
Volume 58, Number 1, January 2008
ISSN 0749-596X
ELSEVIER
Journal of
Memory and
Language
Editor
Kathryn Bock
Associate Editors
Robert Greene
Evan Heit
James McQueen
David Plaut
Editorial Board
Gerry T.M. Altmann
Sally Andrews
Julie E. Boland
Gordon Brown
Charles Clifton, Jr.
Cynthia M. Connine
Gary S. Dell
Kara D. Federmeier
Fernanda Ferreira
Victor Ferreira
Cynthia Fisher
Kenneth I. Forster
Carol A. Fowler
David A. Gallo
Alan Garnham
Simon Garrod
Richard J. Gerrig
Zenzi M. Griffin
Robert J. Hartsuiker
Larry L. Jacoby
Colleen Kelley
Judith F. Kroll
Wido La Heij
Stephen J. Lupker
Michael E.J. Masson
Brian McElree
Antje S. Meyer
Akira Miyake
Gregory L. Murphy
James S. Nairne
James H. Neely
Martin J. Pickering
Suparna Rajaram
Keith Rayner
Henry L. Roediger III
Arthur G. Samuel
Patrick Sturt
Daniel Swingley
Marcus Taft
Rebecca Treiman
John Trueswell
Jos J.A. van Berkum
Mieke Verfaillie
Andrew Yonelinas
Available online at
ScienceDirect
www.sciencedirect.com

Available online at www.sciencedirect.com
ScienceDirect
Journal of Memory and Language 58 (2008) 3–18
Journal of Memory and Language
www.elsevier.com/locate/jml

How metastrategic considerations influence the selection of frequency estimation strategies ☆

Norman R. Brown *

Department of Psychology, University of Alberta, Edmonton, AB., Canada T6G 2E9

Received 23 May 2006; revision received 7 September 2007
Available online 7 November 2007

Abstract

Prior research indicates that enumeration-based frequency estimation strategies become increasingly common as memory for relevant event instances improves and that moderate levels of context memory are associated with moderate rates of enumeration [Brown, N. R. (1995). Estimation strategies and the judgment of event frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 1539–1553; Brown, N. R. (1997). Context memory and the selection of frequency estimation strategies. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 898–914]. Two experiments, one an RT study, the second a protocol study, were conducted to determine what happens to when context memory is very good. Both compared an *exemplar-generation* condition to a *non-generation* condition. As expected, generation improved context memory and reduced estimation bias. Importantly, Experiment-1 RTs increased with frequency at the same rate in both conditions, which implied that generation had no effect on enumeration. Experiment-2 protocols provided converging evidence for this conclusion as enumeration rates were equal across conditions. These findings are consistent with a metastrategic account of strategy selection; on this view, people mix enumerated and non-enumerated responses in a way that optimizes performance at the task level rather than the trial level.

Keywords: Frequency judgments; Multiple strategies; Strategy selection; Metastrategy; Generation effect

Introduction

Judgments of event frequency are solicited by two types of researchers: those interested in understanding how repeated experiences affect memory and performance (Dougherty, Gettys, & Ogden, 1999; Greene, 1989; Hasher & Zacks, 1984; Hintzman, 1988; Howell, 1973; Johnson, Hashtroudi, & Lindsay, 1993; Logan, 1988; Sedlmeier & Betsch, 2002; Watkins & LeCompte, 1991; Williams & Durso, 1986; Zacks & Hasher, 2002) and those who hope to obtain accurate information about the frequency of potentially important, but undocumented, behaviors, and feelings (Blair & Burton, 1987; Bogart et al., in press; Brown & Sinclair, 1999; Brown, Williams, Barker, & Galambos, 2007; Burton & Blair, 1991; Conrad, Brown, & Cashman, 1998; Edwards, Thomsen, & Toroitich-Ruto, 2005; Garry, Sharman, Feldman, Marlatt, & Loftus, 2002; Means & Loftus, 1991; Menon, 1993; Menon, Raghubir, &

☆ This project has been supported by the author's NSERC operating grant. I'd like to thank Amy L. Anderson for her assistance in conducting these experiments and Peter Lee and Bob Latimer for their comments on the manuscript.
* Fax: +1 780 492 1768.
E-mail address: norman.brown@ualberta.ca

0749-596X/$ - see front matter
doi:10.1016/j.jml.2007.09.004

Schwarz, 1995; Robinson & Clore, 2002). Over the past decade, these two very different lines of research have converged on an understanding of frequency processing that has been labeled the *Multiple Strategies Perspective* (Brown, 2002a). The main tenets of this approach are: (a) People represent frequency information in a number of different ways; (b) they use a variety of estimation strategies to generate frequency judgments; (c) event properties and encoding factors are systemically related to the way that frequency information is represented; (d) representation and strategy selection are also related; and (e) strategy selection can affect both effort and accuracy.

Although there is good support for these claims, a number of important and interesting issues remain. This article focuses on one of these, strategy selection. It is obvious that strategy selection is *restricted* by the task-relevant contents of memory. For example, rate-based responses, which are produced by multiplying rate-of-occurrence facts retrieved directly from memory, are ruled out when such facts unavailable (Conrad et al., 1998; Menon, 1993). Likewise, familiarity-based strategies are useless when a probe fails to access a coherent memory representation (Barsalou & Ross, 1986; Conrad, Brown, & Dashen, 2003; Freund & Hasher, 1989; Pandelaere & Hoorens, 2006). What is less obvious is how (and why) people select a strategy when they have access to multiple sources of potentially relevant information. This is a central issue in the strategy-differences literature (LeMaire, 2005; Payne, Bettman, & Johnson, 1993; Rieskamp & Otto, 2006; Siegler & Shipley, 1995) and is of particular concern to psychologists who study frequency processing (Betsch & Sedlmeier, 2002; Brown, 2002a).

This concern is driven in large part by the fact that people often *enumerate* when they estimate event frequencies (Blair & Burton, 1987; Brown, 1995, 1997; Brown & Sinclair, 1999; Bruce & Van Pelt, 1989; Burton & Blair, 1991; Conrad et al., 1998; Conrad et al., 2003; Marx, 1985; Means & Loftus, 1991; Menon, 1993; Menon et al., 1995; Pandelaere & Hoorens, 2006). In other words, they retrieve and count relevant instances and respond with a number that is equal to (*simple enumeration*) or greater than (*enumeration and extrapolation*) the derived count. What is interesting here is that enumeration is slow and often produces biased estimates. Yet when people use enumeration-based strategies, they often use them in preference to other strategies that are considerably easier to execute and about as accurate (Brown, 1995, 1997). On the face of it, this tendency appears suboptimal, if not willfully perverse. After all, why would a person use a difficult strategy when he or she could use an easy one?

In this article, I consider three possible answers to this question, answers that correspond to three different approaches to strategy selection. The first, termed the *research-limits* account, assumes that people are strongly biased to enumerate and that they will use enumeration-based strategies whenever they can. The second account, termed the *cost-benefits* account, also assumes that people have a preference for enumeration, though this preference is tempered by the recognition that enumeration can be demanding and by a willingness to use less effortful strategies when enumeration is deemed impractical. Both these accounts treat strategy selection as a trial level phenomenon—on both views, the decision to enumerate is made on a case by case basis and is conditioned by the availability of task-relevant memory traces (see below). The third account shifts the focus from the trial level to the task level; this *metastrategy* approach holds that there is a metastrategic aspect to strategy selection which, under some conditions, enables people to mix difficult-to-execute strategies (e.g., enumeration) with easy-to-execute strategies, in a way that optimizes performance over trials.

In the next section, I describe these three approaches in greater detail, explain how each accounts for established findings in the frequency literature, and argue that it should be possible to select between them by observing enumeration rates when context memory is very good. The two experiments reported below were conducted to enable this selection. Both used an *exemplar-generation* task to achieve the required level of memory performance, and both provided evidence indicating that metastrategic considerations play an important role in the selection of frequency estimation strategies.

Approaches to strategy selection and their relation to enumeration-based frequency judgments

The resource-limits approach

The resource-limits approach to strategy selection, which is the simplest of the three considered here, is applied when there is a single preferred strategy. The preferred strategy may be optimal in the sense that it is both more accurate and less effortful than its comparators, but it needed not be. It is enough that people execute this strategy whenever they can. When there is a preferred strategy, strategic variability is introduced when requisite declarative or procedural knowledge is lacking or when mnemonic or attentional resources are inadequate.

