listed category members. Only frequently listed one- and two-syllable category members served as context items in the different-context condition. In the same-context condition, each category label was paired multiple times with the most frequently listed of its exemplars. Each category was assigned to a single level of presentation frequency, depending on the number of suitable context items available in the norms.

The first four word pairs presented in the study list served as a primacy buffer, and the last four as a recency buffer. Like the other stimulus pairs, each buffer pair consisted of a category label and a category exemplar, though these category labels were not repeated elsewhere in the list. The stimulus pairs were allocated to the remaining 252 list positions so that repetitions of target items were evenly distributed across the list. To do this, the study list was divided in half. Each half list was then divided into eight blocks (six blocks with 16 items/block and two blocks with 15 items/block). Each Category 16 target item appeared once in each block, and each block also included either 4 or 5 Category 12 target items, 3 Category 8 target items, either 1 or 2 Category 4 target items, and either 0 or 1 Category 2 target item. In addition, the repetitions were dispersed across blocks so that no target item appeared more than once per block, and each target item appeared equally often in both halves of the study list.

A separate study list was created for each different-context participant. Each list began with a unique assignment of target items to block, consistent with the constraints just described. The blocks were then randomized within list half, and the target items were randomized within block. Finally, a random ordering was created for the context items associated with each target item. This ordering picked out the specific context word that would be paired with the target item on each repetition. Thus, 1 participant might see the target word *CITY* first paired with *Boston,* then *Cleveland,* and then *London*; a second participant might see *CITY* paired with *Paris,* then *Memphis,* and then *Dallas.* A same-context participant was yoked to each different-context participant. The target items were presented in the same order to both participants in the yoked pair.

Test lists comprised 30 target items and six category labels that did not appear in the study list. The latter served as 0-frequency catch trials. A different test list was constructed for each different-context participant in the following manner. First, the list was divided into six blocks, with one target item from each frequency level (0, 2, 4, 8, 12, 16) randomly assigned to each block. The target words were then randomly ordered within blocks. Each same-context participant received the same test list as his or her different-context counterpart.

Procedure. Except for the nature of the context items, the experimental procedure was identical in the same- and different-context

[2] In this study, participants were first presented with a list of CVC-word pairs and then were timed as they estimated the list frequency of the CVC target items. In general, participants produced their estimates rapidly (under 3 s), and there was a tendency for response times to increase slightly and then decrease across the range of presentation frequencies. These results suggest that Voss et al.'s participants relied on nonenumeration strategies (see Experiments 2 and 3). It should also be noted that response times have been used to study continuous frequency estimation (Hockley, 1984) and comparative frequency judgments (Hintzman & Gold, 1983; Hintzman, Grandy, & Gold, 1981), though this research is not directly relevant to the issues addressed in this article.

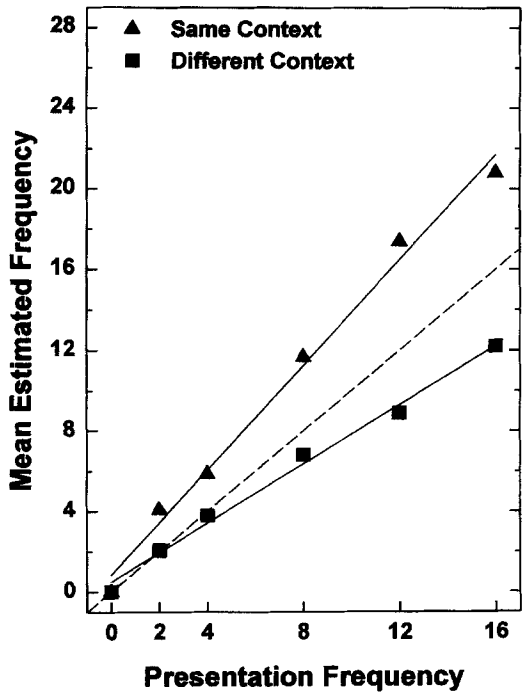

Figure 2. Mean estimated frequencies for different- and same-context participants in Experiment 1. The solid lines represent the best linear fit for the means, and the dashed line represents the actual frequencies.

conditions. Before the presentation of the study list, all participants were told that they would see 260 word pairs and that each pair would include a category name and a category exemplar. They were instructed to study the pairs for a later memory test, but they were not informed of the nature of the test. A computer-controlled video monitor presented the study list. The target–context pairs were displayed one at a time. In all cases, the target item appeared in the center of the screen in uppercase letters, and the context item appeared two lines beneath in lowercase letters. Each pair was displayed for 5.5 s. The screen was then erased and remained blank (except for markers indicating the screen positions of the target and context word and a trial counter) for .5 s; the next target–context pair was then displayed.

After the presentation of the study list, participants read instructions for the test phase. These instructions informed participants that they would be presented with 36 category names and that their task was to estimate as accurately as possible the number of times that each name had appeared in the study list. Participants were also told that they would be required to think aloud as they formulated their estimates, and they were warned that they would be prompted to say something if they fell silent for more than a few seconds. The instructions did not provide an upper bound for the response range but did provide an implicit lower bound by informing participants that some of the test items did not appear in the study list.

The presentation of the test words was self-paced. The participant initiated a trial by pressing the enter key on the computer keyboard. A category label then appeared in the center of the display, along with a response field two lines beneath. The participant read the category name aloud, described his or her thoughts to the experimenter, and then entered an estimate at the computer keyboard. Finally, the participant pressed the enter key, causing the current category name and estimate to be erased and replaced by a message informing the participant to initiate the next trial.

All verbal responses were tape recorded, though only the last 30 were analyzed, as the first 6 served as a practice block. During the practice trials, the experimenter actively encouraged participants to speak and frequently prompted them if they did not.

Participants. Forty University of Alberta undergraduates took part in this study. Half were randomly assigned to the different-context condition, and half to the same-context condition. Participants were individually tested in sessions lasting about 45 min and received course credit for their cooperation.

Results

Frequency estimates. Data from the first 6 test trials were eliminated because they made up the practice block: the remaining 30 trials were analyzed. This set included five target items from each of six levels of presentation frequency (0, 2, 4, 8, 12, 16). From these data, for each participant and for each level of presentation frequency, two means were computed. One was simply the mean of the estimated frequencies, and the other was the mean of the absolute errors (i.e., |estimated frequency − actual frequency|). These data were submitted to separate Context (same vs. different) × Presentation Frequency (0, 2, 4, 8, 12, 16) analyses of variance (ANOVAs).[3] Two additional measures were computed for each participant. One was the rank-order correlation between estimated and actual frequency for the 30 test trials.[4] The other was a regression slope obtained by fitting estimated frequency against actual frequency. The former provided a measure of relative accuracy, and the latter indicated the degree to which participants were biased to overestimate or underestimate event frequencies.

Figure 2 presents the mean estimated frequency plotted against presentation frequency for different-context and same-context participants. These data indicated that different-context participants tended to underestimate event frequencies, that same-context participants tended to overestimate them, and that the tendency to underestimate or overestimate increased with presentation frequency. This Context × Presentation Frequency interaction was statistically significant, $F(5, 190) = 7.0$, $p < .0001$, $MS_E = 18.37$, as were the main effects for context, $F(1, 38) = 13.9$, $p < .001$, $MS_E = 80.67$, and for presentation frequency, $F(5, 190) = 86.3$, $p < .0001$, $MS_E = 18.37$. Regression slopes reflected the same pattern of underestimation and overestimation. Specifically, the average slope was .73 in the different-context condition and 1.30 in the same-context condition, $t(38) = 3.0$, $p < .01$, $MS_E = 0.19$, indicating that estimated frequency increased less rapidly than

[3] In all three experiments, within-subject medians were computed for estimated frequency and absolute error. In addition, both means and medians were computed for response times collected in Experiments 2 and 3. Analyses based on these medians are not reported because, in all cases, means and medians were similar in size and displayed the identical pattern of effects.

