

Revisiting the role of recollection in item versus forced-choice recognition memory

GABRIEL I. COOK and RICHARD L. MARSH
University of Georgia, Athens, Georgia

and

JASON L. HICKS
Louisiana State University, Baton Rouge, Louisiana

Many memory theorists have assumed that forced-choice recognition tests can rely more on familiarity, whereas item (yes–no) tests must rely more on recollection. In actuality, several studies have found no differences in the contributions of recollection and familiarity underlying the two different test formats. Using word frequency to manipulate stimulus characteristics, the present study demonstrated that the contributions of recollection to item versus forced-choice tests is variable. Low word frequency resulted in significantly more recollection in an item test than did a forced-choice procedure, but high word frequency produced the opposite result. These results clearly constrain any uniform claim about the degree to which recollection supports responding in item versus forced-choice tests.

When the encoding episode is held constant, memory performance will often covary with the querying procedure used at test. Tulving (1985) found the usual advantages of recognition memory over cued recall and of cued recall over free recall. He also demonstrated, using the remember–know procedure, that the relative amount of recollection was greatest in free recall, next greatest in cued recall, and least in recognition. The opposite pattern was found with know responses. Over the years since its introduction, the remember–know procedure has evolved into one method by which to separate the relative contributions of recollection and familiarity to memory performance (e.g., Jacoby, Yonelinas, & Jennings, 1997; Kishiyama & Yonelinas, 2003; Yonelinas, Dobbins, Szymanski, Dhaliwal, & King, 1996). According to such dual-process models, recollection (a remember response) reflects a threshold-like process, whereas familiarity (a know response) reflects an assessment of evidence (like familiarity) that falls along a continuous dimension. Moreover, the remember–know procedure is not limited to studies of recognition memory, as Tulving showed, but rather, has been used in diverse memory tasks ranging from cued recall (e.g., Lindsay & Kelley, 1996) to autobiographical recollection (e.g., Rubin, Schrauf, & Greenberg, 2003) and source monitoring (e.g., Guttentag & Carroll, 1997; Hicks, Marsh, & Rit-schel, 2002).

The present study used the remember–know procedure to revisit the issue of whether the contribution of recol-

lection is different in item versus forced-choice recognition. Gardiner, Java, and Richardson-Klavehn (1996) first used the remember–know procedure in two-alternative, forced-choice recognition. In conjunction with a levels of processing manipulation, they argued cross-experimentally that the remember–know procedure generalizes to forced-choice tests because they obtained more remember responses after deeper rather than shallower encoding. Khoe, Kroll, Yonelinas, Dobbins, and Knight (2000) directly compared the level of recollection obtained in item (yes–no) versus forced-choice recognition and found no recollective differences in the two procedures. That null outcome was obtained across a variety of encoding manipulations and held true for amnesic patients as well. Kroll, Yonelinas, Dobbins, and Frederick (2002, Experiment 3) also obtained equivalent amounts of recollection in the two kinds of tests. The issue is actually quite important because forced-choice tests are often assumed to rely less on recollection than item tests do (e.g., Aggleton & Shaw, 1996; Nolde, Johnson, & D'Esposito, 1998; Parkin, Yeomans, & Bindschaedler, 1994). In the forced-choice procedure, participants can ostensibly weigh the relative familiarity of both alternatives and choose the one that is greatest (e.g., Glanzer & Adams, 1990; Glanzer & Bowles, 1976). Consequently, unlike item recognition, recollection need not be consulted in a forced-choice design. Khoe et al.'s and Kroll et al.'s findings are important and surprising insofar as they contravene what may otherwise be a tacit and often-held assumption about the differences between forced-choice and item recognition.

Unfortunately, equivalence of recollection across the two test formats has not been replicated. Whereas Khoe et al. (2000) used either 100 or 200 medium-frequency words as study material, Bastin and Van der Linden (2003) used 18 pictures of novel faces. They found that

The authors thank Christopher Berardi, Justin Waldorf, and Todd Lindsey for their dedicated help in collecting the data. Correspondence concerning this article should be addressed to Richard L. Marsh, Department of Psychology, University of Georgia, Athens, GA 30602-3013 (e-mail: rlmars@uga.edu).