Many strategy-related phenomena examined in the developmental, aging, expertise, and memory literatures have been linked to resource limitations of one form or another (e.g., Chi, Feltovich, & Glaser, 1981; Dunlosky & Hertzog, 1998; Ericsson & Lehmann, 1996; Siegler, 1987; Siegler & Lemaire, 1997). The development of basic math skills is perhaps the best studied example of resource-limited strategy selection. For one-digit addition problems, the direct retrieval of the answer from memory is the preferred (and optimal) strategy. To use this strategy effectively, the child needs access to a set of strong associations that link the addends and the operand to the solutions. When these associations are weak or absent, children rely on ''back-up'' strategies (e.g., the ''Min'' strategy, which involves counting up from the larger addend the number of times indicated by the smaller addend) to generate their responses. Once the requisite associations are in place, these back-up strategies drop from the child's repertoire and he or she comes to rely on direct retrieval exclusively (Siegler & Jenkins, 1989).

One way to think about enumeration is to assume that people are strongly biased to enumerate (at least in the absence of a prestored tally or rate) and that they will do so whenever they have access to readily retrievable task-relevant event instances. This bias may arise from the belief that enumeration produces more accurate estimates than the alternatives, or it may reflect a strong preference to ground frequency estimates with concrete information. Regardless of its origins, a bias of this sort fits well with a resource-limits account. On this view, enumeration is identified as the preferred strategy and the availability of readily retrievable event traces is the limiting factor. It follows that the use of enumeration should *increase* when conditions foster retention and recall of these traces. Consistent with this prediction, a study examining taxonomic frequency estimates (Brown, 1997) produced a monotonic relationship between context memory (i.e., memory for the words paired with the target items during study) and response time (RT, an index of enumeration). In this study, the RT data indicated that enumeration was rare when context memory was very poor, and that it became increasingly common as memory for event instances improved.[1]

[1] In the standard taxonomic frequency paradigm (Brown, 1995, 1997; Bruce, Hockley, & Craik, 1991; Conrad et al., 2003; Greene, 1989; Watkins & LeCompte, 1991; Williams & Durso, 1986), participants study word pairs consisting of a target word (a category label, e.g., METAL) and a context word (a category exemplar, e.g., gold). Although target words are presented multiple times across the list, each context word is presented only once. Because each word pair is unique, the recall of a particular context word at test is taken is evidence that the studied word pair has been remembered. Throughout this article, I use the term *context memory* in a paradigm-specific manner—it refers to memory for the category instances that were paired with the category labels during study. However, it is also assumed that people treat the unique word pairs of this sort in much the same way as they treat distinctive event instances they experience outside of the lab.

The cost-benefit approach

When there is no preferred strategy (or when the prerequisites for executing the preferred strategy are not met), selection requires a choice between procedures that may differ along at least two relevant dimensions: (perceived) processing difficulty and (perceived) response quality. A cost-benefit approach is applicable to strategy selection when accuracy and effort are (believed to be) positively correlated across the set of candidate strategies, i.e., when more effortful strategies (are thought to) produce more accurate outcomes than less effortful ones. When these conditions are met, it seems likely that the strategy selection involves a decision which takes both costs (effort) and benefits (accuracy) into consideration.

Cost-benefit research has been carried out mainly in the judgment and decision-making area, with the multiple-attribute choice task providing the primary test-bed (e.g., Chu & Spires, 2003; Payne et al., 1993; Russo & Dosher, 1983; Swait & Adamowicz, 2001). This research has demonstrated that stimulus factors (e.g., number of attributes, number of options) and task demands (e.g., accuracy pay-offs, time pressure) jointly and sensibly affect strategy selection. For example, people tend to use simple one-reason decision strategies (e.g., *Take the Best*, Gigerenzer & Goldstein, 1996) when they are under time pressure (Rieskamp & Hoffrage, 1999) or when information acquisition is costly (Bröder, 2000). In contrast, people often consider multiple cues when they are not pressed for time or when they have ready access to task-relevant information (Newell & Shanks, 2003; Newell, Weston, & Shanks, 2003).

Like the resource-limit position, a cost-benefit treatment of strategy selection for frequency estimation assumes that people prefer enumerated estimates. However, it is also assumed that this preference is not absolute. Instead, on this view, people choose to enumerate when they believe they can retrieve and count and a representative set of task-relevant instances in a reasonable amount of time, and they select an alternative strategy when they decide that the generation of an accurate enumerated estimate may require time or effort.

It seems likely that people are aware that the effort required for a response increases with the number of instances that need to be retrieved and counted. If so, the cost-benefit approach predicts that the relation between event memory and enumeration rate should be ∩-shaped. In other words, the tendency to enumerate should increase as event memory improves across conditions, but only up to a point, after which it should decline (Keller & Staelin, 1987). In addition, when actual frequency is manipulated, but event memory is very good, enumerated estimates may be less common for high-frequency items than low-frequency ones. Both predictions assume that people will shy away from enumeration when they are confronted with too many potentially recoverable traces.

The metastrategic approach

The resource-limits and cost-benefit approaches focus on strategy selection at the trial level. In contrast, the metastrategic approach contends that task-level factors can also affect strategy selection. This is because difficult-to-execute strategies can sometimes confer benefits that extend beyond a single trial and because between-trial contingencies often exist which can make easy-to-execute strategies more accurate than they would be if performed in isolation. These considerations raise the possibility that there may be an optimal mix of strategies and that people prefer this mix when conditions allow.

In the present context, this position suggests that participants may mix enumeration-based responses with non-enumerated ones, even when context memory is good enough to support enumeration on every trial. A *strategies-mix* metastrategy of this sort might work well here because the common estimation strategies have complementary strengths and weaknesses. On the one hand, enumeration-based strategies produce estimates that are based on credible numerical information, and they provide a non-arbitrary basis for defining and partitioning the response range. But, they also require a good deal of time and effort to execute. In contrast, non-enumeration strategies are easy to execute, but they do not deliver concrete information. Instead, these strategies depend on non-numerical information (i.e., retrieved vague quantifiers, a sense of familiarity, similarity-based echo strength, etc; Brown, 2002a) and require an uncertain mapping between this information and the response range. On this view, participants take the time to enumerate, in part, because it helps them ground their range assumptions, in part, because they prefer concrete information to vague intuitions, and in part, because enumeration mitigates the problems associated with an exclusive reliance on non-enumeration strategies (e.g., a lack of concreteness, response range uncertainty).

When a strategies-mix approach is taken, people should be able to map an ordinal judgment on to a response scale that has already been defined by and calibrated to a set of enumerated estimates. Also, they may sometimes "reuse" prior enumeration-based responses when available non-numerical information indicates the current target item occurred about as often as a previous one. More concretely, consider a participant who has recently provided an enumeration-based estimate for a frequently presented target word and who has just been presented with another target that he or she believes to have also been presented "many times." In this case, it might make sense to state this prior response rather than to attempt to enumerate all available context words. Alternatively, this individual may use a standard memory assessment strategy (Brown, 2002a) and select a value from the appropriate portion of the enumeratively-grounded response range. Either way, these non-enumeration strategies are fast and easy to execute, and when used along with enumeration, produce responses that are ordinally consistent with and comparable in magnitude to enumeration-based estimates. Thus, although it is true that enumeration increases with context recall (Brown, 1997), it might also be true that this relation holds only up to a point. Beyond this point, participants may continue to mix enumeration and non-enumeration strategies in a way produces a well-grounded, credible set of frequency estimates in an efficient manner.

Experiment 1

Prior research indicates enumeration is very uncommon when context memory is poor, and that moderate levels of context memory (e.g., 40%, Brown, 1997; Experiment 1) are associated with moderate rates of enumeration (e.g., 57%, Brown, 1995; Experiment 1). These findings leave open the question of how enumeration rates might be affected if context memory were very good. For reasons stated above, the resource-limits approach predicts that enumeration rates will increase; the cost-benefits approach predicts that they will decrease; and the metastrategic approach predicts that improved context memory will have little if any effect on enumeration rate.

The challenge in assessing these competing predictions was to find a manipulation that produced a marked increase in context memory relative to an established baseline condition and that left the timed frequency judgment paradigm intact (Brown, 1995, 1997; Conrad et al., 2003). *The exemplar-generation* task provided a way of meeting this challenge. Both this experiment and the next included an *exemplar-generation* condition and a *non-generation* condition. During the study phase, participants in the generation group were required to produce unique category exemplars (e.g., Yellowknife, Moose Jaw, Red Deer...) in response to repeatedly presented category labels (e.g., CITY). Participants in the non-generation group saw the same set of multiply presented category labels. However, as in prior studies (Brown, 1995, 1997), each label was paired with an experimenter-provided category exemplar (e.g., New York, Paris, London...). Following the study phase, all participants were presented with each category label and were timed as they estimated how often it had appeared during the study phase; then during a final phase, participants attempted to recall all context words presented with or generated for each target word.