[4] In this experiment and the following ones, rank-order correlations were transformed using Fisher's *r*-to-*z* method before being submitted to statistical tests. The mean correlations reported below were obtained by back-transforming the corresponding *z*-score means.

presentation frequency in the different-context condition and more rapidly in the same-context condition.

In brief, context and presentation frequency influenced the magnitude of participants' estimations. These variables also affected estimation accuracy. There are two ways to measure accuracy. One can consider how close estimated frequency is to actual frequency or how sensitive a set of estimates is to differences in actual frequency. Absolute error provides a measure of the former, and the rank-order correlation between estimated and actual frequency provides a measure of the latter (Flexser & Bower, 1975; Naveh-Benjamin & Jonides, 1986). Both of these measures indicated that different-context participants were somewhat more accurate than same-context participants. Absolute error was significantly smaller in the different-context condition ($M = 2.7$) than in the same-context condition ($M = 5.3$), $F(1, 38) = 14.0, p < .001, MS_E = 29.69$, and the rank-order correlation was significantly larger, $t(38) = 3.8, p < .001, MS_E = .12$. The mean rank-order correlation was .95 in the different-context condition and .87 in the same-context condition. As is typical in frequency estimation studies, absolute error increased with frequency, $F(5, 190) = 50.9, p < .0001, MS_E = 9.21$. In addition, there was a significant Context × Frequency interaction, $F(5, 150) = 4.0, p < .05, MS_E = 9.21$, indicating that absolute error increased with frequency more rapidly in the same-context condition than in the different-context condition (see Table 1).

Protocols. Two naive judges, working together, coded the verbal reports. These judges used a common coding scheme to score both same-context and different-context responses and were instructed to reach a consensus on all decisions. The coding scheme had five elements. First, the coders recorded the number of context words contained in a response. Second, they determined whether the protocol contained a vague quantifier (Wright, Gaskell, & O'Muircheartaigh, 1994). These were phrases like "It occurred a lot," "There weren't too many of those," and "It showed up quite often" that expressed a general impression of event frequency in nonnumerical terms. Third, they noted whether the participant asserted that the target word had not appeared in the study list, and fourth, they noted when the participant mentioned frequency-relevant information not covered by other categories in the coding scheme. For example, participants occasionally recalled a previous estimate and used it to anchor the current one. Finally, responses were coded as *unjustified* when participants offered an estimate without providing a rationale or mentioning any type of frequency-relevant information.

These codes were used to assign each response to at least one of the following seven response-type categories: general impressions, new target, miscellaneous, unjustified, simple enumeration, enumeration and extrapolation, and (single) context retrieval. The first four categories were applied to both different-context and same-context responses and corresponded directly to the coders' decisions regarding the relevant elements in the coding scheme. The last three categories were used to classify responses that included one or more context words. The enumeration-based categories applied to different-context responses only and were mutually exclusive. Specifically, a different-context response was assigned to the simple-enumeration category when the frequency estimate produced by the participant equaled the number of context words mentioned in his or her protocol. When the former exceeded the latter, the response was assigned to the enumeration-and-extrapolation category. The context-retrieval classification was used when a same-context participant mentioned the target item's context word in a response.[5]

Table 1
Mean Absolute Error at All Levels of Presentation Frequency for Different- and Same-Context Conditions From Experiments 1 and 2

Presentation frequency	Experiment 1			Experiment 2		
	Different	Same	*M*	Different	Same	*M*
0	0.0	0.0	0.0	0.2	0.1	0.2
2	0.6	2.4	1.5	0.6	1.4	1.0
4	1.2	2.9	2.1	1.4	2.7	2.1
8	3.1	6.4	4.8	3.5	5.0	4.3
12	4.6	9.7	7.2	5.3	6.8	6.1
16	6.7	10.7	8.7	7.8	10.3	9.1
M	2.7	5.3	4.0	3.1	4.4	3.8

Data summarizing the protocol analysis are presented in Table 2. The values listed in this table represent the proportion of trials assigned to each of the response types just described.[6] This table also includes data for two aggregate categories. In the different-context condition, simple enumeration and enumeration-and-extrapolation were summed to create a total enumeration score. This score provides a measure of the likelihood that different-context participants used either of the enumeration-based strategies listed in Figure 1. An aggregate *uninformative response* category was defined for the same-context condition. Two types of same-context responses were placed in this category: unjustified responses and responses that involved only the retrieval of the target's context word. The latter were considered to be uninformative because context retrieval implies only that the target appeared on the study list. Because this is true for all nonzero targets, retrieving a context provides very little information about presentation frequency.

A set of one-way ANOVAs was performed on the protocol data. In each analysis, presentation frequency served as the independent variable, and counts representing the number of

[5] Two things should be noted about this classification scheme. First, same-context participants could have enumerated by recalling aspects of the context, other than the context word, that differed across presentations of a given target word. However, because there was no evidence for this in the protocols, enumeration-based categories were not used to classify same-context responses. Second, different-context participants occasionally reported only one context word (i.e., produced a response that could have been scored as a context retrieval). However, because such responses were uncommon and could not be distinguished from enumeration-based responses, the (single) context-retrieval category was not used to classify different-context responses.

[6] Responses to the catch trials were excluded from Table 2 and from the ANOVAs because most of the categories used to classify responses were inapplicable to the estimates elicited by these items. Not surprisingly, over 90% of the catch trials were assigned to the new-target category, and the remaining responses were unjustified.

Table 2
Proportion of Verbal Reports Assigned to Response Types in Experiment 1

Response type	Presentation frequency 2	4	8	12	16	*M*
Different context						
Total enumeration[a]	.55	.59	.63	.62	.47	.57
Simple enumeration	.45	.34	.26	.29	.06	.28
Enumeration and extrapolation	.10	.25	.37	.33	.41	.29
General impression	.07	.14	.14	.26	.40	.20
Unjustified	.33	.29	.24	.19	.21	.25
Miscellaneous	.00	.00	.04	.02	.03	.02
New target	.06	.03	.00	.00	.00	.02
Same context						
Uninformative[b]	.83	.66	.71	.67	.58	.69
Context retrieval	.53	.60	.67	.63	.67	.62
Unjustified	.39	.26	.27	.24	.23	.28
General impression	.13	.28	.18	.26	.36	.24
Miscellaneous	.04	.04	.10	.08	.09	.07
New target	.03	.03	.01	.00	.00	.01

[a]Total enumeration is simple enumeration plus enumeration-and-extrapolation. [b]Uninformative is unjustified responses plus responses based on context retrieval only.

responses identified as instances of a given response type served as the dependent variable. In addition, the results of an Enumeration Type (simple vs. extrapolated) × Presentation Frequency ANOVA are reported below; the dependent variable here was a count representing the number of times that participants used simple enumeration or a enumeration-and-extrapolation strategy.

The data presented in Table 2 indicate that different-context participants relied heavily on enumeration-based strategies; 57% of the responses to nonzero target words involved either simple enumeration (28%) or enumeration and extrapolation (29%). It is also apparent that simple enumeration became less common as presentation frequency increased and that enumeration and extrapolation became more common. This Enumeration Type × Presentation Frequency interaction was significant, $F(4, 156) = 8.9, p < .0001, MS_E = 0.93$, as was the main effect of presentation frequency, $F(4, 156) = 3.1, p < .05$, $MS_E = 0.17$. A set of Fisher's adjusted least significant difference tests indicated that participants were significantly less likely to enumerate when responding to items presented 16 times (47%) than to those presented at other frequencies. In addition, they were significantly less likely to enumerate when the target items had appeared twice (55%) than when they had appeared 8 times (63%).

