ELSEVIER

Acta Psychologica 113 (2003) 23–44

acta psychologica

www.elsevier.com/locate/actpsy

Reducing bias in frequency judgment by improving source monitoring

Michael R.P. Dougherty *, Ana M. Franco-Watkins

Department of Psychology, University of Maryland, College Park, MD 20742, USA

Received 23 April 2002; received in revised form 22 October 2002; accepted 15 November 2002

Abstract

A common finding in judgment and decision making is that people's frequency judgments often fail to map onto objective frequencies. The present research examined the possibility that one source of bias in frequency judgment is attributable to people's inability to screen out irrelevant memory traces. We used a two-list source-monitoring paradigm to investigate whether frequency judgments are influenced by "extra-experimental" experiences and whether enhancing source monitoring improves judgment accuracy. Across four experiments we found: (1) frequency judgments regarding one list were biased by the second, (2) manipulating encoding between lists improved source monitoring and resulted in more accurate judgments, (3) manipulating item context between lists improved source monitoring and resulted in more accurate judgments, but only when the context was item specific, and (4) manipulating simple-background context between lists was ineffective at improving source monitoring.

PsycINFO classification: 2340
Keywords: Availability heuristic; Source monitoring; Frequency judgment; Judgment and decision making

1. Introduction

Many real world estimation tasks require participants to discriminate different sources of information. For example, imagine that you learn a contingency between variables A and B in context X, and a different (e.g., opposite) contingency between A and B in context Y. In judging the contingency between A and B, it is advantageous to discriminate the sources of the two relationships, since the accuracy of your judgment

* Corresponding author.
E-mail address: mdougherty@psyc.umd.edu (M.R.P. Dougherty).

0001-6918/03/$ - see front matter
doi:10.1016/S0001-6918(02)00149-X

depends crucially upon it. For example, your judgment in context X ought to be based on your observations of A and B in that context, without regard for the relationship in context Y. The case is similar for judgments of frequency. Since frequentistic information often co-varies with context (e.g., shark attacks are greater in Florida than in Maine, etc.), judged frequency ought to be context dependent. A failure to completely discriminate between frequencies learned in different contexts might be one source of bias found in frequency judgments.

Tversky and Kahneman (1973) proposed the availability heuristic as a descriptive account of how people make frequency and probability judgments. The essence of their account was that "ease of retrieval" could be used as a surrogate for assessing frequency or probability: judged frequency was assumed to be based on how easily relevant exemplars could be generated from memory. Because ease of retrieval was assumed to be affected by factors *unrelated* to objective frequency, frequency judgments often deviated systematically from objective frequency. For example, making rare events more salient could increase participants' perceived frequency of occurrence of the rare events, because increasing salience leads to increased availability. For example, in one study on the availability heuristic Tversky and Kahneman (1973, see also Lewandowsky & Smith, 1983) presented participants with a list of 39 names, 19 of which corresponded to famous males and 20 of which corresponded to less famous females. At test, participants were asked whether the list contained more male or female names. Eighty-one percent of the participants erroneously believed the list contained more male names than female names. Tversky and Kahneman proposed that the famous names were judged as more frequent because they were easier to retrieve from memory than were the less famous names, resulting in an availability bias.

While research investigating the availability heuristic certainly led to a better understanding of the factors affecting frequency and probability judgments, little progress was made in specifying the processes underlying availability and biased frequency judgments (Sedlmeier, Hertwig, & Gigerenzer, 1998). Indeed, our review of the literature revealed several interpretations of the term "availability", including number of instances retrieved in a short period of time (Tversky & Kahneman, 1973), subjective ease of retrieval (Schwarz et al., 1991), and familiarity (Dougherty, Gettys, & Ogden, 1999; Hintzman, 1988). Thus, stating that judgments were made via the availability heuristic fails to explicate the underlying cognitive processes, since any one of several "availability-type" processes could be used.

Dougherty et al. (1999) proposed a novel interpretation of the results of the famous names study presented above in terms of Minerva-DM, an adaptation of a global matching model to account for frequency and probability judgments, and the source-monitoring framework. [1] They showed that the results of the famous names

[1] The global matching perspective is one whereby all traces in long term memory contribute at the time of judgement. However, the degree of match between the cue used at test and memory traces affects how much each individual trace contributes to the overall output. Traces learned prior to an experiment will tend to contribute little to output, partly because of factors that might affect how well those traces are stored in memory (decay) and because many of these "extra-experimental" traces will tend to have different-context components. Both of these factors reduce the overall match between the cue and trace, and hence the degree to which extra-experimental traces contribute to the overall output.

study were consistent with an account in which participants failed to completely discriminate between names learned during the experiment and those learned prior to the experiment. Although participants were presented with only 19 famous names, they likely had experienced each of the famous names hundreds of times prior to the experiment (e.g., the name John Wayne was likely experienced many times prior to the experiment), but had experienced the less famous names only a few times prior to the experiment. In fact, Dougherty et al. (1999) examined the citation frequencies in a major newspaper database (Chicago Tribune, New York Time, and Washington Post) for a set of highly famous names (famous actors) and a list of less famous names (less known actors). The ratio of citations for famous male actors to less famous male actors was 48:1. For famous female actors and less famous female actors the ratio was 26:1. Thus, when judging the frequency of the famous names, their judgments could have been influenced, at least to some degree, by their pre-experimental experiences of the famous names. The inability to completely screen out pre-experimental traces would be due to a failure of accurate source monitoring.

Johnson, Hashtroudi, and Lindsay (1993) and Johnson and Raye (1981) proposed the source-monitoring framework to understand how people distinguish between difference sources of information. Source monitoring is an extension of the reality monitoring framework (Johnson & Raye, 1981), which was developed to study how people distinguish internally generated memories (e.g., the imagining of an event) from externally generated memories (e.g., the actual experiencing of an event). Johnson et al. (1993) proposed that source and reality monitoring decisions are made on the basis of characteristics of memory traces, such as the perceptual, contextual, or semantic details, or details about the affective state or cognitive operations present during storage of the trace. Source-monitoring processes enable people to discriminate between different subsets of memories on the basis of contextual features (e.g., color, spatial, and temporal details) or features about the cognitive operations used at encoding (e.g., rehearsal strategies; Henkel, Franklin, & Johnson, 2000; Johnson et al., 1993).