There is an extensive literature documenting the existence of the *generation effect*, which refers to the tendency for people to have a better memory for self-generated material than for experimenter-selected material (Mulligan & Lozito, 2004; Slamecka & Graf, 1978).

It follows that memory for context words should be better in the generation condition than in the non-generation condition. If this prediction is correct, then RT and frequency estimates can be compared across conditions to determine the possible effects of improved context memory on strategy selection and estimation performance.

In this experiment, both study-phase generation and study-phase verbalization were manipulated between groups. During the study phase, half of the non-generation participants read each of the word pairs aloud, and half studied the word pairs in silence. Similarly, half of the participants in the exemplar-generation condition read each target word out loud and then stated an appropriate category instance; the other half generated their responses silently. Study-phase verbalization was introduced for three reasons. First, it made it possible to determine whether participants in the generation condition were able to perform the task assigned to them. Second, a record of the generated words was necessary to assess the accuracy of responses in the cued recall test. Third, the verbalization manipulation made it possible to determine whether this requirement influenced frequency judgments and performance on the cued-recall test.

The non-generation conditions were included in this experiment to provide baseline data. It was expected that these data would replicate those reported in prior studies that made use of different-context lists like the one used here (Brown, 1995, 1997; Conrad et al., 2003). These studies have consistently demonstrated that such lists produce: (a) moderate memory for the context words, (b) a RT function that increases steeply with actual event frequencies, and (c) a strong tendency to underestimate event frequencies. The latter two findings reflect to use enumeration-based estimation strategies. In this paradigm, RT increases with presentation frequency because enumeration is a serial process. Assuming that the time to retrieve an instance is constant (or increases with each additional instance, Bousfield & Sedgewick, 1944; Indow & Togano, 1970) and that participants retrieve more instances when responding to high-frequency items than to low-frequency items, it follows that RT should increase with presentation frequency. Enumeration is associated with underestimation because instances are sometimes permanently forgotten, because output interferences renders others temporally inaccessible, and because participants may choose to terminate their memory search before all relevant instances have been retrieved.

The three strategy-selection approaches outlined above make unique predictions regarding possible between-group differences in RTs. These predictions assume: that context memory will be considerably better in the generation condition than the non-generation condition; that the non-generation groups will produce the usual pattern of performance; and that the steepness of the RT function provides a valid index of the enumeration rate (Brown, 1997). Given these assumptions, the resource-limits approach predicts that the RT function should be steeper (and mean RT higher) in the generation condition than in the non-generation condition; this would indicate that enumeration-based strategies were used more often when context memory was very good than when it was not. The cost-benefit approach makes the opposite prediction—the RT function should be shallower (and mean RT lower) in the generation condition, indicating a decreased reliance on enumeration. Finally, it could be that the RT functions (and RT means) are the same in the two conditions; this would indicate that the enumeration rate was unaffected by the increase in context memory—a result consistent with the metastrategic account.

The three strategy-selection approaches do not make unique predictions concerning the estimation biases. However, it did seem likely that participants in the generation group would produce larger, less biased estimates than those produced by participants in the non-generation group. This would indicate that the memory advantage conferred by study-phase generation enabled participants in the generation condition to retrieve and count a larger percentage of relevant event instances when they enumerate. A mitigation of the underestimation bias would be expected regardless of how increased context memory affects enumeration rates, provided generation participants enumerated some of the time. This is because enumeration-based estimates play an important role in defining the response range, even when enumeration is not used on every trial (Brown, 1995; Menon et al., 1995).

Methods

Materials

In the non-generation condition, each participant was presented with a list of 260 word pairs. Each pair consisted of a target word that identified a taxonomic category and a context word that identified a member of that category. Published category norms were used to create the stimulus pairs (Battig & Montague, 1969; McEvoy & Nelson, 1982). Categories selected from these norms met two criteria: First, each category could be described by a single noun (e.g., WOMAN, COUNTRY, METAL); second, each category had a reasonable number of frequently listed category members. These one-word category names served as target items, and frequently listed one- and two-syllable category members as context items. Each category was assigned to a single level of presentation frequency depending on the availability of suitable context items.

The first four pairs and the last four pairs on the study list served as filler items. Words used in the filler pairs were presented only once and the target words were not included in the frequency and cued-recall tests. Target words that appeared in the body of the list were presented 2, 4, 8, 12, or 16 times. Six target words were assigned to each level of presentation frequency. Because each target word was paired with a different context word on each presentation, participants were exposed to 252 unique word pairs in the body of the study list.

A unique study list was created for each participant in the non-generation condition. These lists were constructed so that repeated presentations of a given target word were evenly distributed across the list, and so that target items repeated at a given level of presentation frequency were not over- or under-represented in any portion of the list. Each study list also had the target items arranged in a unique random order, and each used a different random ordering of the context items (for details, see Brown, 1995).

Both the frequency judgment test and the cued-recall test contained 36 items. Thirty of these were the multiply presented target items, and six were category labels that did not appear in the study list. The latter served as zero-frequency catch trials. A different test frequency judgment list was constructed for each non-generation participant. To do this, the list was first divided into six blocks, with one target item from each frequency level (0, 2, 4, 8, 12, 16) randomly assigned to each block. Then, the target words were randomly ordered within blocks. The same scheme was used to create the test booklets for the cued-recall test. Each participant in the non-generation condition was yoked to one in the generation condition. The order used to present the target words during study and at test was the same for both participants in each yoked pair.

Procedure

Prior to the presentation of the study list, all participants were informed that they would see a list consisting of 260 items and that they should study these items for a later memory test. In addition, participants in the generation groups were told that each of the presented items would be a category label, that most labels would appear more than once, and that they were required to think of a different example of the category each time a given label appeared on the list. This requirement was emphasized several times in the instructions. In addition, participants assigned to the verbalized generation condition were instructed to read the category label out loud and then state their example; those in the verbalized non-generation condition were instructed to read the word pairs aloud as they were presented.

During the study phase, the target-context pairs (non-generation condition) or the target words were displayed one at a time, for 5.5 s on a computer-controlled video monitor. Regardless of condition, the target item always appeared in the center of the screen in upper-case letters, and in the non-generation condition, the context item always appeared two lines beneath in lower-case letters. After the 5.5 s study interval, the screen was erased and remained blank (except for markers indicating the screen positions of the target and context word, and a trial counter) for 0.5 s.

The frequency task followed the study phase. Participants were informed that they would be presented with 36 target words, one at a time and that they would have to estimate as accurately as possible the number of times each had appeared in the previous list. They were also informed that decision times would be collected, and they were warned that some test items did not appear during the study phase. This set the lower bound of the response range to zero; the upper bound of the range was unspecified.

Participants pressed the ENTER key on the computer keyboard to initiate each trial. This caused a target word to appear at the center of the computer display. After the word appeared, the participant decided how many times it had occurred during the study list. Participants were instructed to press the keyboard's SPACE BAR as soon they had "a single numerical response in mind." Pressing the SPACE BAR caused a response field to appear. Participants then entered their estimates on the keyboard, and pressed the ENTER key to indicate that they had completed the trial. Pressing the ENTER key caused the test word, frequency estimate, and response field to be erased and to be replaced by a message prompting the participant to initiate the next trial.

Each trial yielded a *decision time*, an *initiation time*, and an *entry time*. The decision time interval began with the presentation of the test word and ended when the participant pressed the SPACE BAR. This interval was particularly important because it indicated the time required to formulate an estimate. The initiation interval began as soon as the SPACE BAR was pressed and ended as soon as the first digit was typed, and the entry interval began when with the termination of the initiation interval and ended when the ENTER key was pressed. In combination, initiation and entry time indicated the time required to enter an estimate once it had been formulated.

The first six trials were treated as practice. During these trials, the experimenter sat with the participant and made sure he or she understood the task and the test procedure.

A cued-recall test followed the frequency test. Each participant was given a booklet that listed the 36 target words, 12 to a page. Participants were reminded that most of the target words had been paired with multiple context words during the study phase, and they were instructed to recall as many of these context words as they could. Specifically, participants were asked to work through the booklet one target word at a time at their own pace, and to respond by listing all the context words that were paired with it during the study phase.