Although different-context participants often enumerated, it was also common for them to use general impression statements (20%) and to provide unjustified responses (25%). The percentage of responses that included a general impression statement increased from 7% to 40% as presentation frequency increased from 2 to 16, $F(4, 76) = 8.5, p < .0001$, $MS_E = 0.99$, and the percentage of unjustified responses decreased slightly from 33% to 21%, over the same range, $F(4, 76) = 2.0, p > 1.0, MS_E = 0.84$.

As noted above, not all response types were mutually exclusive. Thus, it was possible for a response to be classified as an instance of more than one. In the different-context condition, 6% of the responses were judged to involve more than one type of frequency-relevant information. These responses typically included both enumeration and general impression statements and were more common at high frequencies than at low frequencies, $F(4, 76) = 3.7, p < .01, MS_E = 0.33$; for Frequencies 2, 4, 8, 12, and 16, the percentages of these responses were 1%, 3%, 5%, 10%, and 13%, respectively.

In the same-context condition, 62% of the responses included reference to the target item's context word, 28% were unjustified, and 24% included a general impression statement. As these numbers suggest, uninformative responses (69%) were much more common than informative ones (31%). As in the different-context condition, presentation frequency and probabilities associated with the main response types were related. On the one hand, the probability of mentioning the target word's context increased with presentation frequency (from 53% to 67%), $F(4, 76) = 3.4, p < .05, MS_E = 0.50$, as did the probability of stating a general impression (from 13% to 36%), $F(4, 76) = 5.4, p < .001, MS_E = 0.75$. On the other, both uninformative responses and unjustified responses were less common at higher frequency than at lower frequencies. The percentage of uninformative responses dropped from 83% to 58% across the range of presentation frequencies, $F(4, 76) = 4.8, p < .01, MS_E = 0.88$, and the percentage of unjustified responses dropped from 39% to 23% across the same range,[7] $F(4, 76) = 4.0, p < .01, MS_E = 0.53$.

Discussion

The results of Experiment 1 revealed large between-group differences in the contents of participants' verbal reports and the magnitude of their frequency estimates. As predicted, enumeration played an important role in the different-context condition. Almost 60% of the responses produced by the different-context participants were enumerated, with simple enumeration predominating at the smaller frequencies, and enumeration and extrapolation at the higher ones. In contrast, same-context participants did not retrieve and count multiple event instances and were generally unwilling or unable to provide a rationale for their estimates; 69% of the same-context responses were uninformative. These results, which have been replicated using a different stimulus set, study list structure, and subject pool (Conrad & Brown, 1994; see also Marx, 1985), suggest that experimental participants, like survey respondents, are often willing to enumerate when instances of the target event are distinctive and that they depend

[7] Participants were much less likely to produce responses that included both retrieved-contexts and general impression statements at low frequencies than at high frequencies. For example, at Frequency 2, only 6% of the responses involving context retrieval also included a general impression statement; at Frequency 16, 42% did. The tendency to justify a larger percentage of context-retrieval responses at high frequencies than at low frequencies, in conjunction with the tendency to produce fewer unjustified responses, explains how the percentage of uninformative responses can decrease with presentation frequency while the percentage of context-retrieval responses increases.

on nonenumeration strategies when they are not (Burton & Blair, 1991; Conrad et al., 1993; Menon, 1993).

Although the verbal protocols provided evidence for between-group differences in enumeration, they also indicate that different-context participants did not always enumerate and that same-context participants frequently retrieved context words. In addition, general impression statements were about as common in the different-context condition as in the same-context condition, and, in both conditions, they were more common at higher frequencies than at low ones. As noted above, over 40% of the different-context responses did not appear to involve enumeration, over 60% of the same-context responses involved the retrieval of the target's context, and general impression statements appeared almost as often in the different-context condition (20%) as in the same-context condition (24%). The first of these findings suggests that enumeration is not mandatory, even when relevant event instances are highly distinctive. Apparently, strategy selection is restricted by the contents of memory but not dictated by them. The second finding suggests that target and context are often so closely linked in the same-context condition that accessing the former often led to the effortless retrieval of the latter. Finally, findings concerning the use of general impressions are of interest because they suggest that participants, regardless of condition, sometimes encode and retrieve facts about event frequency cast in qualitative terms (Alba et al., 1980; Brooks, 1985; Jonides & Jones, 1992; Watkins & LeCompte, 1991).[8]

Presentation context had a pronounced effect on the magnitude of participants' frequency estimates as well as on the contents of their verbal reports. Participants in the different-context condition tended to underestimate event frequencies, and participants in the same-context condition tended to overestimate them. This is consistent with prior research demonstrating that frequency judgments are often smaller when a target item appears in multiple contexts than when it appears multiple times in the same context (Hintzman & Stern, 1978; Rose, 1980; Rowe, 1973).

A fundamental difference between the enumeration-based strategies, favored by different-context participants, and the nonenumeration strategies, used by same-context participants, can account for the between-group differences in estimation performance observed in this experiment. Enumeration produces numerical information in the form of counts. This information can be used to determine an event's exact or approximate frequency, and it can also be used to draw inferences about the response range and to anchor subsequent estimates. In contrast, nonenumeration strategies typically do not produce numerical information. Rather, these procedures provide a qualitative or relative evaluation of event frequency that must be converted to an appropriate numerical value before participants can respond with a frequency judgment.

The underestimation observed in the different-context condition can be traced to participants' preference for enumeration. In general, the number of instances retrieved will be less than the number of instances presented because participants are more likely to forget or fail to retrieve relevant instances than they are to import irrelevant ones or confabulate a new one. This means that estimates based on simple enumeration will often be underestimated. Estimates that involve extrapolation should also be underestimated because adjustments to specific numerical anchors (i.e., the enumerated counts) are generally "insufficient" (Tversky & Kahneman, 1974).

It is probable that different-context participants use numerical information generated by enumeration-based strategies to draw inferences about the statistical properties of the response range and to determine how frequently individual items appeared in the study list (Brown & Siegler, 1993). Because enumeration leads to conservative frequency estimates, it should foster a biased set of range assumptions, one in which the upper bound of the response range is smaller than the upper bound of the stimulus range, the subjective mean is smaller than the objective mean, and so forth. These range assumptions are very important when participants do not enumerate. As mentioned above, nonenumeration strategies typically yield information about relative event frequency. That is, they indicate whether a given item was very common, very uncommon, or somewhere in-between. Having determined an item's relative frequency, participants must convert from a relative value to an absolute one. This can be done in a number of ways. For example, participants might consider the range of possible responses and select a value from that portion of the range that corresponds to the target's relative frequency, they might recall a prior response to a comparable item and use it as a quantitative anchor or reference point, or they might distribute their responses around a number assumed to represent the central value of the target dimension. Regardless of the details of the conversion process, estimates based on relative frequency information should reflect beliefs about the response range; other things being equal, estimates will be relatively large when participants believe the response range encompasses large values and relatively small when they do not (Anderson, 1982; Poulton, 1982; Stevens, 1975; see also Rowe & Rose, 1977; Smith, Hager, Palphreyman, & Jobe, 1992). In the different-context condition, for reasons just described, participants are likely to adopt a conservative response range and/or set of quantitative reference points, and hence to underestimate event frequencies even when they do not enumerate.