Relatively few studies have directly examined source-monitoring processes in judgment tasks. Johnson, Taylor, and Raye (1977) examined people's ability to discriminate frequency information based on external sources (words occurring on lists of words that were studied) from internal sources (items that had been recalled from the lists across successive trials). They found that judgments of how often words occurred on the lists (the external source) were influenced by how often those words had been recalled during a memory test (the internal source). Participants gave higher frequency judgments for items that had been recalled more often on successive recall tests, suggesting that they failed to complete discriminate between traces that were self-generated and traces resulting from study sessions. More recently, Hockley and Cristi (1996) revealed that participants' estimates of how often words had occurred as part of a word pair were relatively unaffected by how often that word had occurred as a singleton. Hockley and Cristi suggested that source-monitoring processes might enable participants to distinguish between different sources of frequency information, in particular associative and non-associative frequency information.

One way to conceptualize how source-monitoring processes might operate is in terms of a multiple-trace memory model (Dougherty et al., 1999; Hintzman, 1988). According to the multiple-trace view, each experienced event is encoded as a separate memory trace in memory. Thus, each trace in memory can include features that correspond to the to-be-judged item, contextual information, and features of the encoding processes. These contextual and encoding features can serve as cues that can be used to separate out different subsets of information in memory. For example, imagine that you learn a contingency between variables A and B in context X, and a different contingency between A and B in context Y. Judgments of the contingency between A and B would be highly inaccurate if contextual information were ignored. However, if context information could be used to discriminate between those instances in which A and B occurred together in context X versus when they occurred together in context Y, then judgments should be much more accurate.

One way to achieve context discrimination in the above task is to assume a two-part conditional memory process, where the first stage involves the discrimination of events that occurred in the two different contexts. For example, one could probe memory with context X to activate all traces in memory that have an "X" component. The second stage of the process would involve assessing the contingency between A and B among only those traces that occurred in context X. Thus, contextual information serves as the cue for separating instances of A and B occurring together in context X from those occurring in context Y. The accuracy of participants' judgments should depend, to some extent, on how well they can discriminate between context X and context Y. According to source-monitoring theory, any variable that enhances the differences between those events (e.g., contextual differences or differences in encoding operations), should in turn improve judgment accuracy. Such a two-stage process is embodied in recent applications of memory models, such as Dougherty et al.'s (1999) Minerva-DM model and Shiffrin and Steyvers's (1996) REM model.

The purpose of the present research was to examine bias in frequency judgment within the context of source-monitoring processes. We wished to examine whether one source of bias in frequency judgment, which is often attributed to use of the availability heuristic, might be due to *failures* in source monitoring. We also wished to examine whether taking steps to enable better source monitoring might decrease the amount of bias. Our goal was to shed light on the processes that might mediate biased frequency judgments often reported in the judgment and decision making literature.

The source-monitoring framework suggests that people should be able to discriminate between different sources of frequentistic information—whether a particular event is more frequent in one context or another—and that factors that improve source monitoring should reduce the extent to which items learned in one context affect judgments of frequencies for items learned in a different-context (cf. Begg, Maxwell, Mitterer, & Harris, 1986). Two factors that should improve source monitoring are: (1) the dissimilarity of the contexts and (2) differences in encoding operations between two sources of items (Johnson et al., 1993). In essence, these two variables increase the discriminability of the cues used to separate two sources of information.

Source monitoring should be better when the contexts in which two sets of items are learned differ considerably, and when the cognitive operations used to encode the items differs between the two contexts (even if the perceptual characteristics of the contexts are identical). In general, source monitoring capitalizes on the differences between the sources of memory traces: anything that can be used to discriminate the source of memory traces is valuable to the source-monitoring process. Thus, if *contexts* between two-lists of items differ, the source-monitoring process can use contextual differences to differentiate source. If the *encoding* operations differ for two-lists of items, then the source-monitoring process can use this information as a means of differentiating source. Thus, cross-context contamination should be minimized when contexts differ and when the encoding operations between two-lists differ.

2. Overview of experiments

In the present set of experiments, we examined frequency judgments using a two-list paradigm. Each experiment was divided into three phases: (1) a varied-frequency list study session (hereafter known as *varied list*), (2) a constant frequency list study session (hereafter known as *constant-list*), and (3) frequency judgment test. The varied-list is analogous to what might be called extra-experimental experiences, experiences learned prior to, or after the to-be-judged items were learned. Participants assessed the frequency of targets from the constant-list.

2.1. General method

2.1.1. Materials

Varied and constant-lists were constructed from the Batting and Montague (1969) and Toronto Word Pool (Friendly, Franklin, Hoffman, & Rubin, 1982) word norms. Single words were displayed in large black font in the middle of the computer screen for Experiments 1a, 1b, and 3 with a simple-visual background used as context information whereas word pairs were used in Experiment 2 with the second word in the pair explicitly defined as context information.

Varied-list: This list consisted of target and filler words. Two target words were allocated per frequency (0, 2, 4, 8, and 16) and counterbalanced across level of frequency, such that across participants, each target occurred equally often in each level of varied frequency.

Constant-list: This list consisted of target and filler words. The targets were identical to those presented in the varied-list and *always* appeared with a frequency of 4 on this list.

2.1.2. Procedure

The experimental sequence was constructed so that the varied-list preceded the constant-list (except where noted) with a brief distractor task implemented after

each list. At the beginning of each list, participants were informed that their memory for the words would be tested at a later juncture in the experiment, and specific instructions for remembering each list was provided. Additionally, they were notified that there were several parts to the experiment; however, they were unaware that they would see multiple lists.

At test, participants were instructed to estimate the frequency that each target occurred on the constant-list. They were instructed that their frequency estimate should be for the constant-list only. The function keys on the computer keyboard were labeled with numbers 1 through 12. Participants made their judgment by pressing one of these labeled keys.