Participants

One-hundred twenty undergraduates, recruited from the University of Alberta subject pool, took part in this experiment. Thirty of these were randomly assigned to each group. Participants were tested individually in sessions lasting about 45 min, and received course credit for their cooperation. Data from one participant in the verbalized generation group were discarded because his cued-recall test included 73 intrusion errors; this was taken as an indication that this individual did not understand the task requirements.

Results

A set of preliminary analyses indicated that the verbalization manipulation had no effect on cued recall performance, estimated frequency, or RT. Thus, unless otherwise noted, the analyses reported below were collapsed over verbalization and treated generation as the only between-group variable.

Exemplar-generation

Before turning to the effects of the generation on frequency judgments and cued recall, it is necessary to determine whether participants in the generation groups were able to produce unique category instances during the study phase. The study-phase protocols obtained from the generation participants in the verbalization condition provide data relevant to this issue. On average, these people produced 214.6 responses to the 252 target words, a success rate of 85%. On occasion, a participant produced the same context word more than once. As a result, the average number of unique words (202.2) was slightly lower than the total number of words produced. Thus, participants in the verbalized generation condition produced a unique response for 80% of the target words.

Cued-recall

Responses recorded during the cued-recall test were used to compute a mean *total recall* measure for each participant and each level of presentation frequency. In the non-generation condition, the counts that went into this measure included both correctly recalled instances and intrusion errors; in the generation condition, all items listed on the cued-recall test were included. It was necessary to rely on total recall because it was not possible to determine whether the items recalled by silent-generation participants were the ones that they produced during the study phase. Intrusion errors, however, were flagged in the other three conditions and proved to be uncommon: Only 1.7% of the words recalled by participants in the verbalized generation group were intrusions. Comparable figures for the verbalized non-generation group and the silent non-generation were 1.5% and 1.2%, respectively.

The total recall means were submitted to a Group (generation vs non-generation) × Frequency (0, 2, 4, 8, 12, 16) Anova. This analysis yielded main effects of group, ($F(1,117) = 106.69$, $MSE = 4.28$, $p < .0001$, and frequency, $F(5,585) = 1133.57$, $MSE = 0.86$, $p < .0001$), and a reliable Group × Frequency interaction $F(5, 585) = 63.98$, $p < .0001$. The group means are presented in Table 1 and the interaction in Fig. 1 (left panel). It is clear from these data that participants in the generation condition recalled more items (mean = 167.47 words) than those in the non-generation condition (mean = 109.78 words). In other words, in this experiment, participants recalled 83% of the unique instances they generated, but only 52% of the instances they read—this difference represents the predicted exemplar-generation effect.

Response times

The cued-recall data indicated that participants who generated their own category exemplars recalled over 50% more context words than those who studied experimenter-selected exemplars. Having obtained this exem-

Table 1
Mean number of words recalled, estimated frequencies, absolute errors, slopes, correlations, and RTs (in seconds) for Experiments 1 and 2

Encoding task	Words recalled	Estimated frequency	Absolute error	Slope[a]	Correlation[b]	RT
Experiment 1						
Generation	167.48 (3.41)	8.01 (0.47)	3.24 (0.37)	1.08 (0.08)	0.93 (0.01)	5.84 (0.56)
Non-generation	109.78 (4.40)	4.85 (0.20)	2.86 (0.11)	0.62 (0.03)	0.92 (0.01)	5.90 (0.49)
Experiment 2						
Generation	171.05 (5.50)	6.56 (0.31)	1.90 (0.15)	0.88 (0.06)	0.95 (0.01)	—
Non-generation	101.10 (8.36)	5.00 (0.58)	3.44 (0.22)	0.66 (0.11)	0.83 (0.01)	—

The standard errors for these means are presented in parenthesis.
[a] Estimated against actual frequency.
[b] Rank-order correlations between estimated and actual frequency.

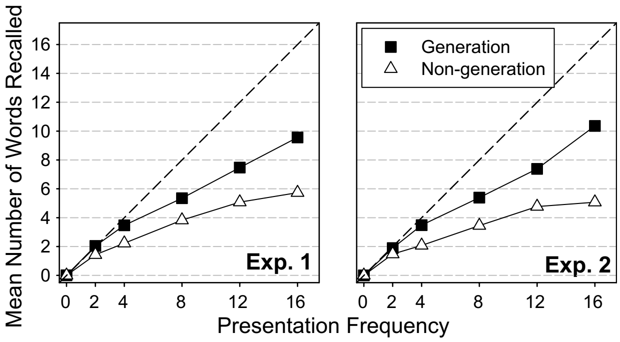

Fig. 1. Mean number of words recalled as a function of presentation frequency in Experiments 1 and 2.

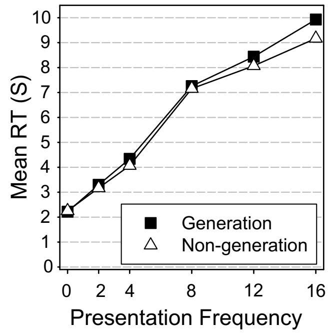

Fig. 2. Mean response time (RT) as function of presentation frequency in Experiments 1.

plar-generation effect, it is possible to test the prediction that very good context memory and enumeration are related. Again, the resource-limits account predicts that the generation condition should yield a steeper RT function than the non-generation condition, the cost-benefits account predicts the RT function should be shallower, and the metastrategic position predicts no difference.

To address these predictions, six means, one for each level of presentation frequency, were computed for each participant. Because the first six trials were treated as a practice block, each mean represented the average of five, rather than six, responses. Although each response was composed of several measurable durations, only the decision time analysis is reported. However, it is worth noting that ANOVAs performed on the decision time measure and on the total time measure yielded a similar pattern of means and produced identical patterns of statistical effects.

RT is presented as a function of presentation frequency and generation condition in Fig. 2. As expected, RT increased steeply with presentation frequency. This is consistent with prior results and indicates that participants in both groups made use of enumeration-based estimation strategies. Consistent with the metastrategic account, RT was unaffected by study-phase generation; mean decision time was same for both groups (see Table 1), and this was true regardless of the level of presentation frequency. These observations were confirmed by the ANOVA, which indicated that only the frequency effect was reliable; $F(5,585) = 78.35$, $MSE = 13.47$, $p < .0001$; Fs < 1.0 for the main effect of group and the Group × Frequency interaction.

Frequency estimates

Several dependent variables were used to assess estimation performance and compare it between groups. These included: (a) estimated frequency, (b) absolute error (i.e., |estimated frequency – actual frequency|), (c) the slope obtained by regressing estimated frequency against actual frequency, and (d) the rank-order correlation between estimated and actual frequency. The estimate means and the regression slope reflect estimation bias; the rank-order correlation indicates how well-participants discriminate between the frequency levels; and absolute error provides a general index of estimation accuracy (Brown, 1995; Flexser & Bower, 1974; Naveh-Benjamin & Jonides, 1986).

Because the first six trials were treated as a practice block, these estimates were discarded. Two sets of six means were computed for each participant from the remaining 30 responses. One set consisted of the estimated frequency means, which were derived by averaging over the estimates provided for the five targets at each level of presentation frequency; the other set was composed of the corresponding absolute error means. These data were submitted to separate Group × Frequency ANOVAs. A separate regression slope was obtained for each participant. This was done by fitting estimates elicited by the 30 target items to their true frequencies. Slopes less than 1.0 indicate a tendency to underestimate event frequencies, and those greater than 1.0 indicate a tendency to overestimate them. Individual rank-order correlations were also computed from each participant's 30 frequency judgments.