The protocol data suggested that same-context participants

[8] In principle, both direct retrieval and memory assessment strategies can give rise to general impression statements. Such statements may appear when a participant has retrieved a previously stored nonnumerical fact or when he or she verbalizes quantitatively imprecise intuitions resulting from an evaluation of availability, similarity, or trace strength. There are, however, two reasons for favoring the former interpretation. First, if participants produce general impression statements only when they lack other things to say or only when they have used a memory assessment strategy, these statements should have been more common in the same-context condition than in the different-context condition, and they were not. Second, in both conditions, general impression statements were more common at the higher frequencies than at the lower ones. One way to explain this is to assume that frequency information is encoded or updated probabilistically. If so, frequently presented items are more likely to have frequency information associated with them than rarely presented items, and thus participants are more likely to encounter prestored facts when responding to the former than to the latter.

rarely if ever enumerate. This has two interesting consequences. First, it means that range assumptions should play a more prominent role in the same-context condition than in the different-context condition. Second, it means that same-context participants must establish their range assumptions without the benefit of enumeration-based counts. At this point it is unclear how participants do this, but there is no a priori reason to believe that they are more likely to define a response range that is broader than the stimulus range than to define one that is narrower. Nonetheless, estimated frequency was considerably greater than actual frequency in this condition. There is a simple explanation for this finding. Participants can adopt a response range that is very much larger than the stimulus range but not one that is very much smaller. As a result, frequencies can be grossly overestimated but not grossly underestimated. Consistent with this view, 8 of the 20 same-context participants produced at least one estimate that was greater than 32 (i.e., at least twice the size of the largest actual frequency). In contrast, only one same-context participant produced estimates that were never larger than 8 (i.e., no larger than half the magnitude of the largest actual frequency). The connections between context, strategy selection, and numerical conversion are investigated further in Experiment 3.

In brief, it can be argued that the underestimation observed in the different-context condition was a necessary consequence of enumeration and that the overestimation observed in the same-context condition occurred because participants relied on nonnumerical strategies and were given no information about the upper bound of the response range. This is not the only way to explain the effect of context variability on the magnitude of frequency judgments. For example, Hintzman (1988) was able to simulate this effect using a modified version of his MINERVA 2 model. This model assumes that all frequency judgments are produced by a single memory assessment process. This process computes the similarity between a probe and each trace stored in memory and delivers a large value when the probe resembles many traces and a small one when it does not. According to this model, context variability effects occur because instances of same-context events resemble one another more than instances of different-context events, and hence same-context probes produce more "intense echoes" than different-context probes.

The MINERVA 2 model provides a parsimonious explanation for the context variability effect and for many other findings reported in the frequency literature. However, because this model assumes that all frequency estimates are generated by a single memory assessment process, it cannot account for the presence of enumeration-based responses in the different-context condition. Of course, it might be argued that participants enumerated in this study only because they felt compelled to provide a verbal justification for their responses and that they would not do so if they were not required to think aloud. This issue is addressed in Experiment 2.

Experiment 2

In Experiment 1, verbal protocols produced by different-context participants were very different from those produced by same-context participants; different-context participants tended to retrieve and count category exemplars, and same-context participants did not. This was taken as evidence that different-context participants rely on enumeration-based strategies and that same-context participants use a variety of nonenumeration strategies. This interpretation is consistent with prior research on behavioral frequency estimation (e.g., Conrad et al., 1993; Menon, 1993) and helps explain between-group differences in estimation performance. However, converging evidence is still necessary to support these conclusions because verbal protocols do not always reflect normal cognitive processing in an accurate manner (Nisbett & Wilson, 1977; Russo, Johnson, & Stephens, 1989; Wilson, 1994). This experiment was designed to provide this evidence.

In this experiment, participants were timed as they generated their frequency estimates. If different-context participants often enumerate and same-context participants depend on nonenumeration strategies as the protocol data suggest, then response times should increase with event frequency in the different-context condition but not in the same-context condition. Response times should increase in the different-context condition because enumeration involves the serial retrieval of category exemplars. Thus, it should take participants more time to retrieve two exemplars than one, more time to retrieve three than two, and so forth (Bousfield & Sedgewick, 1944; Gruenewald & Lockhead, 1980; Indow & Togano, 1970). More generally, participants who enumerate should respond more slowly when they retrieve many instances before answering than when they retrieve only a few (Conrad et al., 1993; see also Hartley, 1977, 1981). Nonenumeration strategies do not engage a serial retrieval process. Instead, participants determine the target item's relative frequency (by retrieving a fact from memory or evaluating some aspect of memory performance) and then convert this information to a numerical response. There is no reason to believe that presentation frequency will affect the speed with which either of these operations is performed, and hence no reason to expect that response times and presentation frequency will be related when participants depend on nonenumeration strategies. It follows that presentation frequency should not affect response times in the same-context condition.

The context manipulation used in Experiment 1 was related to the magnitude of participants' estimates and to the content of their verbal reports; different-context participants tended to underestimate event frequencies, and same-context participants tended to overestimate them. This difference should be replicated in Experiment 2 if protocol participants and response time participants estimate event frequencies in the same way.

Method

Design, materials, and procedure. With one major exception, the design, materials, procedure, and instructions used in this study were identical to those used in Experiment 1. This exception involved the procedure followed during the test phase—participants in this experiment were timed as they generated their estimates, but did not describe their thoughts. As in Experiment 1, participants were told that they would be presented with 36 category names and that they would be required to estimate as accurately as possible the number of

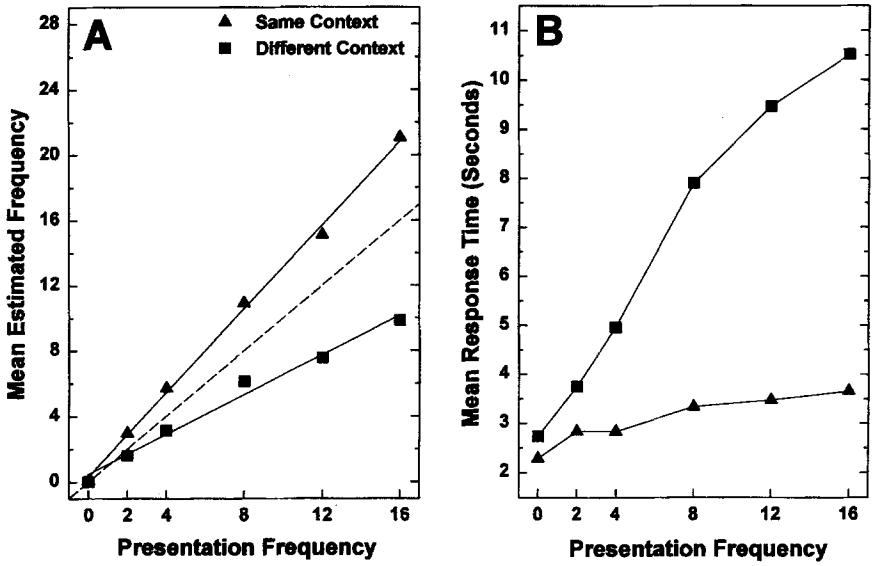

Figure 3. Mean estimated frequencies (Figure 3A) and mean response times (Figure 3B) for different- and same-context participants in Experiment 2. In Figure 3A, the solid lines represent the best linear fit for the means, and the dashed line represents the actual frequencies.

times each name had appeared on the previous list. Participants were also informed that their decision times would be recorded, though the instructions emphasized accuracy over speed.

During the test phase, participants initiated a trial by pressing the enter key on the computer keyboard. This caused a target word to appear in the center of the computer display. Participants were required to read the word and decide how many times it had appeared in the study list. They were instructed to press the keyboard's space bar just as soon as they had "a single numerical response in mind," but not before. When the space bar was pressed, a response field appeared two lines beneath the test word. At this point, participants entered an estimate at the keyboard and then pressed the enter key. This caused the current display to be erased and replaced by a message prompting participants to initiate another trial.

Each trial was divided into three intervals, and a separate response time was recorded for each. The first interval began with presentation of the test word and ended when the participant pressed the keyboard's space bar. This interval indicated how long it took participants to generate a frequency estimate. The second interval began with the space bar response and ended when the participant entered the first digit of his or her estimate; the third began when the first digit was entered and ended when the participant pressed the enter key. Together, these intervals measured how long it took participants to enter their estimates after they had reached a decision.

As in Experiment 1, the first six trials were treated as a practice block. During these trials, the experimenter sat with the participant and made sure he or she understood the task and the test procedure.