The amount of bias in participants' judgments can be illustrated by the degree to which their estimates deviated from the true frequency of 4 as a function of the number of varied-list frequency exposures. We hypothesized that the amount of bias exhibited by participants would be affected by the number of times the target item occurred in the varied list, which should manifest in a significant judged frequency by varied frequency trend: participants should be more biased as the frequency of varied-list exposure increases. Additionally, in accord with the source-monitoring framework, we hypothesized that increasing the dissimilarity between the varied-list and target list context should lead to more accurate judgments. In Experiments 1a and 1b, we examined whether manipulating background context affected the ability to monitor targets in each list. In Experiment 2, linguistic context (using word pairs) was manipulated to demonstrate how verbal context facilitates the ability to monitor targets between the two-lists. In Experiment 3, we examined to what degree encoding strategy affected monitoring between lists. According to the source-monitoring framework, both contextual features and features of the encoding operations are used in source-monitoring decisions (Johnson et al., 1993). Thus, we predicted that judgment accuracy would be best when contextual features differed between lists and when encoding operations differed between lists.

3. Experiment 1a

The purpose of Experiments 1a and 1b was twofold. First, we sought to examine the extent to which participants' judgments might be biased by items learned prior to the to-be-judged items. Previous research using a two-list paradigm has demonstrated that discrimination is reduced for conditions where targets are presented in both lists, however, participants are able to recall frequencies for List 1 despite the fact that the same words were presented in List 2 (Reichardt, Shaughnessy, & Zimmerman, 1973). Second, assuming that such a finding holds, we wanted to examine whether simple-visual background context affected participants' ability to discriminate between items from the varied and constant-lists. In both experiments participants studied the varied and constant-lists with either the same background contexts or different background contexts. In Experiment 1a, participants were given as much time as needed to respond with their frequency judgment. In Experiment 1b participants were instructed to respond as quickly as possible and to give

their first impression. In both cases, we hypothesized that participants who studied the two-lists with the same background context would show more bias than participants who studied the two-lists with different background contexts because source discrimination should be better in the different-context condition. Thus, the different background context condition should be less influenced by the varied-frequency list.

3.1. Method

3.1.1. Participants

Thirty-five undergraduates at the University of Maryland participated in the experiment and received partial credit towards fulfillment of course requirements.

3.1.2. Materials

All words were displayed on one of two background contexts: context 1 consisted of a green and yellow geometric background and context 2 consisted of a gray background with a thin red border. The varied-list consisted of 10 target words (two per frequency) and 22 filler words presented once. Varied-list words were displayed using context 1. For the constant-list, all targets were presented with a frequency of 4 with 16 filler words presented once. Constant-list words were displayed on either context 1 or context 2 background.

3.1.3. Design and procedure

The design was a 2 (constant-list context: same, different) $\times$ 5 (varied-list frequency: 0, 2, 4, 8, or 16) mixed factorial. Constant list context was manipulated between participants and varied-list frequency was manipulated within participants. For both lists, participants were instructed to read each word as it appeared on the screen and word presentation was self-paced. The varied-frequency list context was identical for all participants. In the constant-list, participants were randomly assigned to one of two conditions: same-context ($n = 18$), or different-context ($n = 17$). The same background was used for both the varied and constant-lists in the *same-context* condition whereas the *different-context* condition viewed the varied-list with context 1 and constant-list with context 2.

At test, all target words were presented using the constant-list background (context 1 for the same condition, and context 2 for the different condition). For the frequency judgment, there was no restriction placed on response time, although most participants responded within 5–8 s.

3.2. Results and discussion

The 0-frequency condition was included in our design as a control condition to ensure that participants' judgments of items that did not occur on the varied-frequency list were unaffected by our manipulation. In this experiment, as well as

all the remaining experiments, judgments for 0-frequency items were never affected by our independent variables, in this case background context similarity, all p's > 0.20 across all experiments. As a consequence, we choose to analyze the 0-frequency items separately from the 2, 4, 8, and 16 frequency conditions. This also allows for a straightforward interpretation of the interaction terms of the ANOVA, since we can be certain that any interactions would not be due solely to the null effect in the 0-frequency condition and additive affect for the remaining conditions.

Fig. 1a presents the mean frequency judgments for targets presented during the constant-list. Consistent with our hypothesis, participants' judgments of the constant-list items were affected by the varied-list exposures. Overall, there was a main effect of the varied-list on judged frequency $F(3, 31) = 6.53$, $p < 0.05$. However

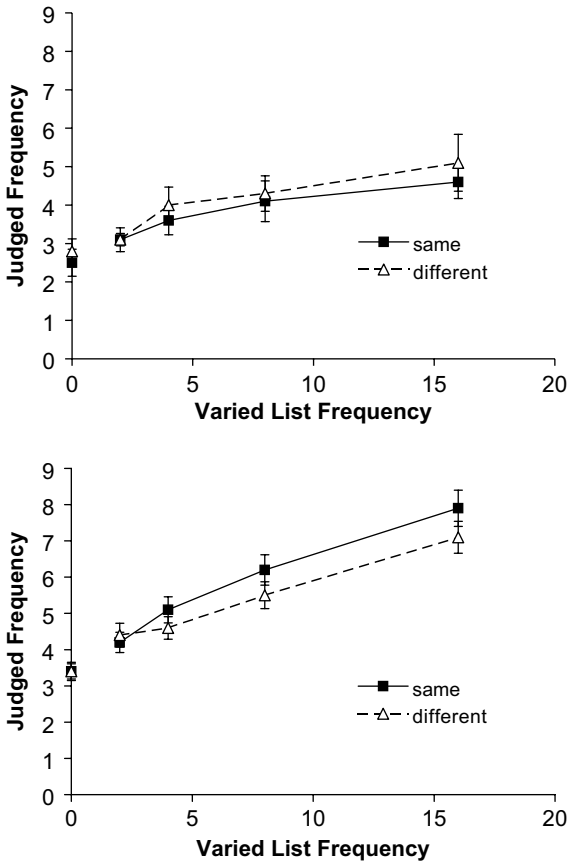

Fig. 1. (a) Top panel: mean estimated frequency and SEM of constant-list as a function of varied-list frequency and background context (without speeded judgment). (b) Bottom panel: mean estimated frequency and SEM of constant-list as a function of varied-list frequency and background context (with speeded judgment).

neither the main effect of context nor the interaction of frequency × context were significant, $F(1,31) = 0.23$, and $F(3,31) = 0.20$, $p > 0.05$, respectively. Trend analyses revealed a significant linear trend due to varied frequency, $F(1,33) = 20.03$, $p < 0.05$ ($\omega^2 = 0.35$), but the linear trend did not differ between same and different-contexts. Thus, contrary to our initial hypothesis, background context similarity did not affect the degree to which participants' estimates were biased by the varied-list.