Table 1
The Proportion of Items Labeled Remember Versus
Known by Condition

Condition	Cell <i>N</i>	Hits		False Alarms		Corrected Recollection
		Remember	Know	Remember	Know	
Two Alternative						
Low WF	31	.40	.35	.04	.18	.36
High WF	30	.37	.32	.06	.20	.30
Item						
Low WF	30	.49	.27	.06	.17	.42
High WF	31	.28	.35	.09	.32	.19

Note—"Cell *N*" denotes the number of participants. "Corrected Recollection" equals hits called R minus false alarms called R. WF, word frequency.

recollection was greater in item recognition when compared with the forced-choice testing procedure. The stimulus materials were probably not directly responsible because Kroll et al. (2002) had participants study either 280 or 720 pictures and found equivalent recollection. Rather, what may differ between these studies is the overall level of recollection contributing to the recognition response. Discriminability (d') was lower in Khoe et al.'s experiments (under .80) compared with Bastin and Van der Linden's (above 1.5). Less reliance on recollection in two-alternative forced choice (because the judgment is a relative one) may be found only when the encoding conditions result in items that have a great deal of recollection. Therefore, Bastin and Van der Linden's finding of more recollection with item testing may have occurred because of the large amount of recollection their participants were able to recruit for an 18-item study sequence. The difference between studies may also be attributable to the specific recollective details that participants had. For example, the Khoe et al. study used an orienting task at encoding, whereas the Bastin and Van der Linden study did not. Therefore, in the former case, recollection may have been about the orienting task, whereas in the latter case, it may have constituted more idiosyncratic reactions to the items (e.g., seeing *shot* may have triggered a memory of receiving an immunization). In an analysis of subjective reports of what constitutes recollection, Gardiner, Ramponi, and Richardson-Klavehn (1998) found that personal, idiosyncratic reactions could account for significant recollection. Moreover, Bodner and Lindsay (2003) have shown that what constitutes recollection for an identical set of items can depend critically on what other items are present on the study or test lists.

Our motivation for the present study was predicated on a single speculation—namely, that idiosyncratic recollective details may be stronger or more enduring than those engendered by orienting instructions. By using a class of stimuli for which the basis of recollection consisted more of idiosyncratic responses to the words themselves than of recollections related to an orienting task, we predicted that we should be able to replicate the recollection advantage that Bastin and Van der Linden (2003) found in item recognition over forced choice. Khoe et al. (2000) used medium-frequency words, which may not support strong levels of recollection on an item test. By

contrast, Gardiner and Java (1990) showed that low-frequency words result in much more recollection than high-frequency words. For the present purposes, it matters little what theory predicts why low-frequency words lead to better recollection (e.g., attention likelihood theory [Glanzer & Adams, 1990] vs. recollective details [Joordens & Hockley, 2000]). It matters more that our prediction is that a recollection advantage in item tests over the forced-choice procedure may be obtained only with very memorable items such as low-frequency words and small numbers of studied pictures. Consequently, word frequency was manipulated in our study with the prediction that low-frequency words would support more recollection on an item test, but high-frequency words would be unaffected or perhaps show the opposite pattern.

We also explored an ancillary issue with two-alternative recognition. Hicks and Marsh (1999) demonstrated that three-alternative responses (remember, know, new) yield a more liberal response bias in item recognition tests than did two sequential judgments (old–new, followed by remember–know; see also Eldridge, Sarfatti, & Knowlton, 2002). Because a remember–know judgment was being used after a two-alternative choice, we wanted to rule out that taking the remember–know judgment somehow changed the way a forced-choice response was being made (cf. Kroll et al., 2002). For example, taking the remember–know judgment may augment the basis for a forced-choice response to include recollection that otherwise would not have been considered. If this possibility is true, performance should be better when remember–know judgments follow the two-alternative judgment as compared with when they do not because more information is contributing to the judgment. Kroll et al. addressed this issue, but did so with a cross-experimental comparison. Khoe et al. (2000) also did so, but only using a shallow orienting task. As a consequence, we believed it prudent to address this issue again with our stimulus materials.

METHOD

Design

Six between-subjects conditions were tested. For the fundamental question of interest, four conditions orthogonally crossed word frequency (high vs. low) with the type of test (item vs. two alternative). Thus, two conditions tested item recognition, but the study and test items differed for the two groups: One group studied high-

frequency words and was tested with high-frequency distractors whereas the other group studied low-frequency words as stimuli. In a similar fashion, two conditions had the exact same study conditions but received a two-alternative test format. In all four of these conditions, remember-know judgments were taken. To address the ancillary question of whether taking remember-know judgments affects two-alternative tests, we tested two conditions without remember-know judgments and compared them with the forced-choice condition with remember-know judgments.