Participants. Fifty undergraduates were recruited from the University of Alberta subject pool. Half were randomly assigned to the different-context condition, and half to the same-context condition. Participants were tested individually in sessions lasting about 35 min and received course credit for their cooperation.

Results

Frequency estimates. The frequency estimates collected in this study were processed and analyzed like those collected in Experiment 1. The relevant means are presented in Figure 3A and Table 1. In general, the estimates provided by participants in the two experiments were very similar. As in Experiment 1, different-context participants tended to underestimate event frequencies; same-context participants tended to overestimate them, and these tendencies increased with event frequency. This resulted in a reliable Context × Presentation Frequency interaction, $F(5, 240) = 16.2, p < .0001, MS_E = 13.52$, as well as significant main effects of context, $F(1, 48) = 29.2, p < .0001, MS_E = 54.23$, and presentation frequency, $F(5, 240) = 125.6, p < .0001, MS_E = 13.52$. Consistent with this pattern of underestimation and overestimation, the regression slopes were shallower in the different-context condition ($M = 0.61$) than in the same-context condition ($M = 1.28$), $t(48) = 4.5$, $p < .0001, MS_E = 0.15$.

As in Experiment 1, frequency judgments were quite accurate in both conditions, and absolute error was significantly smaller in the different-context condition ($M = 3.1$) than in the same-context condition ($M = 4.4$), $F(1, 48) = 4.1, p < .05$, $MS_E = 29.02$. In this experiment, relative accuracy was no better in one condition than in the other; the mean rank-order correlation between estimated and actual frequency was .92 for the different-context participants and .91 for the same-context participants, $t(48) < 1, MS_E = 0.08$. Finally, the Context × Frequency interaction for absolute error was not significant, $F(5, 240) < 1.0, MS_E = 8.15$, though the main effect of frequency was, $F(5, 240) = 70.4, p < .0001, MS_E = 8.15$ (see Table 1).

Decision times. Three response latencies were collected on each test trial. The first measured decision time, and the second and third measured the time to initiate and enter a numerical response. Preliminary Context × Presentation Frequency ANOVAs were performed on all three measures and

on a total time measure (the sum of the three). Although only the decision time ANOVA is reported here, three things should be noted about the other timing measures. First, both initiation time (M = 1.0 s) and entry time (M = 0.8 s) were relatively brief. Second, both initiation time and entry time increased slightly (no more than 0.5 s) across the range of presentation frequencies. Third, the total time analysis displayed the same pattern of effects as the decision time analysis.

Mean decision time is plotted against presentation frequency for both different-context and same-context participants in Figure 3B. The data displayed in this figure indicate that presentation frequency had a strong effect on response times in the different-context condition but not in the same-context condition. This Context × Presentation Frequency interaction was significant, $F(5, 240) = 14.6, p < .0001, MS_E = 6.24$, as were the main effects of group, $F(1, 48) = 15.3, p < .001, MS_E = 59.32$, and frequency, $F(1, 48) = 27.2, p < .0001, MS_E = 6.24$.

Discussion

As predicted, response time increased steeply with presentation frequency in the different-context condition but not in the same-context condition. This result provides converging evidence for the claim that different-context participants tend to enumerate and that same-context participants depend on nonenumeration strategies, and it rules out the possibility that enumeration occurs only when participants are required to think aloud. In addition, frequency estimates collected in this study closely resemble those collected in Experiment 1; again, different-context participants tended to underestimate event frequencies, same-context participants tended to overestimate them, and these tendencies increased with presentation frequency. Taken together, these results indicate that the protocols collected in Experiment 1 were an accurate source of information about estimation strategies and that the requirement to verbalize did not interfere with the selection or execution of these strategies.

In addition to providing converging evidence for between-group differences in strategy use, response times reported above indicate that different-context participants worked much harder to produce their estimates than same-context participants. Surprisingly, the large between-group differences in response time obtained in this experiment were not matched by large between-group differences in accuracy; different-context participants were only slightly more accurate than same-context participants. This raises an interesting question: Why do different-context participants enumerate if enumeration-based strategies require more effort than nonenumeration strategies but are not necessarily more accurate? There are two possibilities. First, participants may choose to enumerate when they can because enumeration generates explicit numerical information. As mentioned above, this type of information is attractive because it provides a concrete, credible basis for a response. A second possibility is that participants generally prefer nonenumeration strategies to enumeration-based strategies. (After all, the former are less effortful than the latter and almost as accurate.) This view assumes that a separate trace is created each time a target item appears in a unique, memorable context and that standard memory assessment processes have difficulty gauging item frequency when the relevant traces are highly distinctive. According to this view, different-context participants rely on enumeration-based strategies only because they are unable to use nonenumeration strategies effectively. The current research was not designed to select between these two positions. However, it is worth noting that different-context participants in Experiment 1 made use of both enumeration-based and nonenumeration strategies. This finding is more consistent with the view that participants are able to choose between enumeration-based and nonenumeration strategies when event instances are distinctive than with the view that participants are compelled to enumerate because of the way that distinctive instances are represented in memory.

Experiment 3

This experiment was designed to determine whether range information influences frequency estimation. As in the prior experiments, half of the participants were assigned to the different-context condition, and half to the same-context condition. Within each condition, participants in one group, the *Boundary 16* group, were informed that no item appeared more than 16 times on the study list; participants in a second group, the *Boundary 24* group, were informed that no item appeared more than 24 times; and participants in a third group, the *control* group, were told nothing about the upper bound of the response range.

The boundary manipulation was expected to have a strong effect on the magnitude of the estimates produced by same-context participants and little, if any, effect on those produced by different-context participants. Same-context participants should be sensitive to differences in the way that the range is defined because they rely on nonenumeration strategies. These strategies provide information about relative event frequency (e.g., the word *CITY* appeared "many times" in the study list) but not absolute frequency. As a result, participants must somehow map a relative value on to a response range before they can provide a numerical estimate. Other things being equal, this mapping process yields smaller estimates when operating with a narrow range than with a wide one (Anderson, 1982; Poulton, 1982; Stevens, 1975). Thus, if same-context participants depend on nonenumeration strategies as the prior experiments suggest, and if they use boundary information provided to them to define their response range, then frequency estimates should be smaller in the Boundary 16 condition than in the Boundary 24 condition. In addition, estimates produced by the same-context control participants should be larger than those produced by the same-context participants in the Boundary 24 group. There is an empirical justification for the latter prediction; 80% of same-context participants in the prior experiments produced maximum estimates greater than 24. If a similar percentage of same-context control participants select numbers greater than 24 to bound their response ranges, estimates produced by these participants should tend to be larger than those produced by participants in the Boundary 24 group.

There were two reasons for predicting that the boundary

manipulation would not affect estimates in the different-context condition. First, range assumptions do not play a very important role when participants use enumeration-based strategies. This is because enumeration produces counts that provide a direct, numerical, indication of item frequency. These counts enable participants to produce reasonable estimates independent of their knowledge of the response range or of the target's relative position within the range. Second, different-context participants may use their enumeration-based estimates to draw inferences about the response range (Brown & Siegler, 1993). If so, these participants are likely to adopt similar response ranges, regardless of the boundary information provided to them, and hence are likely to produce similar responses even when they do not enumerate.

In predicting that boundary facts will interact with context (and presentation frequency), it has been assumed that exposure to these facts will not affect strategy selection; regardless of boundary condition, different-context participants should typically enumerate, and same-context participants typically should not. Response times collected in this study provide a means of verifying these expectations. As in Experiment 2, decision times collected in the different-context condition should increase with presentation frequency, and those collected in the same-context condition should be relatively flat. Boundary facts, however, should not affect these tendencies.