In addition to collecting frequency judgments, we also collected reaction times (RTs) for the frequency judgments. In a recent series of papers, Brown (1995, 1997) proposed a multiple-strategy approach to frequency judgment. Two primary mechanisms within this framework are the enumeration process and the familiarity process. Brown demonstrated that people use enumeration under very specific circumstances. For example, enumeration tends to be used only when the target word is a category label and the context is an exemplar from the category. Under conditions in which the contexts are not distinct, or when category-exemplars are not used, participants tend to use a familiarity strategy to assess frequency. Thus, because our experiment did not contain either category labels or distinct contexts for each presentation, we would not expect enumeration to be used. The RT data are presented in Table 1. Overall, there was no effect of frequency on RT, context, or a context by frequency interaction (all p's > 0.05). That there was no effect of frequency on RT's indicates that participants were not using an enumeration strategy (Brown, 1995).

Why did context similarity fail to affect judged frequency? One possibility is that simple-visual background context is a poor cue for discriminating between lists. Because participants saw the same-context over and over, they may become inoculated against the context and may have not included the background context in the memory trace. It is possible that participants largely ignored the background context after

Table 1
Mean RT data for Experiments 1–4 (ms)

Experiment	Varied-list frequency				
	0	2	4	8	16
Experiment 1a					
Same-context	3784.3	4440.5	5202.8	4799.6	4566.9
Different-context	5136.5	4743.4	5891.2	5049.9	5948.9
Experiment 1b					
Same-context	1884.6	2212.4	2063.5	2203.8	2075.2
Different-context	1893.0	1781.5	1837.8	1864.1	1862.2
Experiment 2					
Same-context	2565.2	2611.0	2638.1	2545.6	2590.2
Different-context	2362.5	2710.7	2633.7	2510.8	2525.2
Experiment 3					
Rote	2141.7	2059.1	2134.0	2321.4	2256.4
Elaborative	1909.3	1844.8	1853.3	1990.8	2224.5

only a few words. If this were the case, then one would not expect a difference between the two contexts.

Given that we had predicted differences between the same and different context conditions a priori, we were surprised that we were unable to find such an effect. Thus, the purpose of Experiment 1b was to replicate this finding, to ensure that it was not peculiar to that particular sample of participants. Our only modifications were to increase our sample size to increase the power of our statistical test and to implement a time restriction for responding in the frequency judgment phase. We implemented the time restriction to prevent participants from engaging in an adjustment strategy. We reasoned that participants might sense that they were influenced by the varied-list, and more so when the contexts were identical, and therefore adjust their estimates downwards. In fact, this might explain our failure to find an effect of context in this experiment: the increase in adjustment could offset any effect context had on judged frequency. If participants were indeed using metacognitive awareness to adjust their estimates downwards, we would expect the magnitude of the bias (the degree to which participants are affected by the varied-list) to increase when participants are required to respond more quickly. Moreover, if the adjustment were greater for the same-context condition, we would expect the difference between same and different-context conditions (with same > different) to emerge with the time restriction.

4. Experiment 1b

4.1. Method

4.1.1. Participants

Seventy undergraduates at the University of Maryland participated in the experiment and received partial credit towards fulfillment of course requirements.

4.1.2. Materials

Varied and constant word lists were identical to those in Experiment 1a.

4.1.3. Design and procedure

The design and procedure were identical to Experiment 1a except for the test phase instructions. Participants were randomly assigned to either the same ($n = 35$) or different ($n = 35$) context conditions. Participants were instructed to use their first impression and respond within 2 s of reading the word on the screen when making their frequency judgment.

4.2. Results and discussion

Fig. 1b presents the mean judged frequency of targets during the constant-list as a function of varied-list frequency. Consistent with Experiment 1a, the varied-list

affected judged frequency: the main effect of frequency was significant, $F(3,66) = 46.00$, $p < 0.05$. Also consistent with Experiment 1a, there was no effect of context on judged frequency or a context × frequency interaction, $F(1,68) = 1.11$, and $F(3,66) = 1.13$, $p > 0.05$, respectively. Finally, trend analyses revealed a significant linear trend due to varied-frequency, $F(1,68) = 132.55$, $p < 0.05$ ($\omega^2 = 0.64$), but no effect of same versus different-context on steepness of the trend.

Analysis of the RT data revealed no effect of frequency, however, participants in the different-context condition responded significantly quicker than participants in the same-context condition, $F(1,68) = 6.93$, $p < 0.05$, which did not interact with frequency. Yet, for both context conditions, participants managed to respond closely to our instruction of responding within 2 s of the word being presented on the screen.

Note that two patterns emerge in comparisons of Fig. 1a and b. First, participants were much more influenced by the varied-list in Experiment 1b. For example, in Experiment 1a for varied-list frequency of 16, the mean estimates in the same and different-context conditions were 4.6 and 5.1, respectively. In Experiment 1b the corresponding estimates were 7.9 and 7.1 for the same and different-context conditions. This pattern was consistent across all levels of varied-frequency, including the 0-frequency words—words that did not even occur in the varied-list. Second, in contrast to Experiment 1a where participants showed *less* bias in the same-context condition, participants showed more bias in the same-context condition in Experiment 1b. Although neither of these last two comparisons was significant within experiment, it did suggest that the time restriction exacerbated the degree to which participants were affected by the varied list. For this reason, we included the time restriction for responding in our remaining experiments.

Experiments 1a and 1b demonstrated that judgments were biased by the varied-list. However, in neither experiment did we find evidence that background context was effective at improving source monitoring. This suggests either that context does not aid in discriminating between lists or that simple-visual background context is not sufficient to induce context effects—possibly because it is largely ignored at encoding. In Experiments 2, we manipulated context by providing each word with a unique context, rather than having a single context for the entire list. We assumed that participants would attend more to the context information at encoding if it were unique for each word, which would in turn make context information a better cue for discriminating between lists.