In Fig. 3 (left panel), mean estimated frequency is plotted against presentation frequency as function of

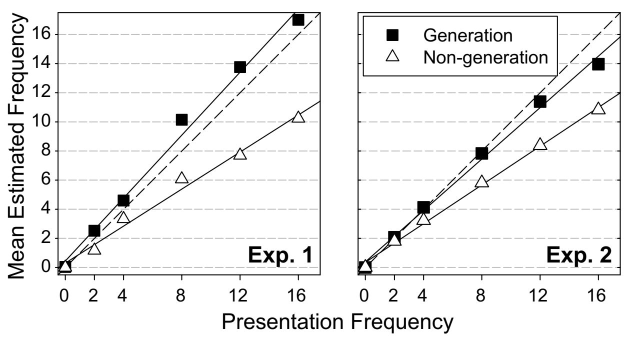

Fig. 3. Mean estimated frequency as function of presentation frequency in Experiments 1 and 2. The solid lines represent the best linear fit for each group's means and the dashed diagonal lines represent the actual frequencies.

generation condition. It is clear from this figure and from the means listed in Table 1 that estimated frequency was affected by the nature of the study phase. As in several prior experiments, participants in the non-generation condition tended to underestimate event frequencies, and this tendency increased with presentation frequency. In contrast, in the generation condition, estimated frequencies were slightly greater than the corresponding presentation frequencies. Consistent with these claims, the regression slopes obtained from participants in the generation group were significantly steeper than those obtained from participants in the non-generation group; $F(1, 117) = 30.74$, $p < .0001$; $MSE = .20$. In addition, separate follow-up *t*-tests indicated that the non-generation slopes differed significantly from 1.0, $t(59) = -11{,}65$, $p < .0001$, but that the generation slopes did not, $t(58) = 1.04$; $p > .1$. The ANOVA performed on the estimate means makes the same point; the Group × Frequency interaction was highly significant, $F(5, 585) = 22.66$, $p < .0001$, $MSE = 10.82$, as were the main effects of group, $F(1, 117) = 38.25$, $p < .0001$, $MSE = 46.71$, and frequency, $F(5, 585) = 304.70$, $p < .0001$, $MSE = 10.82$.

Although the estimates produced by non-generation participants were, on average, more biased those produced by generation participants, both groups were equally good at assigning larger numbers to more frequently presented target words; the mean rank-order correlation between estimated and actual frequency was .93 for the generation condition and .92 for the non-generation condition; $F(1, 117) = 1.00$, $p > .1$, $MSE = .003$. Moreover, the nature of the generation task did not affect absolute error (see Table 1); $F(1, 117) < 1.0$ for the main effect of group; $F(5, 585) = 1.10$, $p > .1$, $MSE = 6.56$ for the Group × Frequency interaction. However, absolute error did increase with presentation frequency, $F(5, 585) = 153.19$, $p < .0001$, $MSE = 6.56$. Collapsing over generation condition, the absolute error means for frequency levels 0, 2, 4, 8, 12, and 16 were.02,.56, 1.36, 3.65, 5.36, and 7.33, respectively.

Discussion

In this experiment, the non-generation condition replicated prior findings (Brown, 1995, 1997). As in the past, participants in this condition demonstrated moderate memory for the category instances; they produced a RT function that increased sharply with presentation frequency; and they tended to underestimate event frequencies. Also, as predicted, generation participants were much better at recalling context words than were non-generation participants, and they produced less biased estimates. RTs also increased with presentation frequency in this condition, but there was no difference between the RT function obtained from the generation group and the one obtained from the non-generation group.[2]

The increasing RT functions indicate that participants in both groups often relied on enumeration, and the equivalence of the functions, which was predicted by the metastrategic position, implies enumeration rates were unaffected by the large difference in context memory brought about by the generation task. If this interpretation is correct, it indicates that the between-group difference in estimated frequency must be related to strategy execution rather than to the particular mix of strategies settled on by participants in the two groups. Experiment 2, a protocol study, was designed to address this issue and also to provide converging evidence for the claim that enumeration rates are comparable in the two conditions.

Experiment 2

In this experiment, as in Experiment 1, participants in one group generated unique category instances during the study phase, and those in a second group read word pairs composed of multiply presented category labels and unique category instances. In addition, participants in both groups provided verbal protocols at test as well as at study. This method made it possible to obtain direct evidence concerning the relationship between study-phase generation and strategy selection and to examine possible between-group differences in strategy execution. If strategy selection is unaffected by the study-phase generation, as the RT data reported in Experiment 1 implied, then the protocols should indicate that enumeration rates are similar in the two conditions. In addition, prior research has demonstrated that simple enumeration becomes less common as presentation frequency increases, and estimates involving enumeration and extrapolation become more common (Brown, 1995). This trade-off should be present for both the non-generation condition (where it has been observed before) and the generation condition.

[2] This pattern of increasing RTs and relatively unbiased estimates observed in the generation condition has been reproduced twice in my laboratory. In one unpublished study, the generation condition ($n = 30$) yielded RTs that increased from 2.88 s to 8.33 s as presentation frequency increases from 2 to 16; in this case, the mean estimation slope was 1.02. In a second unpublished study, generation condition ($n = 30$) RT increased from 3.12 s to 11.35 s across the same range of presentation frequencies, and the estimation slope averaged 1.10.

It was also expected that a protocol analysis would isolate between-group differences in the strategy execution and that these differences would account for the between-group difference in estimated frequencies. If enumeration rates are similar in the two conditions and the estimates are not, it follows that strategies are executed differently by participants in the two groups. One possibility is that generation participants retrieve more items when they enumerate than non-generation participants. A second possibility is that extrapolation is more aggressive in the generation condition than the non-generation condition; and a third is that participants in the generation condition produce larger estimates when they make use of non-enumeration strategies than do their counterparts in the non-generation condition. Finally, it could be that there are several between-group differences and that these differences are explicably related to one another. For example, suppose more instances are retrieved during enumeration in the generation condition than in the non-generation condition. This might lead generation participants to believe that the response range is relatively wide and those in the non-generation condition to believe that it is relatively narrow. If so, these range differences should affect the size of the non-enumerated estimates (Brown, 1995, Experiment 3) and could also influence what people do when they extrapolate.

At the outset of this experiment, there were no grounds for selecting between these possibilities. However, when protocols are collected at study and test, one can determine how many items are recalled when people enumerate, how far they extrapolate when they produce estimates that differ from their enumerated counts, and what values are produced when enumeration-based strategies are not used. Thus, in addition to providing information about the strategies people use when they generate frequency judgments, the protocols make it possible to locate between-group differences in strategy execution and to link these differences to between-group differences in estimation performance.

Finally, it should be acknowledged that Experiment 1 data alone do not provide unambiguous support for the metastrategic position nor do they conclusively refute the cost-benefit account. Again, RTs reported above imply that enumeration rates were equivalent across the two conditions, and this conclusion, in conjunction with the between-group differences in estimated frequency, implies the existence of between-group differences in strategy execution. This line of reasoning depends critically on the assumption that RTs index enumeration in the same way in both conditions. Although this assumption is a parsimonious one, it need not be correct. Suppose, for example, that participants in the generation condition enumerated less often than their counterparts in the non-generation group, but spent more time on task when they did enumerate (reflecting their better memory for enumerable context words). In combination, these two between-group differences, one a difference in strategy selection consistent with the cost-benefit position, and the other a difference in strategy execution, could have produced the observed pattern of RT means and estimation differences. If this alternative account is correct, verbal protocols collected in this experiment should yield a between-group difference in enumeration rate as well as information about differences in the way the selected strategies are executed.

Methods

Materials and procedure

Materials used in this experiment were the same as those used in Experiment 1, and participants in this experiment followed the study-phase procedures that were used in the verbalization conditions in Experiment 1.

The procedure followed during the test phase was identical to the procedure used in Experiment 1, except that participants in both groups were instructed to think aloud as they generated their estimates. They were also warned that they would be prompted to say something if they fell silent for more than a few seconds. All verbal responses were tape recorded, though only the last 30 were analyzed as the first six served as a practice block. During the practice trials, the experimenter actively encouraged participants to speak and frequently prompted them if they did not. The cued-recall test used in this experiment was identical to one used in Experiment 1.

Participants

Forty undergraduates were recruited from the University of Alberta subject pool and randomly assigned in equal numbers to the generation and non-generation conditions. Participants were tested individually in the presence of a researcher and received course credit for sessions lasting about 45 min.