Method

Design, materials, and procedure. Data were collected from three groups of different-context participants and three groups of same-context participants. Participants in the Boundary 16 group were informed that no test word appeared more than 16 times, participants in the Boundary 24 group were informed that no test word appeared more than 24 times, and participants in the control group were told nothing about the upper bound of the response range. Participants in the two experimental groups learned about the boundary limits in the instructions to the test phase. Specifically, in the Boundary 16 condition, the opening paragraph of the test instructions included the following sentences: "No category name appeared in the study list more than 16 times. Thus, your estimates should be no larger than 16." This point was reiterated once in the middle of the instructions, and a third time at the end. These sentences were appropriately altered in the Boundary 24 condition and deleted from the instructions given to the control participants.

Other than the modification to the instructions just described and the addition of boundary as a between-subjects variable, the design, materials, procedure, and instructions used in this experiment were identical to those used in Experiment 2.

Participants. One hundred and fifty University of Alberta undergraduates took part in this study, with 25 students randomly assigned to each of the six groups. Participants were individually tested in sessions lasting about 35 min and received course credit for their participation.

Results

A between-subjects variable, boundary, with three levels (Boundary 16, Boundary 24, or control), was included in all ANOVAs. Otherwise, frequency estimates and decision times collected in this experiment were processed and analyzed like those collected in Experiment 2.

Frequency estimates. Frequency estimates indicated that same-context participants were strongly influenced by boundary facts and different-context participants were not. This can be seen in Figure 4, in which estimate means are plotted again presentation frequency. As predicted, in the same-context condition, control participants provided larger estimates and displayed steeper regression slopes than Boundary 24 participants, and Boundary 24 participants provided larger estimates and displayed steeper slopes than the Boundary 16 participants. Specifically, in the same-context condition, mean estimates were 11.0, 7.7, and 5.6, and mean regression slopes were 1.65, 1.09, and .70 for the control group, the Boundary 24 group, and the Boundary 16 group, respectively. Also as predicted, in the different-context condition, there was very little difference across the three groups in either the magnitude of the estimates or the steepness of the regression slopes. Here, the mean estimates were 5.5, 5.2, and 5.2, and the regression slope means were .70, .68, and .65 for the control, Boundary 24, and Boundary 16 groups, respectively. Consistent with the described pattern, an ANOVA performed on the frequency judgments indicated that the Boundary × Context × Presentation Frequency interaction was reliable, $F(10, 720) = 8.5, p < .0001, MS_E = 11.92$, as were all other main effects and interactions. Similarly, for the regression slopes, the Context × Boundary interaction was significant, $F(2, 144) = 12.1, p < .0001, MS_E = 0.21$, as were the main effects.

Table 3 lists the mean rank-order correlations and absolute errors for the six groups of participants. As in Experiment 2, participants in all conditions displayed an accurate understanding of the relative ordering of the test items; the mean rank-order correlation between estimated frequency and actual frequency was .93 in the different-context condition and .90 in the same-context condition, $F(1, 144) = 9.7, p < .01, MS_E = 0.11$. The size of these correlations was unaffected by the presence or nature of a boundary fact; for the main effect of boundary, $F < 1$, and for the Boundary × Context interaction, $F(2, 144) = 1.2, p > .1$. Boundary information did, however, have an impact on absolute accuracy. In particular, exposure to boundary facts reduced absolute error in the same-context condition but not in the different-context condition, $F(1, 144) = 6.1, p < .01, MS_E = 26.85$; see Table 3. It should also be noted that there was a reliable Boundary × Context × Presentation Frequency interaction, $F(10, 720) = 4.4, p < .01, MS_E = 8.33$, indicating that absolute error increased more rapidly with presentation frequency for the same-context control participants than for participants in other groups.

Decision times. Although three latencies (a decision time, an initiation time, and an entry time) were collected on each test trial and were subsequently analyzed, only the decision time data are discussed. However, as in Experiment 2, both initiation time ($M = 0.8$ s) and entry time ($M = 0.7$ s) were brief, initiation times and entry times increased slightly across the range of presentation frequencies, and an ANOVA

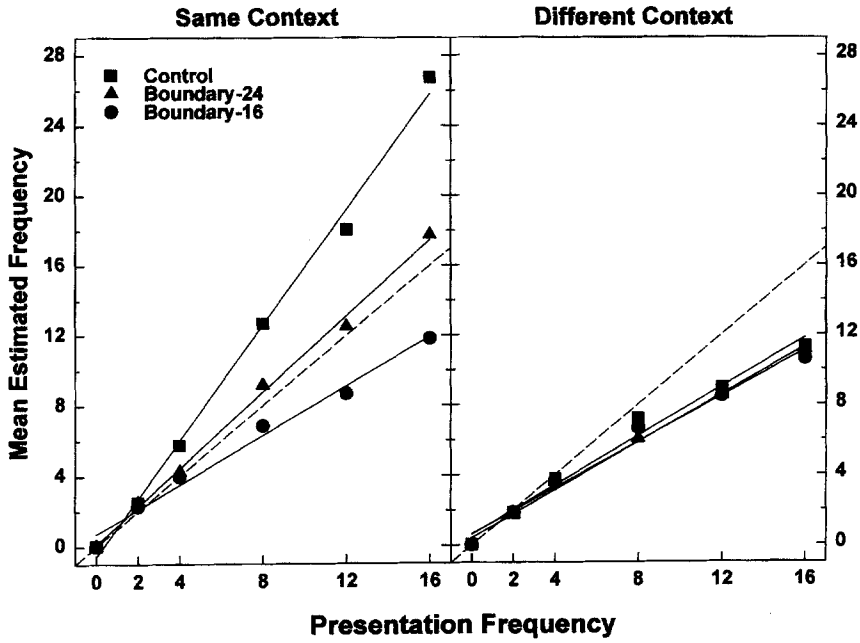

Figure 4. Mean estimated frequencies for control, Boundary 16, and Boundary 24 participants in both the same-context (left panel) and different-context conditions (right panel) in Experiment 3. The solid lines represent the best linear fit for the means, and the dashed lines represent the actual frequencies.

performed on a total time measure (the sum of the three timings) displayed the same pattern of effects as the decision time ANOVA.

Mean decision times for each group of participants are plotted against presentation frequency in Figure 5. As in Experiment 2, different-context participants in the three boundary groups took longer to estimate frequencies for categories that had been presented many times than to those that had only been presented a few times. In contrast, same-context participants in the different boundary conditions responded quite rapidly to all items, regardless of their presentation frequencies. Specifically, in the different-context condition, collapsing over boundary condition, mean decision times for Presentation Frequencies 0, 2, 4, 8, 12, and 16 were 2.5 s, 3.0 s, 4.5 s, 6.5 s, 7.2 s, and 8.3 s, respectively. In the same-context condition, the comparable means were 2.0 s, 2.8 s, 3.0 s, 3.5 s, 3.5 s, and 3.5 s. This reliable Context × Presentation Frequency interaction, $F(5, 720) = 28.1, p < .0001, MS_E = 4.45$, replicated the one obtained in Experiment 2, as did the significant main effects of context, $F(1, 144) = 37.5, p < .0001$, $MS_E = 31.76$, and presentation frequency, $F(5, 720) = 71.6$, $p < .0001$, $MS_E = 4.45$. Boundary facts appeared to have no

Table 3
Mean Rank-Order Correlation Between Estimated and Actual Frequency and Mean Absolute Error for Different- and Same-Context Conditions From Experiment 3

	Rank-order correlation			Absolute error		
Boundary	Different	Same	*M*	Different	Same	*M*
16	.94	.89	.92	2.4	2.3	2.4
24	.93	.90	.92	2.6	3.1	2.9
Control	.92	.91	.92	3.0	5.7	4.4
M	.93	.90	.92	2.7	3.7	3.2