5. Experiment 2

As indicated previously, the source-monitoring framework leads to the prediction that source discrimination is better when the two studied lists are studied in different-contexts, and we were quite surprised by our failure to find such effects in Experiments 1a and 1b. As mentioned in the discussions of Experiments 1a and 1b, one possible explanation for our failure to find context effects was that perhaps participants largely ignored the background context once it had been presented a few times.

If this were the case, it would be possible to induce context effects by varying context within the lists. Rather than having a single context for an entire list, each word on the list was presented with either the same-context or a different-context on each presentation. In this experiment, we manipulated context using random word pairs rather than simple-background. One word was designated the target and the other as the context. In the same-context condition, target words from the varied and constant-lists were always paired with the same-context word. Thus, if *dog* was paired with *shoe* in the varied-frequency list, it was also paired with shoe in the constant frequency list. In the different-context condition the target words were presented with a randomly generated context in the varied-list, but always occurred with the same-context word in the constant frequency list. Thus, dog would appear with a randomly generated word each time it occurred in the varied-list, but always appeared with shoe in the constant list. Our use of words as context (often referred to as linguistic context) is not new, as it has been used widely in the recognition memory literature (see Dalton, 1993; Tulving & Thomson, 1973). Linguistic context provides a convenient way to manipulate contextual information at the item level.

This experimental design enabled us to examine whether context effects in frequency judgment would emerge if we could entice participants to attend to the contexts associated with each word. If participants discriminate source on the basis of context, then we would expect participants in the different-context condition to be less influenced by the varied-list exposures than participants in the same-context condition. An effect of context similarity would indicate that context information is an effective cue for discriminating lists, at least when the context is unique for each word.

5.1. Method

5.1.1. Participants

Thirty-eight University of Maryland undergraduates participated in the experiment and received partial credit towards fulfillment of course requirements.

5.1.2. Materials

Random word pairs were used for each list. In the word pair, the left word representing the target word was displayed using red text and the right word representing the context word was displayed using black text with the pair appearing in the center of a white computer screen. In the varied-list, target pairs were presented randomly with either the same-context or different-context word per frequency. Thus, half of the targets were paired with the same-context word (e.g., *rabbit–lamp*, *rabbit–lamp*, etc.) while the other half were paired with a different-context word each time the target was presented (e.g., *canary–stove*, *canary–book*, etc.). Two target words were allocated per frequency and 30 filler word pairs were presented once. In the constant list, the targets were identical to those presented in the varied-list with the same-context word used for each presentation. Target pairs were presented

four times and 30 filler pairs (10 old fillers from varied list and 20 new filler pairs) were presented once.

5.1.3. Design and procedure

The experimental design was similar to the previous experiments except for the fact that words rather than backgrounds were used as context. The design was a 2 (constant-list context: same, different) × 5 (varied-list frequency: 0, 2, 4, 8, or 16) factorial with both context and varied-list frequency manipulated within participants. Participants were not informed as to which word in the pair represented the target and context words or that there would be multiple lists in the experiment. For both the varied and constant-lists, participants were instructed to sub-vocally repeat each word pair until both words disappeared from the screen. All word pairs were presented for 3 s.

During test, the same-context word pair (as it appeared during the constant-list) was presented on the screen with the target word appearing in red and the context word in black text. Participants were instructed that they should use their first impression and respond as quickly and as accurately as possible when making their frequency judgment. If participants took longer than 4 s to make a frequency judgment, they were prompted to try to respond faster.

5.2. Results and discussion

Fig. 2 presents the judged frequency of targets as a function of context and varied-list frequency. Note that participants were more influenced by varied-list frequency for same-context than for different-context targets. A 2×4 repeated-measures ANOVA revealed a significant interaction for context frequency, $F(3, 35) = 2.99$, $p < 0.05$. Additionally, the main effect of context and varied-list frequency were also

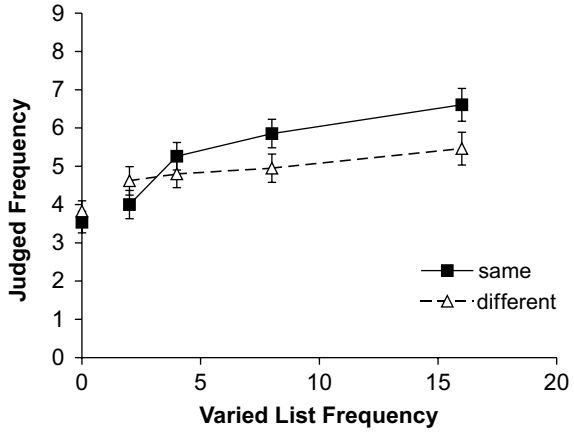

Fig. 2. Experiment 2: mean estimated frequency and SEM of constant-list as a function of varied-list frequency and word context.

significant, $F(1,37) = 8.51$ and $F(3,35) = 12.87$, $p < 0.05$, respectively. Trend analyses revealed an interaction of context by varied-frequency, $F(1,37) = 7.19$, $p < 0.05$ ($\omega^2 = 0.14$), as well as the linear trend of frequency which was present in Experiments 1a and 1b, $F(1,37) = 35.52$, $p < 0.05$ ($\omega^2 = 0.47$). Thus, the linear trend is affected by context associated with targets. Overall, the different-context led to less biased frequency judgments. Thus, contrary to Experiments 1a and 1b where we failed to find context effects when we manipulated background context similarity, the present experiment showed that context can be an effective cue for discriminating lists when the context is unique for each word.

Table 1 presents the mean RT data. Overall, participants responded within 2.4–2.8 s of word presentation and response time was not significantly affected by context or varied-list frequency. As in Experiments 1a and 1b, no evidence of enumeration was revealed, as the RT functions were flat.

The results of Experiment 2 support the idea that context can be an effective cue for source discrimination in frequency judgment if it is unique for each word. Participants were still fairly accurate at discounting the varied-list exposures and only overestimated the true frequency of 4.0 by an average of at most 2.6 for frequency-16 same-context condition. It seems clear that participants were performing some sort of list discrimination for both the same and the different context conditions, but that discrimination was much more accurate in the different-context condition.

In the next experiment, we examined whether better discrimination between lists can be induced by manipulating the type of encoding between lists. Johnson et al. (1993) proposed that source monitoring is better when the cognitive operations used at learning differed between lists. For example, instructing participants to engage in elaborative rehearsal for one list, but rote rehearsal for the other list should improve source monitoring above that which could be achieved if rote rehearsal were used for both lists.