Results

Recall performance and frequency judgments

The frequency and recall data reported in this section were averaged and analyzed the same way as they were in Experiment 1; the relevant means are presented in Table 1 and Figs. 1 and 3. It is clear from these exhibits that the protocol procedure produced little if any reactivity (Russo, Johnson, & Stephens, 1989). Across the measures indicating generation success, cued recall performance, and estimation accuracy and bias, performance in the generation condition looked much the same in this experiment as in Experiment 1. Likewise, performance was comparable in the two non-generation conditions.[3]

Given these similarities, it is not surprising that this experiment produced the same pattern of between-group differences as Experiment 1. As before, participants in the generation condition ($M = 171.05$ words) recalled more words on the cued-recall task than did participants in the non-generation condition ($M = 101.1$; $F(1,38) = 49.44$, $p < .0001$, $MSE = 4.58$), and the between-group difference in recall success increased with presentation frequency (for the Group × Frequency interaction, $F(5,190) = 42.58$, $p < .0001$, $MSE = 0.85$). In addition, estimated frequencies were higher in the generation condition ($M = 6.56$) than in the non-generation condition ($M = 5.01$; $F(1,38) = 5.63$, $p < .05$, $MSE = 25.88$), and these estimates were also less biased and more accurate in both an absolute and relative sense. Specifically, the slope of the estimation function was closer to 1.0 in the generation condition ($M = 0.88$) than in the non-generation ($M = 0.66$, $t(38) = 1.76$, $.10 > p > .05$; $MSE = 0.12$); absolute error was smaller in the generation condition ($M = 1.89$; for the non-generation condition, $M = 3.44$; $F(1,38) = 33.05$, $p < .0001$; $MSE = 4.31$); and the rank-order correlation between estimated and actual frequency was larger (for the generation condition, $M = 0.95$; for the non-generation condition, $M = 0.89$; $t(38) = 4.08$, $p < .001$; $MSE = 0.02$).[4]

Strategy usage

The primary goal of this experiment was to determine whether study-phase generation affected strategy selection, strategy execution, or both. As a first step, it was necessary to code the verbal protocols collected during the estimation task. To do this, each response was assigned to one of three mutually exclusive categories: *Simple enumeration*, *enumeration and extrapolation*, and *other* (*non-enumeration*). A response was considered to involve enumeration when the participant mentioned at least one category instance prior to stating his or her frequency estimate. Enumerated responses were coded as "simple" when the number of items mentioned was equal to the magnitude of the estimated frequency; they were coded as "extrapolated" when the latter was larger than the former. Responses that did not involve enumeration were assigned to the "other" category.

Except for the six responses collected during the practice block, all protocols were scored by two teams of coders; the few disagreements that did occur were resolved through discussion. Next, for each participant and each level of presentation frequency, these codes were used to determine the percentage of all responses that were based on simple enumeration and the percentage that involved extrapolation. These percentages were submitted to an Enumeration Type (simple vs extrapolated) × Group (generation vs non-Generation) × Frequency (2, 4, 8, 12, 16) ANOVA.

Fig. 4 plots the strategy use percentages for simple enumeration, enumeration and extrapolation and total enumeration (the sum of simple enumeration and enumeration and extrapolation) as a function of presentation frequency and generation condition. Consistent with prior research (Brown, 1995; Experiment 1), these data indicate: (a) Enumeration-based estimates were very common, occurring on 59% of all trials. (b) Enumeration was common at all levels of presentation frequency; collapsing over condition, total enumeration percentages for the frequency levels 2, 4, 8, 12, and 16 were 55%, 57%, 62%, 61%, and 62%, respectively, ($F(4,152) = 1.29$, $p > .10$, $MSE = 170.37$, for the main effect of presentation frequency). (c) Simple enumeration traded-off with enumeration and extrapolation across

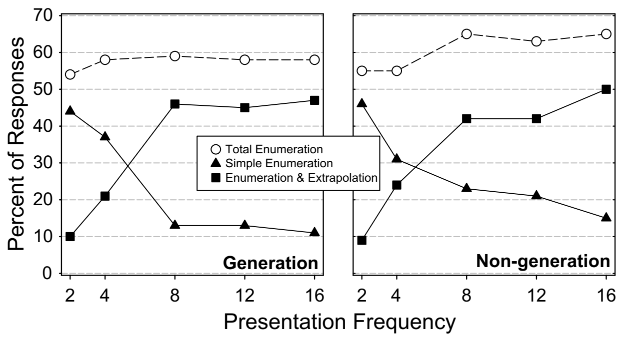

Fig. 4. Percent of enumerated responses as a function of presentation frequency in the Experiment 2 protocols. Total enumeration at a give level of presentation frequency (indicated by the open marks) represents the sum of simple enumeration (indicated by the black triangles) and enumeration-and-extrapolation (indicated by the black squares) for that level.

[3] Data presented in Fig. 3 seem to suggest that estimates collected in Experiment 1 might differ from those collected in Experiment 2. However, an Experiment (Experiment 1 vs Experiment 2) × Group (generation vs non-generation) ANOVA performed on the slopes indicated that neither the main effect of experiment ($F(1,155) < 1.0$) nor the Experiment × Group interaction ($F(1,155) = 2.28$, $p > 0.10$, $MSE = 0.19$) was reliable.

[4] As in Experiment 1, the 2-way ANOVAs performed on cued recall, estimated frequency, and absolute error means indicated that these measures reliably increased with presentation frequency (for recall, $F(5,190) = 385.88$, $p < .0001$, $MSE = 0.85$; for estimated frequency, $F(5,190) = 114.57$, $p < .0001$, $MSE = 7.87$; for absolute error, $F(5,190) = 127.80$, $p < .0001$, $MSE = 2.38$) and that the rate of increase was affected by the study-phase processing (for recall, $F(5,190) = 42.58$, $p < .0001$, $MSE = 0.85$; for estimated frequency, $F(5,190) = 2.35$, $p < .05$, $MSE = 7.87$; for absolute error, $F(5,190) = 8.59$, $p < .0001$, $MSE = 2.38$). Absolute error means for the six levels of presentation frequency were 0.00, 0.24, 0.65, 1.85, 3.51, and 5.12 in the generation condition, and.01, 0.70, 1.41, 3.47, 6.17, and 8.86 in the non-generation condition.

the five levels of presentation frequency. Collapsing over condition, the percentage of simple enumeration responses for the five levels of presentation frequency were 45%, 34%, 18%, 17%, 13%; comparable values computed for enumeration and extrapolation were 10%, 23%, 44%, 44%, and 48%. ($F(4,152) = 26.78$, $p < .0001$, $MSE = 689.89$, for the Enumeration Type × Presentation Frequency interaction).

Importantly, enumeration was no more common in the generation condition (57%) than in the non-generation condition (61%; $F < 1.0$, for the main effect of group). Moreover, study-phase generation had no effect on the type of enumeration strategy used; simple enumeration was used on 24% of trials in the generation condition and on 27% of the trials in the non-generation condition; the usage figures for enumeration and extrapolation were 34% and 33% for the generation and non-generation conditions, respectively, ($F(1,38) = 2.37$, $p > .10$; $MSE = 2833.59$, for the main effect of Enumeration Type; $F < 1.0$ for the Group X Enumeration Type interaction). Likewise, generation had no reliable effect on the Enumeration Type trade-off ($F < 1.0$, for the Enumeration Type × Group × Frequency interaction).

Strategy execution

Given that enumeration-based estimates were as common in the non-generation condition as in the generation condition, it follows that the observed differences in judged frequency must be related to differences in strategy execution. The data presented in Fig. 5, which presents estimate means computed as a function of strategy type for both conditions, support this implication and indicate where the differences lie. As this figure illustrates, participants in the two groups produced comparable estimates only when they used simple enumeration (for generation, $M = 4.14$, for non-generation, $M = 3.68$; Wilcoxon 2-sample Test $Z = .70$, $p > .10$). Otherwise, participants in the generation group produced larger frequency estimates. When enumeration and extrapolation was used, estimates averaged 9.58 and 6.53 in the generation and non-generation conditions, respectively, (Wilcoxon 2-sample Test $Z = 7.83$, $p < .0001$); comparable values for non-enumerated estimates were 8.60 and 7.16 (Wilcoxon 2-sample Test $Z = 4.02$, $p < .0001$).

In addition, each enumeration-and-extrapolation response was decomposed into an enumerated portion and an extrapolated portion—the value of the enumerated portion was determined by counting number of items recalled during the estimation process; the value of the extrapolation portion was determined by taking the difference between the final estimate and the enumerated portion. As is apparent in Fig. 5, participants in the generation condition recalled more words (5.79 vs 3.25; Wilcoxon 2-sample Test $Z = 8.57$ $p < .0001$) and extrapolated more aggressively (3.78 vs 3.28; Wilcoxon 2-sample Test $Z = 3.05$, $p < .01$) than did participants in the non-generation group.