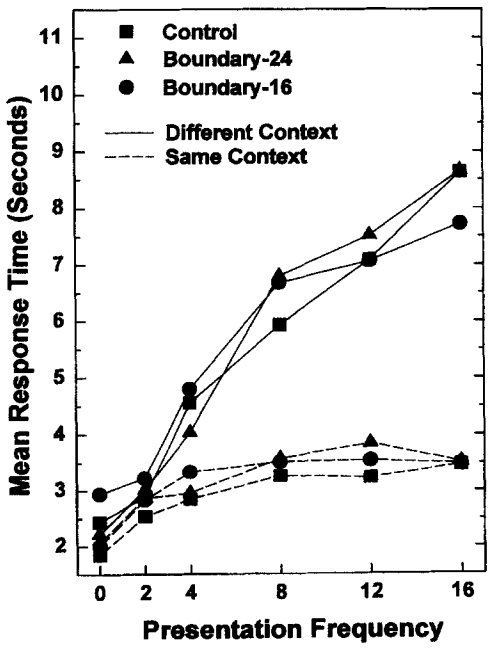

Figure 5. Mean response times for control, Boundary 16, and Boundary 24 participants in both the same-context and different-context conditions in Experiment 3.

effect on response times in either the same- or different-context conditions; the F value for the Boundary × Context × Presentation Frequency interaction was less than one, as were the F values for the main effect of Boundary, and for the Boundary × Context and Boundary × Presentation Frequency interactions.

Discussion

As predicted, exposure to boundary facts affected the frequency judgments of same-context participants, but not those of different-context participants. In the same-context condition, control participants overestimated event frequencies, Boundary 16 participants underestimated them, and estimates provided by Boundary 24 participants fell in-between. In contrast, in the different-context condition, participants underestimated event frequencies to the same extent, regardless of what they were told about the response range. Also as predicted, decision times were unaffected by the boundary manipulation; as in Experiment 2, response times increased steeply with presentation frequency in the different-context condition but not in the same-context condition.

These results are consistent with earlier ones indicating that different-context participants tend to enumerate and that same-context participants rely on nonenumeration strategies. In other words, different-context participants often estimate event frequencies by retrieving and counting relevant instances, and same-context participants perform the same task by assessing the relative frequency of the target event and converting it from a relative value to a numerical one. Experiment 3 provided evidence that these assessment and conversion processes operate independently of one another. This can be seen in the way that boundary facts in the same-context condition affected estimated frequencies (and consequently, the absolute errors and the regression slopes) but not the rank-order correlations between estimated and actual frequency. Apparently, the assessment process assigned the same ordering to the target events in all conditions, and the conversion process selected numerical responses so that they maintained the relative ordering of items and more or less spanned a response range defined by the boundary facts.

The assessment-and-conversion approach to quantitative estimation implied by the current set of results is widely applicable. Indeed, similar frameworks have been proposed to account for performance on tasks ranging from psychophysical and social judgments (e.g., Anderson, 1982; Stevens, 1975) to real-world estimation (Brown & Siegler, 1993). Of course, as the current set of experiments has demonstrated, participants do not always take an assessment-and-conversion approach to quantitative estimation; they may count instances, they may use multiple numerical and nonnumerical reference points, or they may decompose a problem into numerically tractable subproblems. It seems that the crucial factor in determining whether participants depend on an assessment-and-conversion approach is the presence or absence of relevant, credible, numerical information; participants must rely on assessment and conversion when numerical information is absent but can use other types of estimation strategies (e.g., enumeration, reconstruction, decomposition) when it is present.

General Discussion

This research has produced four main findings. First, verbal protocols collected in Experiment 1 indicated that different-context participants often retrieved and counted category exemplars, and same-context participants did not. Second, in Experiments 2 and 3, response times increased steeply with presentation frequency in the different-context condition but not in the same-context condition. Third, in Experiments 1 and 2, and in the control conditions in Experiment 3, different-context participants tended to underestimate event frequencies, and same-context participants tended to overestimate them. Fourth, in Experiment 3, boundary facts affected the magnitude of estimates produced by same-context participants, but they did not affect the magnitudes of those produced by different-context participants, nor did they affect rank-order correlations in either condition.

These findings provide grounds for the following conclusions. (a) Participants use multiple strategies to estimate event frequencies. The protocols indicated that participants use simple enumeration, enumeration and extrapolation, and nonenumeration strategies. There is also a suggestion that qualitative direct retrieval is sometimes used and that memory assessment strategies are quite common. (b) Strategy selection is related to event properties. The protocols and response times indicated that participants favor enumeration-based strategies when event instances are distinctive and rely on nonenumeration strategies when they are not. (c) Strategy selection can affect the magnitude of participants' frequency judgments. In particular, underestimation appears to be a necessary consequence of enumeration, especially when presentation frequency is high. (d) Strategy selection determines whether a separate conversion stage is required and hence whether range information will affect participants' frequency estimates. In general, conversion is necessary when strategies deliver nonnumerical information and unnecessary when they produce a numerical output.

These conclusions will have to be incorporated into a complete theory of frequency estimation, but they do not in themselves constituent such a theory. In addition to recognizing that participants used multiple strategies and that strategy selection is related to event properties, it will be necessary to understand how and why participants choose the strategies that they do, to describe more precisely the nature of the information participants use when they select and execute a given strategy, to identify specific event properties that predict the encoding of various types of frequency-relevant information, and to determine when participants use more than one kind of frequency-relevant information to produce an estimate and how these different types of information are weighted.

It seems likely that the methods described above can be extended to address some of these issues. For example, results from the preceding experiments indicate that participants often enumerate when target items are paired with a different context item on each presentation. This finding can be interpreted broadly as indicating that people enumerate when event instances are unique. According to this view, unique events produce distinctive memory traces, which in turn foster enumeration because they can be readily retrieved and easily

distinguished from one another. These results can also be interpreted more narrowly. It might be that different-context participants were able to enumerate not because the stimulus events were unique per se but because the context items were always typical exemplars of the category identified by the target words. The categorical relation between target and context may have facilitated the encoding, retrieval, and reconstruction of relevant event instances (Bower, Clark, Lesgold, & Winzenz, 1969) and thus promoted enumeration. The narrow interpretation predicts that different-context participants who study target words in the context of random nouns (i.e., *CITY–tree, CITY–cup, CITY–tape,* etc.) may not be able to enumerate, whereas the broad interpretation predicts that they will. If the broad interpretation is correct, then response times should increase with presentation frequency, event frequencies should be underestimated, and estimates should be unaffected by range information. In contrast, if the narrow interpretation is correct, response times should not increase with presentation frequency, and frequency estimates should be affected by the presence and nature of range information.

The point of this example is to demonstrate that response times, estimation biases, and range effects can be used in conjunction to refine our understanding of how and when participants use different estimation strategies. Specifically, situations that promote enumeration will yield one pattern of performance (i.e., a steep response time function, underestimation, and no range effects), and those that hamper it will yield a very different pattern (i.e., a flat response time function and large range effects). Thus, it should be possible to manipulate a variety of factors that may influence the way that events are encoded, stored, and retrieved from memory (e.g., presentation time, test delay, list length, target–context relatedness, response deadlines, motivation, elaboration, etc.) and determine how these factors affect the way that participants generate their frequency judgments. These findings, in turn, should help us better understand ways that people represent event frequency in memory.

References

Aaronson, D., & Ferres, S. (1986). Reading strategies for children and adults: A quantitative model. *Psychological Review, 93,* 89–112.

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

Anderson, N. H. (1982). *Methods of integration theory.* New York: Academic Press.

Barsalou, L. W., & Ross, B. H. (1986). The role of automatic and strategic processing in the sensitivity of superordinate and property frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 116–134.

Battig, W. F., & 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).

Begg, I., Maxwell, D., Mitterer, J. O., & Harris, G. (1986). Estimates of frequency: Attribute or attribution? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 496–508.

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

Bousfield, W. A., & Sedgewick, C. H. (1944). An analysis of sequences of restricted associative responses. *Journal of General Psychology, 30,* 149–165.