6. Experiment 3

The purpose of Experiment 3 was to examine the degree to which having participants use an elaborative rehearsal strategy (mental imagery) for the varied-list words affected judged frequency of the constant-list words, which were learned using rote rehearsal. According to the source-monitoring framework, source discrimination should be better when the two-lists are studied under different encoding conditions. As Johnson et al. (1993) argue, source discriminations can be made on the basis of the cognitive operations used at encoding. On the one hand, elaborate rehearsal should increase participants' memory for items studied under better encoding. This is because improved encoding should affect the strength with which words respond when prompted with the retrieval cue. On the other hand, if two-lists are studied under different encoding conditions, the difference in encoding operations should provide a cue for discriminating different sources (lists) of information. Thus, the source-monitoring framework predicts that using elaborative rehearsal of the

varied-list words, but rote rehearsal of the constant-list words should lead to less bias in the judged frequency of the constant-list words. This is because participants should be able to better discriminate between the two-lists when they were studied under different encoding conditions.

6.1. Method

6.1.1. Participants

Forty-six University of Maryland undergraduates participated in the experiment and received partial credit towards fulfillment of course requirements.

6.1.2. Materials

Single words were displayed in one of two backgrounds (backgrounds used in Experiments 1a and 1b). The varied-list contained 10 targets (two per frequency) and 22 filler words presented once with all words presented on context 1 background that was used in Experiments 1a and 1b. The constant-list contained the 10 targets from the varied-list and 20 new filler words. For all participants, the words were presented in background context 2 used in Experiments 1a and 1b.

6.1.3. Design and procedure

The experimental design consisted of a 2 (encoding quality: elaborate, rote) $\times$ 5 (varied-list frequency: 0, 2, 4, 8, or 16) mixed factorial. Encoding was manipulated between participants and varied-list frequency was manipulated within participants. Encoding quality was manipulated during the varied-list using either rote rehearsal or mental imagery instructions, and participants were randomly assigned to either encoding condition. Participants in the rote rehearsal condition were told to repeat each word over and over sub-vocally for the duration (2 s) that it was presented on the screen. Microphones were placed in front of participants in order to encourage participants to repeat the words sub-vocally. Unbeknown to the participants these microphones were unplugged. Participants in the imagery condition were instructed to form a mental image of each word, and to form the same image if the same word occurred more than once. Participants provided a brief written description of each image by typing their image into the computer; thus, word presentation for imagery condition participants was self-paced. The encoding strategy used for the constant-list was identical for all participants with participants instructed to repeat each word sub-vocally for the 2 s that each word appeared on the screen.

Participants were instructed to use their first impression and respond within 2 s of reading the word on the screen when making their judgments. All words presented during test appeared with the constant-list background.

6.2. Results and discussion

Fig. 3 presents the mean judged frequency as a function of the rehearsal strategy used and the frequency of occurrence of targets in the varied-list. As can be seen, there

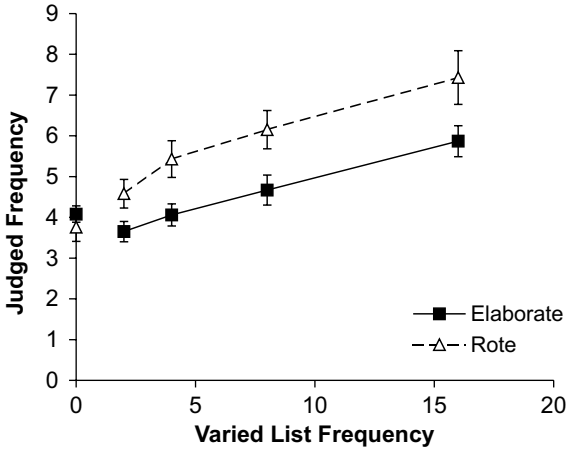

Fig. 3. Experiment 3: mean estimated frequency and SEM of constant-list as a function of varied-list frequency and encoding instruction.

was a monotonic increase of judged frequency as a function of varied-list frequency. The elaborative rehearsal participants estimated targets as occurring less frequently during the constant-list than participants in the rote rehearsal condition. Thus, there was less bias (deviation from the true frequency of 4) under elaborative rehearsal conditions. A mixed factorial ANOVA revealed significant main effects for rehearsal strategy, $F(1,44) = 7.31$, $p < 0.05$, and varied-list frequency, $F(3,42) = 24.82$, $p < 0.05$. The interaction term was non-significant, $F(3,42) = 0.74$, $p > 0.05$. Trend analyses revealed a significant linear trend of varied frequency, $F(1,44) = 74.14$, $p < 0.05$ ($\omega^2 = 0.61$), however, encoding did not affect this linear trend.

Inspection of the RT data indicated that enumeration was not used, as the RT functions were remarkably flat with no reliable effects of frequency or encoding (see Table 1).

The results of Experiment 3 suggest that use of an elaborative rehearsal strategy in the varied-list leads to less influence of extra-list items on judged frequency of constant-list items. However, in addition to there being differences between the two encoding conditions in terms of the cognitive operations, there were other differences. For example, only the imagery condition was required to type in their image, and the imagery condition was also given more time to study the individual words during the varied-list. Thus, this experiment does not provide a clean test of whether the reduction of bias in the imagery condition is due to the formation of images. The improvement in accuracy could be due to the fact that imagery participants were required to type their image for the varied-list (but not the constant-list), or due to a combination of imagery and typing. In short, the imagery condition had much richer information on which to base their source-monitoring decisions.

Experiments 2 and 3 found two interesting results. In Experiment 2, we found that frequency judgments were more accurate when the two-lists occurred in different-contexts and in Experiment 3, we found that judgments were more accurate when

the two-lists received different types of encoding. However, an interesting finding was that we found a context similarity by varied-frequency interaction in Experiment 2, but no interaction between rehearsal strategy and varied-frequency in Experiment 3. Since these were separate experiments with slightly different methods, it is difficult to compare the two sets of results. Hence, we attempted to replicate these findings in a fourth experiment. The basic methodology was identical to that of Experiment 2, where we used words as context. In addition to manipulating context similarity, we also manipulated whether participants engaged in rote rehearsal for both lists, or engaged in rote rehearsal for the constant-list and elaborative rehearsal for the varied-list. Context similarity was manipulated within participants, and rehearsal strategy between participants. A total of 89 participants completed the experiment, with participants randomly assigned to levels of encoding.