These between-group differences were not confounded with differences in underlying presentation frequency. There were no between-group differences in the mean presentation frequency of the items that elicited simple enumeration responses ($M = 5.69$ and 6.56 for the generation and non-generation conditions, respectively, Wilcoxon 2-sample Test $Z = 1.30$, $p > .10$), enumeration-and-extrapolation responses ($M = 10.44$ and

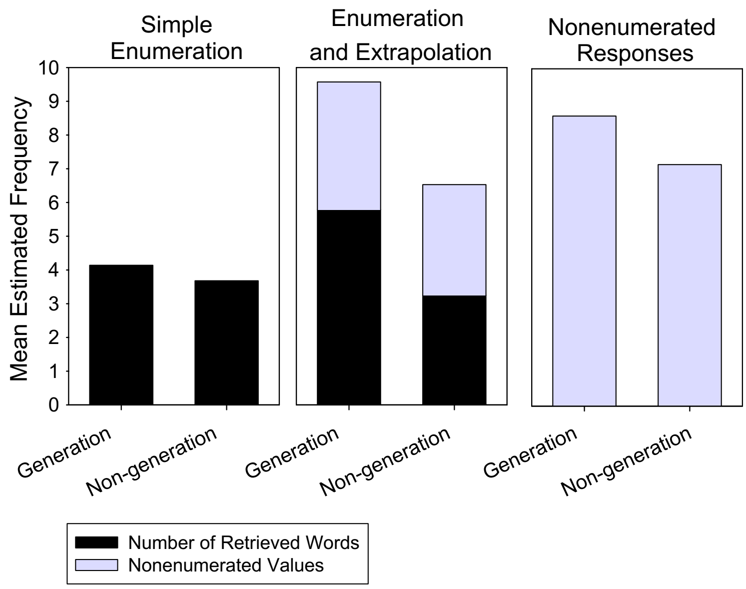

Fig. 5. Mean number of words retrieved during enumeration (dark bars in left and center panels), mean value of extrapolation (gray bars in center panel), and mean value of non-enumerated estimates (gray bars in right panel) from Experiment 2 protocols.

10.50 for the generation and non-generation conditions, respectively, Wilcoxon 2-sample Test $Z = 0.17$, $p > .10$), or non-enumerative responses ($M = 8.28$ and 7.89 for the generation and non-generation conditions, respectively, Wilcoxon 2-sample Test $Z = 0.73$, $p > .10$).

Discussion

This experiment replicated the effects of study-phase generation on cued recall and estimation performance. As before, recall was much better in the generation condition than non-generation condition, and frequency judgments were less biased. Also, consistent with the RT data collected in Experiment 1, the verbal reports indicated that participants in both conditions made frequent use of enumeration and that study-phase generation had no effect on strategy selection. In addition, these protocols indicated that the trade-off between simple enumeration and enumeration and extrapolation took the same form in both cases. These data provide converging support for the metastrategic account and suggest that participants sometime select non-enumeration strategies even when enumeration is possible.

Study-phase generation had no effect on strategy selection, but participants in the generation condition produced larger frequency estimates than those in the non-generation condition. The contents of the estimation reports were used to locate the origins of this global difference in the magnitude of the frequency judgments. These data indicated: (a) When people failed to enumerate, frequency estimates were larger in the generation condition than the non-generation condition. (b) When people made use of the enumeration-and-extrapolation strategy, estimates were larger in the generation condition than in the non-generation condition. In addition, when people used this strategy, (c) participants in the generation condition recalled more instances and extrapolated more aggressively than participants in the non-generation condition. In contrast, (d) when participants used a simple enumeration strategy, study-phase generation had no effect on estimate size, and, in both conditions, the enumerated estimates tended to be relatively small.

Given the robustness of the exemplar-generation effect, it easy to see how these strategy execution differences may have been conditioned by between-group differences in context memory. It appears that when presentation frequency was low (2 or 4), participants in both conditions often recalled (or believed that they had recalled) all relevant instances before terminating the retrieval process. In contrast, at higher frequencies (8 and above), enumerative retrieval was often terminated (see Fig. 4). However, because self-generated context items were more memorable than the ones provided by the experimenter, participants in the generation condition were able to recall and count more items before the retrieval process timed out.[5] It addition, extrapolation was somewhat more aggressive in the generation condition than in the non-generation condition, perhaps because unrecalled self-generated items carry more trace strength than unrecalled experimenter-selected items.

These two tendencies combined to produce enumeration-and-extrapolation-based estimates that were 47% larger in the generation condition than in the non-generation condition. In turn, this difference was probably responsible for the fact that estimates produced by non-enumeration strategies were larger in the generation condition. Certainly, this latter difference is consistent with prior research which has demonstrated that participants use the values of large enumeration-based estimates to induce the upper bound for the response range and that these range assumptions play a particularly important role when people use non-enumeration strategies to generate their estimates (Brown, 1995).

One final point. The protocols collected in this experiment provided converging evidence that enumeration rates are unaffected by study-phase generation and in so doing bolster the claim that metastrategic considerations play a role in the selection of frequency estimation strategies. On this view, enumeration-based strategies are used in part because they produce credible well-grounded estimates, and, in part, because they make a contribution to accuracy at the task level. There is evidence that enumerated responses are used to establish and partition the response range (Brown, 1995;

[5] Three additional assumptions are required here. The first is that the retrieval process terminates when the effort or time required to retrieve a context word accedes some threshold; the second is that retrieval difficulty increases as the number of retrieved items increases; and the third is that self-generated items are recalled faster than experimenter-selected items. The first two assumptions are consistent with Bousfield's classic work on category retrieval and output inference (Bousfield & Sedgewick, 1944; Indow & Togano, 1970). There are two reasons for accepting the third assumption. First, generation participants might use a systematic strategy to generate instances during the study phase (e.g., generate the names of relatives and close friends in response to the prompt WOMAN; generate the names of major Canadian cities in response to the prompt CITY). Second, regardless of the strategy they used, generation participants are likely to have retrieved readily accessible, highly familiar instances during the study phase. In contrast, participants in the non-generation condition are presented with an arbitrary set of familiar and less familiar category instances (e.g., Paris, Tampa). Thus, during the estimation task (and the cued-recall task) generation participants may be in the position to recall instances by reviving a recently executed search strategy and/or to restrict their search to familiar, personally relevant, instances. Both factors should facilitate the retrieval of recently generated instances in the generation condition, and the absence of these features should make instance retrieval relatively difficult in the non-generation condition.

Menon et al., 1995). In addition, it seems possible that enumeration might also be used to ensure that estimates remain well-calibrated and consistent over the course of a lengthy testing period.

If range setting is the enumeration's *only* task-level function then there should be a clear order effect; enumeration should be much more common during the first half of the test list than the second. This is because the generation of a consistent set of estimates requires that participants work with a stable set of range assumptions through the test phase. A Half (first half vs second half) × Group (generation vs non-generation) ANOVA[6] was performed in order to determine whether enumeration was more common during the first half of the test than during the second. This analysis indicated the enumeration rate were virtually the same across list halves (58% for the first half and 59% for the second half) and unrelated to study-phase generation ($F < 1.0$ for both main effects and the interaction). Thus, it seems unlikely range setting is enumeration's only task-level function. However, the exact nature of these additional functions remains an issue for future research.

General discussion

In summary, the relation between event memory and strategy selection was examined in two experiments. In Experiment 1, cued recall, RTs, and frequency judgments were obtained from participants who performed the exemplar-generation task and from ones who studied unique experimenter-generated word pairs. As predicted, memory was far better in the generation condition than in the non-generation condition, and frequency judgments were less biased. Also as predicted, RT increased with presentation frequency in both conditions. However, there was no between-group difference in the steepness of the RT functions. In Experiment 2, generation and non-generation participants thought aloud during the frequency judgment task. An analysis of these protocols indicated that enumeration-based estimates were equally common in both conditions. However, when participants in the generation group used the enumeration-and-exploration strategy they tended to recall more instances and extrapolate more aggressively than did participants in the non-generation group, and they tended to produce larger responses when they made use of non-enumeration strategies.

These data, in conjunction with those presented in prior studies (Brown, 1995, 1997; Conrad et al., 2003; Marx, 1985), support the conclusions that (a) enumeration increases with event memory, but only up to a point, and that (b) after that point, increased access to task-relevant event memories has no effect on the mix of enumeration-based and non-enumeration strategies. Consistent with the metastrategic account, this latter conclusion suggests that participants may have been balancing accuracy and effort at the task level, rather than at the trial level.