Bower, G. H., Clark, M. C., Lesgold, A. M., & Winzenz, D. (1969). Hierarchical retrieval schemes in recall of categorical word lists. *Journal of Verbal Learning and Verbal Behavior, 17,* 573–587.

Brooks, J. E. (1985). Judgments of category frequency. *American Journal of Psychology, 98,* 363–372.

Brown, N. R. (1990). The organization of public events in long-term memory. *Journal of Experimental Psychology: General, 119,* 297–314.

Brown, N. R., Rips, L. J., & Shevell, S. K. (1985). The subjective dates of natural events in very long-term memory. *Cognitive Psychology, 17,* 139–177.

Brown, N. R., & Siegler, R. S. (1993). Metrics and mappings: A framework for understanding real-world quantitative estimation. *Psychological Review, 100,* 511–534.

Bruce, D., Hockley, W. E., & Craik, F. I. M. (1991). Availability and category-frequency estimation. *Memory & Cognition, 19,* 301–312.

Bruce, D., & Van Pelt, M. (1989). Memories of a bicycle tour. *Applied Cognitive Psychology, 3,* 137–156.

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. (1994). Strategies for estimating category frequency: Effects of abstractness and distinctiveness. In *American Statistical Association: Proceedings of the section on survey methods research.* (pp. 1345–1350). Alexandria, VA: American Statistical Association.

Conrad, F. G., Brown, N. R., & Cashman, E. (1993). How the memorability of events affects frequency judgments. In *American Statistical Association: Proceedings of the section on survey methods research* (Vol. 2, pp. 1058–1063). Alexandria, VA: American Statistical Association.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press.

Flexser, A. J., & Bower, G. H. (1975). How frequency affects recency judgments: A model for recency discrimination. *Journal of Experimental Psychology, 103,* 706–716.

Greene, R. L. (1989). On the relationship between categorical frequency estimation and cued recall. *Memory & Cognition, 17,* 235–239.

Gruenewald, J. P., & Lockhead, G. R. (1980). The free recall of category examples. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 225–240.

Hanson, C., & Hirst, W. (1988). Frequency encoding of token and type information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 289–297.

Hartley, A. A. (1977). Mental measurement in the magnitude estimation of length. *Journal of Experimental Psychology: Human Perception and Performance, 3,* 622–628.

Hartley, A. A. (1981). Mental measurement of line length: The role of the standard. *Journal of Experimental Psychology: Human Perception and Performance, 7,* 309–317.

Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General, 108,* 356–388.

Hasher, L., & Zacks, R. T. (1984). Automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologist, 39,* 1372–1388.

Hintzman, D. L. (1969). Apparent frequency as a function of frequency and the spacing of information. *Journal of Experimental Psychology, 80,* 139–145.

Hintzman, D. L. (1976). Repetition and memory. In G. H. Bower

(Ed.), *The psychology of learning and motivation* (Vol. 10, pp. 47–91). New York: Academic Press.

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

Hintzman, D. L., & Gold, E. (1983). A congruity effect in the discrimination of presentation frequency: Some data and a model. *Bulletin of the Psychonomic Society, 21,* 11–14.

Hintzman, D. L., Grandy, C. A., & Gold, E. (1981). Memory for frequency: A comparison of two multiple-trace theories. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 231–240.

Hintzman, D. L., & Stern, L. D. (1978). Contextual variability and memory for frequency. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 439–549.

Hockley, W. E. (1984). Retrieval of item frequency information in a continuous memory task. *Memory & Cognition, 12,* 229–242.

Howell, W. C. (1973). Representation of frequency in memory. *Psychological Bulletin, 80,* 44–53.

Indow, T., & Togano, K. (1970). On retrieving sequence from long-term memory. *Psychological Review, 77,* 317–331.

Johnson, M. K., Raye, C. L., Wang, A. Y., & Taylor, T. H. (1979). Fact and fantasy: The role of accuracy and variability in confusing imaginations with perceptual experiences. *Journal of Experimental Psychology: Human Learning and Memory, 5,* 220–240.

Jones, C. M., & Heit, E. (1993). An evaluation of the total similarity principle: Effects of similarity on frequency judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 799–812.

Jonides, J., & Jones, C. M. (1992). Direct coding for frequency of occurrence. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 368–378.

Jonides, J., & Naveh-Benjamin, M. (1987). Estimating frequency of occurrence. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 230–240.

Just, M. A., & Carpenter, P. A. (1985). Cognitive coordinate systems: Accounts of mental rotation and individual differences in spatial abilities. *Psychological Review, 92,* 137–173.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87,* 252–271.

Marx, M. H. (1985). Retrospective reports on frequency judgments. *Bulletin of the Psychonomic Society, 23,* 309–310.

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.

Menon, G., Raghubir, P., & Schwarz, N. (1993). *Rate-of-occurrence and response alternatives: Sources of information for frequency judgments* (Tech. Rep. No. MARK-93-4). New York: New York University, Leonard N. Stern School of Business.

Morton, J. (1968). Repeated items and decay in memory. *Psychonomic Science, 10,* 219–220.

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

Nisbett, R. E., & Wilson, T. D. (1977). Tell more than we can know. *Psychological Review, 84,* 231–259.

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 54–65.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1992). Behavioral decision research: A constructive processing approach. *Annual Review of Psychology, 43,* 87–131.

Poulton, E. C. (1982). Biases in quantitative judgments. *Applied Ergonomics, 13,* 31–42.

Reder, L. M. (1987). Strategic selection in question answering. *Cognitive Psychology, 19,* 90–137.

Rose, R. J. (1980). Encoding variability, levels of processing, and the effects of spacing of repetitions upon judgments of frequency. *Memory & Cognition, 8,* 84–93.

Rowe, E. J. (1973). Frequency judgments and recognition of homonyms. *Journal of Verbal Learning and Verbal Behavior, 12,* 440–447.

Rowe, E. J., & Rose, R. J. (1977). Effects of orienting task, spacing of repetitions, and list context on judgments of frequency. *Memory & Cognition, 5,* 505–512.

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

Schmidt, R. (1978). Frequency estimation in verbal learning. *Acta Psychologica, 42,* 39–58.

Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General, 116,* 250–264.

Simon, H. A., & Reed, S. K. (1976). Modeling strategy shifts in problem solving tasks. *Cognitive Psychology, 8,* 86–97.

Smith, A., Hager, D., Palphreyman, A., & Jobe, J. (1992). Interindividual calibration of frequency estimates. In *American Statistical Association: Proceedings of the section on survey methods research.* Alexandria, VA: American Statistical Association.

Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual, neural, and social prospects.* New York: Wiley.

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

Tversky, A., & Kahneman, D. (1974). Judgments under uncertainty: Heuristics and biases. *Science, 185,* 453–458.

Underwood, B. J. (1969). Attributes of memory. *Psychological Review, 76,* 559–573.

Voss, J. F., Vereb, C., & Bisanz, G. (1975). Stimulus frequency judgments and latency of stimulus frequency judgments as a function of constant and variable response conditions. *Journal of Experimental Psychology: Human Learning and Memory, 3,* 337–350.

Watkins, M. J., & LeCompte, D. C. (1991). Inadequacy of recall as a basis for frequency knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 1161–1176.

Williams, K. W., & Durso, F. T. (1986). Judging category frequency: Automaticity or availability? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 387–396.

Wilson, T. D. (1994). The proper protocol: Validity and completeness of verbal reports. *Psychological Science, 5,* 249–252.

Wright, D. B., Gaskell, G. D., & O'Muircheartaigh, C. A. (1994). How much is 'Quite a bit'? Mapping between numerical values and vague quantiters. *Applied Cognitive Psychology, 8,* 479–498.

Received August 25, 1994
Revision received December 22, 1994
Accepted January 5, 1995 ■