The results replicated those obtained in Experiments 2 and 3. As in Experiment 2, there was an interaction between varied-list frequency and context similarity, and this was the case both for the elaborative rehearsal and rote rehearsal conditions. In addition, as in Experiment 3, there was no such interaction between varied-list frequency and rehearsal strategy, and this was true both for same-context and different-context conditions. [2] Thus, when participants are able to use rehearsal strategy as a source monitoring cue, it produces additive effects. But when the only difference between the two-lists is in the associated context information, it produces multiplicative effects. Our explanation of these findings is presented in Section 7.

7. General discussion

We began this article with a discussion of sources of biases in frequency judgment. In particular, we proposed that failures in source discrimination could account for the common biases in frequency judgment often attributed to use of the availability heuristic (see also Dougherty et al., 1999). The present experiments provide evidence that at least one source of bias in frequency judgments is the failure to discount frequency information from non-target sources. In Experiments 2 and 3, we found that frequency judgments were more accurate when the contexts between the varied and constant-lists were different and when the encoding operations between the two-lists were different. These findings were replicated in a fourth experiment where we manipulated both context and encoding within the same experiment. Not surprisingly, this fourth experiment showed that discrimination was best when both encoding and context were different between the varied and constant lists. Interestingly, simple-background context was not found to be an effective cue for separating the two-lists (Experiments 1a and 1b). We propose that one source of bias in frequency judgment that has been attributed to the availability heuristic is actually the result of fallible source-monitoring processes.

[2] All reported findings were significant at $p < 0.05$.

Several models have been used to describe the availability heuristic and we briefly describe three classes: (1) recall process models, (2) metacognitive models, and (3) familiarity based models. At least three versions of the recall process model (overt or covert) have been proposed: number of instances recalled in a short period of time (Tversky & Kahneman, 1973), time to retrieve the first instance (Sedlmeier et al., 1998), and enumeration (Brown, 1995, 1997). The most well documented of these models is the enumeration model. In accord with Brown, if enumeration were being used to assess frequency, we would anticipate steep RT functions for the different-contexts compared to the same-context conditions as the frequency of pre-experimental exposure increases, but in our experiments we do not have evidence for an overt recall process since participants were instructed to respond as quickly as possible.

The most well known instantiation of the metacognitive models was proposed by Schwarz et al. (1991), who suggested that the subjective ease or difficulty of the recall task is used as information in the frequency judgment process. People presumably infer frequency of occurrence from the difficulty of the retrieval task: the more difficult the retrieval task, the less frequently an event was presumed to have occurred. This type of process lead to the non-intuitive prediction that events of the same frequency will be rated as more frequent when the recall task is viewed as easy than when it is viewed as difficult (Aarts & Dijksterhuis, 1999; Schwarz et al., 1991; Wänke, Schwarz, & Bless, 1995). Also included in the metacognitive class is the idea that frequency judgments might be based on a feeling of knowing (FOK; Hart, 1967) as was originally proposed by Tversky and Kahneman (1973).

Finally, the last class assumes frequency judgments are based on a familiarity strength derived from accessing memory. These models assume that frequency judgments arise from the same processes responsible for recognition memory (Dougherty et al., 1999; Hintzman, 1988), namely the strength of the match between a memory probe and traces stored in memory. Judged frequency is assumed to be based on the strength of the activation from probing memory, such that increases in familiarity for an item lead to an increase in judged frequency for that item. Indeed, there is considerable research consistent with the idea that familiarity drives frequency judgment in many situations (Brown, 1995, 1997; Greene, 1988; Hintzman, 1988; Hockley & Cristi, 1996). The familiarity based model of frequency judgment is embodied in the class of memory models known as global matching models, which assume that all items stored in memory contribute to recognition memory and frequency judgments according to the degree to which a memory probe (a retrieval cue) matches traces in memory. Similarly, neural network models (e.g., Anderson et al., 1977; Sedlmeier, 1999) have also been used for simulating frequency judgments.

Dougherty et al. (1999) illustrated that a multiple-trace memory model could account for the common bias associated with the availability heuristic. Specifically, Dougherty et al. hypothesized that availability biases might arise from the inability to completely discriminate between items learned in different-contexts, which is a failure of source monitoring (Johnson et al., 1993; Johnson & Raye, 1981). According to Dougherty et al. (1999), biased frequency judgments might arise when prior

experience with the to-be-judged events, where the prior experience is within an irrelevant context, is not successfully discriminated.

One interesting, yet perplexing finding is that we found an interaction between varied-frequency and context in Experiment 2, but no such interaction between varied-frequency and encoding in Experiment 3—a finding we replicated in the fourth experiment in which we manipulated both context and encoding in a single experiment. Why did we find an interaction when manipulating context but not encoding? Obviously, that we replicated the findings in a fourth experiment indicates that the results are unlikely due to statistical error, either a type I error for the context manipulation (Experiment 2) or a type II error for the encoding manipulation (Experiment 3). In addition, that the pattern was replicated using random word pairs, indicates that the difference between the findings in Experiments 2 and 3 was not due to subtle differences in methodology. This leaves a third possible explanation as most probable—that source decisions are made differently depending on what type of information is available to separate the two sources of information.

What type of processes might be evoked when discriminating the source of frequency information on the basis of context versus encoding? One possibility is that participants are relying entirely on a familiarity process for discriminating between different-contexts but invoke a recollective process when discriminating between items that received rote rehearsal versus those items for which images had been formed, such as the process proposed by Schwarz et al. (1991).

We propose that the default strategy for assessing frequency of occurrence in our experiments was based on a familiarity process, but that the output of this process can be augmented by secondary processes, such as the subjective ease of retrieval or recollection. Our manipulation of encoding always took place on the varied-frequency list, with some participants receiving elaborate rehearsal instructions and others rote rehearsal instructions. All participants received rote rehearsal instructions for the constant-list. Thus, in the context of our experiments, the subjective ease with which participants could recall the interactive image could serve as a cue for how often the word pair *did not* occur in the constant-list. Judged frequency could be inferred by assessing familiarity, and then discounting the familiarity by taking into account ease of retrieval. If *ease of retrieval* were unaffected by the number of times participants formed the image (i.e., it was independent of frequency of image formation), it would have led to the same magnitude of adjustment across all levels of varied-list frequency.