Although this is a novel proposal, it seems unlikely that this type of strategies-mix strategy is useful only when people estimate event frequencies. Generalizing from the present study, one can speculate that metastrategic considerations may play an important role in strategy selection when the following conditions are met: (a) The task must involve multiple trials; (b) target items can be ordered or categorized with respect to the criterion dimension; (c) there is an implicit or explicit demand for ordinal consistency across the set of items (i.e., the trials are not independent as the response to one item has implications for responses to other items); (d) there are at least two task-relevant strategies; (e) the more demanding strategy is believed to produce responses that are more precise and/or credible than the less demanding strategy. Under these conditions, participants should use the difficult strategy on some trials so that they can ground their responses, define a response range, and/or continually calibrate and refine the mapping between qualitative assessments and numerical responses. However, on other trials, particularly ones where there is a known (or readily inferred) relation between the current target and a prior target or between the current target and a previously considered category, relatively simple strategies should be used to generate relatively accurate responses.

Many real-world estimation tasks possess the properties outlined in the last paragraph (Brown, 2002b; A. Friedman & Montello, 2006; W. J. Friedman, 1993; Hertwig, Pachur, & Kurzenhauser, 2005; LaVoie, Bourne, & Healy, 2002; Lee & Brown, 2004), as do some memory and metamemory tasks (e.g., recency estimations, judgments of learning, judgments of knowing, etc; Hintzman, 2001; Koriat, Bjork, Sheffer, & Bar, 2004; Koriat & Levy-Sadot, 2001). For these tasks, the metastrategy position makes two general predictions. First, people will use a mixture of simple and complex strategies even when the conditions make it possible to perform the complex strategies accurately; second, people should mix the two types of strategies even when accuracy is emphasized. Of course, these predictions need to be tested. However, the main point here is that an accurate understanding of multiple-strategy situations will require researchers to recognize not only that people mix strategies over trials, but also that they can do this in a way that reduces effort without sacrificing accuracy or consistency.

[6] Because the first six trials were excluded (see above), the half way point in the test list fell between test word 21 and test word 22.

References

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

Betsch, T., & Sedlmeier, P. (2002). Frequency processing and cognition: Stock-taking and outlook. In P. Sedlmeier & T. Betsch (Eds.), *Frequency processing and cognition* (pp. 37–53). Oxford: Oxford.

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

Bogart, L. M., Walt, L. C., Pavlovic, J. D., Ober, A. J., Brown, N. R., & Kalichman, S. C. (in press). Cognitive strategies affecting recall of sexual behavior among high-risk men and women. *Health Psychology*.

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

Bröder, A. (2000). Assessing the empirical validity of the "Take-The-Best" heuristic as a model of human probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 1332–1346.

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

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

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

Brown, N. R. (2002b). Real-world estimation: Estimation modes and seeding effects. In B. H. Ross (Ed.). *Psychology of learning and motivation* (Vol. 41, pp. 321–360). New York: Academic Press.

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

Brown, N. R., Williams, R. L., Barker, E. T., & Galambos, N. L. (2007). Estimating frequencies of emotions and actions: A web-based diary study. *Applied Cognitive Psychology, 21*, 259–276.

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

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121–152.

Chu, P. C., & Spires, E. (2003). Perceptions of accuracy and effort of decision strategies. *Organizational Behavior and Human Decision Processes, 91*, 203–214.

Conrad, F. G., Brown, N. R., & Cashman, E. (1998). Strategies for answering behavioral frequency questions. *Memory, 6*, 339–366.

Conrad, F., Brown, N. R., & Dashen, M. (2003). Estimating the frequency of events from unnatural categories. *Memory and Cognition, 31*, 552–562.

Dougherty, R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DC: A memory process model of judgments of likelihood. *Psychological Review, 106*, 180–209.

Dunlosky, J., & Hertzog, C. (1998). Aging and deficits in associative memory: What is the role of strategy use? *Psychology and Aging, 13*, 597–607.

Edwards, M., Thomsen, S. C., & Toroitich-Ruto, C. (2005). Thinking aloud to create better condom use questions. *Field Methods, 17*, 183–199.

Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: Evidence on maximal adaptations on task constraints. *Annual Review of Psychology, 47*, 273–305.

Flexser, A. J., & Bower, G. H. (1974). How frequency affects recency judgments: A model for recency discrimination. *Journal of Experimental Psychology, 103*, 706–716.

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

Friedman, A., & Montello, D. R. (2006). Global-scale location and distance estimates: Common representations and strategies in absolute and relative judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 333–346.

Friedman, W. J. (1993). Memory for the time of past events. *Psychological Bulletin, 35*, 1–36.

Garry, M., Sharman, S. J., Feldman, J., Marlatt, G. A., & Loftus, E. F. (2002). Examining memory for heterosexual college students' sexual experiences using an electronic mail diary. *Health Psychology, 21*, 629–634.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103*, 650–669.

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

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

Hertwig, R., Pachur, T., & Kurzenhauser, S. (2005). Judgments of risk frequencies: Tests of possible cognitive mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 621–642.

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

Hintzman, D. L. (2001). Judgments of frequency and recency: How they relate to reports of subjective awareness. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 1347–1358.

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., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114*, 3–28.

Keller, K. L., & Staelin, R. (1987). The effects of quality and quantity of information on decision effectiveness. *Journal of Consumer Research, 14*, 200–213.

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General, 133*, 643–656.

Koriat, A., & Levy-Sadot, R. (2001). The combined contributions of the cue-familiarity and accessibility heuristics to feelings of knowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 34–53.

LaVoie, N. N., Bourne, L. E., & Healy, A. F. (2002). Memory seeding: Representations underlying quantitative estimations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 1137–1153.

Lee, P. J., & Brown, N. R. (2004). The role of guessing and boundaries in the telescoping of public events. *Psychonomic Bulletin and Review, 11*, 748–754.

LeMaire, P. (2005). Strategic aspects of human cognition: implications for understanding human reasoning. In M. Roberts (Ed.), *Cognitive Strategies and Human Reasoning* (pp. 11–29). London: Psychology Press.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95*, 492–527.

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.

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

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. (1995). Behavioral frequency judgments: An accessibility–diagnosticity framework. *Journal of Consumer Research, 22*, 212–228.

Mulligan, N. W., & Lozito, J. P. (2004). Self-generation and memory. In B. H. Ross (Ed.). *Psychology of learning and motivation* (Vol. 45, pp. 175–214). New York: Academic Press.

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.

Newell, B. R., & Shanks, D. R. (2003). Take the best or look at the rest? Factors influencing "one-reason" decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 53–65.

Newell, B. R., Weston, N. J., & Shanks, D. R. (2003). Empirical tests of a fast-and-frugal heuristic: Not everyone "takes-the-best". *Organizational Behavior and Human Decision Processes, 91*, 82–96.

Pandelaere, M., & Hoorens, V. (2006). The effect of category focus at encoding on category frequency estimation strategies. *Memory and Cognition, 34*, 28–40.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. New York: Cambridge University Press.

Rieskamp, J. & Hoffrage, U. (1999). When do people use simple heuristics, and how can we tell? In G. Gigerenzer, P. M. Todd, & the ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 141–167). New York: Oxford University Press.

Rieskamp, J., & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General, 135*, 207–236.

Robinson, M. D., & Clore, G. L. (2002). Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin, 128*, 934–960.

Russo, J. E., & Dosher, B. A. (1983). Strategies for multiattribute binary choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9*, 676–696.

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

Sedlmeier, P., & Betsch, T. (Eds.). (2002). *Etc. Frequency processing and cognition*. Oxford: Oxford University Press.

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

Siegler, R. S., & Jenkins, E. (1989). *How children discover new strategies*. Hillsdale, NJ: Erlbaum.

Siegler, R. S., & Lemaire, P. (1997). Older and younger adults' strategy choices in multiplication: Testing predictions of ASCM using the choice/no-choice method. *Journal of Experimental Psychology: General, 126*, 71–92.

Siegler, R. S., & Shipley, C. (1995). Variation, selection, and cognitive change. In G. Halford & T. Simon (Eds.), *Developing cognitive competence: New approaches to process modeling* (pp. 21–76). Hillsdale, NJ: Erlbaum.

Slamecka, N. J., & Graf, P. (1978). Generation effect—delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 592–604.

Swait, J., & Adamowicz, W. (2001). The influence of task complexity on consumer choice: A latent class model of decision strategy switching. *Journal of Consumer Research, 28*, 135–148.

Watkins, M. J., & LeCompte, D. C. (1991). Inadequacy of recall as a basis for frequency knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*, 387–396.

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

Zacks, R. T., & Hasher, L. (2002). Frequency processing: A twenty-five year perspective. In P. Sedlmeier & T. Betsch (Eds.), *Frequency processing and cognition* (pp. 21–36). New York: Oxford University Press.