In contrast, the two-part conditional memory matching and global familiarity models predict an interaction between context similarity and the frequency of occurrence on the varied-frequency list. This is because the degree to which items on the varied-list contribute to the overall familiarity should be a function of both context similarity and varied-list frequency. The more similar each item is to the to-be-judged item, the more it should contribute to the overall familiarity signal (Hintzman, 1988). Thus, conditions in which participants study the two lists in the same-context should elicit greater familiarity than the different-context conditions. In addition, because each item studied on the varied-list is assumed to contribute to overall familiarity,

judged frequency should also increase as frequency of exposure on the varied list increases.

One way to conceptualize the process is to assume that the degree to which traces on the varied-list contribute to judgments of the constant list is proportional to the similarity between the to-be-judged item plus context and the stored items plus context. This type of process leads to the prediction that context similarity should interact with varied-list frequency.

One possible explanation for the improvement in judgment accuracy due to the rehearsal strategy manipulation is that simply having participants engage in a mental imagery improves source-monitoring decisions. This explanation leads to the prediction that participants should be more accurate when both lists receive elaborate rehearsal than when one receives elaborate and the other rote. We investigated this possibility through a short pilot study (not reported here) in which participants ran through the equivalent of our same-context conditions of Experiment 2, but where they either receive mental imagery instructions for both lists, or rote rehearsal for one list and elaborate rehearsal for the other. In fact, participants were less accurate (showed greater overestimation) when both lists were studied under mental imagery instructions. Note that the recollective or ease of retrieval process would be entirely uninformative when both lists were studied under mental imagery instructions. Thus, it is not the case that elaborative rehearsal leads to better source discrimination: the crucial factor appears to be that the two-lists elicit different retrieval processes which can then be used to discount those varied-list items.

Our results provide support for Dougherty et al.'s (1999) hypothesis that bias in frequency estimates can arise from the failure to completely discriminate between instances learned in different-contexts. More important, our data suggest that the accuracy of frequency judgments can be improved by taking measures that enhance people's ability to discriminate between different sources of frequency information. In short, source-monitoring processes can be exploited to improve the accuracy of frequency estimates.

All the aforementioned instantiations of the availability heuristic provide plausible process models for how people make judgments of frequency and probability. One problem with much of the research on availability is that many researchers failed to go beyond the vague level of description to describe which version of availability was implemented. We propose replacing the term "availability heuristic" with "availability bias", and using the term to describe data, not process. Our results provide evidence for one possible process underlying availability biases—a fallible source-monitoring process. The present research hopefully will provide a starting point for more fully exploring the implications of source-monitoring processes for judgment and decision making.

Acknowledgements

This research was supported by funds provided by the Department of Psychology, University of Maryland. We thank Jennifer Hunter, Adnan Abbasi, Brandon Abbs,

Kerry Devlin, and Dave Inklekofer for their help with all experiments, and two anonymous reviewers for their helpful comments.

References

Aarts, H., & Dijksterhuis, A. (1999). How often did I do it? Experienced ease of retrieval and frequency estimates of past behavior. *Acta Psychologica, 103*, 77–89.

Anderson, J. A., Anderson, J. A., Brown, U., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: some applications of a neural model. *Psychological Review, 84*, 413–451.

Batting, 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 Monograph, 80*, 1–46.

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.

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

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

Dalton, P. (1993). The role of stimulus familiarity in context-dependent recognition. *Memory & Cognition, 21*, 223–234.

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. (1999). Minerva-DM: a memory process model for judgments of likelihood. *Psychological Review, 106*, 180–209.

Friendly, M., Franklin, P. E., Hoffman, D., & Rubin, D. C. (1982). The Toronto word pool: norms for imagery, concreteness, orthographic variables, and grammatical usage for 1080 words. *Behavior Research Methods & Instruments, 14*, 375–399.

Greene, R. L. (1988). Generation effects in frequency judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 298–304.

Hart, J. T. (1967). Memory and the memory-monitoring process. *Journal of Verbal Learning and Verbal Behavior, 6*, 685–691.

Henkel, L. A., Franklin, N., & Johnson, M. K. (2000). Cross-modal source monitoring confusions between perceived and imagined events. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 32–335.

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

Hockley, W. E., & Cristi, C. (1996). Tests of separate retrieval of item and associative information using a frequency-judgment task. *Memory and Cognition, 24*, 796–811.

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

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review, 88*, 67–85.

Johnson, M. K., Taylor, T. H., & Raye, C. L. (1977). Fact and fantasy: the effects of internally generated events on the apparent frequency of externally generated events. *Memory and Cognition, 5*, 116–122.

Lewandowsky, S., & Smith, P. W. (1983). The effect of increasing the memorability of category instances on estimates of category size. *Memory and Cognition, 11*, 347–350.

Reichardt, C. S., Shaughnessy, J. J., & Zimmerman, J. (1973). On the independence of judged frequencies for items presented in successive lists. *Memory and Cognition, 1*, 149–156.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: another look at the availability heuristic. *Journal of Personality and Social Psychology, 61*, 195–202.

Sedlmeier, P. (1999). *Improving statistical reasoning: theoretical models and practical implications*. Mahwah, NJ, US: Lawrence Erlbaum Associates.

Sedlmeier, P., Hertwig, R., & Gigerenzer, G. (1998). Are judgments of the positional frequency of letters systematically biased due to availability? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 754–770.

Shiffrin, R. M., & Steyvers, M. (1996). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*, 145–166.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in recognition memory. *Psychological Review, 80*, 353–370.

Tversky, A., & Kahneman, D. (1973). Availability: a heuristic for judging frequency and probability. *Cognitive Psychology, 5*, 207–232.

Wänke, M., Schwarz, N., & Bless, H. (1995). The availability heuristic revisited: experienced ease of retrieval in mundane frequency estimates. *Acta Psychologica, 89*, 83–90.