Participants

Undergraduate students from the University of Georgia volunteered to participate in this experiment in exchange for partial credit toward a course research requirement. Each participant was tested individually in sessions that lasted approximately 25 min each. We attempted to test 30 participants in each of the six different between-subjects conditions, but an extra participant was tested accidentally in one of the forced-choice conditions and another in one of the item conditions (see *ns* in Table 1). Therefore, a total of 182 participants were tested and assigned pseudorandomly to the six conditions.

Materials and Procedure

A total of 120 low-frequency and 120 high-frequency words were selected that were roughly equal in length. By the Kučera and Francis (1967) norms, the low-frequency stimuli averaged 10.74 and the high-frequency stimuli averaged 91.09. For each participant, the software randomly selected anew 60 words to be studied and reserved the remaining 60 words as distractors on the recognition task. Prior to studying each word in the center of the computer monitor for 2 sec, participants were told that they would receive an unspecified memory test. For item recognition, the test list was constructed by randomly intermingling anew the study and distractor items. Thus, there were 120 test trials, with one condition receiving all low-frequency words and the other receiving all high-frequency words. Participants responded using one of three labeled keys: remember (R), know (K), and new. For forced choice, a distractor item was paired with a studied item. On half the test trials, the studied item was presented on the left, and on the other half it was presented on the right. Participants responded by pressing one of two labeled keys to indicate whether the studied item was on the right or the left. When remember-know was tested, participants made a second judgment using two labeled keys to indicate whether the chosen alternative was remembered or known (cf. Khoe et al., 2000).¹ Therefore, the alternatives were always of the same word frequency, but word frequency was manipulated between subjects. The instructions for remember-know judgments were ones that we have successfully used on previous occasions (e.g., Hicks & Marsh, 1999).

RESULTS AND DISCUSSION

We first assessed by d' whether taking the remember-know responses affected forced-choice recognition with the current materials and procedures. Following Macmillan and Creelman (1991), the hit rate for forced choice was computed as the old items presented on the left that were claimed to be old, whereas the false alarm rate was computed as erroneous claims that the left item was old when in fact the alternative on the right was the old item (using the reverse mapping leads to the same results). To better equate d' between item and forced-choice tests the $z(\text{Hit}) - z(\text{FA})$ d' value was adjusted down by the $\sqrt{2}$ in the latter conditions. Other corrections and approaches to d' exist (see, for example, Kroll et al., 2002), but because our emphasis is not on corrected recognition, but rather on recollection, this will suffice

for current purposes. Panel A in Figure 1 displays corrected recognition for all six conditions, with standard error bars. The first two bars represent low and high word frequency for forced choice without remember-know responses, and the middle two bars represent performance when remember-know responses were collected. In a 2 (high vs. low frequency) \times 2 (remember-know taken vs. not taken) analysis of variance, performance was better for low word frequency than for high [$F(1,117) = 17.68$]. There was no effect from whether remember-know responses followed or not, and there was no interaction [both $F(1,117) < 1$]. Therefore, this outcome suggests that taking remember-know responses does not change memory performance in a forced-choice test. Furthermore, this analysis replicated Kroll et al.'s conclusions based on their cross-experimental comparison as well as Khoe et al.'s results. The hit and false alarm rates for forced choice without remember-know judgments were .75 and .16, respectively, at low word frequency and .70 and .28 at high word frequency. The comparable data for the remaining conditions can be found in Table 1, which contains the raw remember-know proportions.²

We turn now to comparing d' for item versus forced-choice testing. The 2 (word frequency) by 2 (item vs. forced choice) ANOVA yielded a significant two-way interaction (consult the last four bars of Figure 1A) [$F(1,118) = 14.92$]. Performance was much better for low-frequency words under item recognition [$t(59) = 3.99$], but it was nominally worse for high-frequency words under item recognition [$t(59) = 1.62$, n.s.]. The absence of d' differences at high word frequency replicates Khoe et al. (2000). These data suggest that stimulus characteristics without orienting instructions can strongly affect direct comparisons of item and forced-choice recognition. The critical data concerning recollection are summarized in panel B of Figure 1. This panel summarizes recollection as the proportion of hits labeled remembered (so the complement to unity represents know responses). In the 2 \times 2 ANOVA, a significant interaction was obtained [$F(1,118) = 8.57$]. Item recognition resulted in significantly more recollection than the forced-choice format when word frequency was low [$t(59) = 2.18$], but significantly less recollection when word frequency was high [$t(59) = 1.99$]. In other words, recollection in item recognition can be either greater or smaller than in the forced-choice format depending on word frequency, but notice that word frequency did not affect recollection in the forced-choice format.

Some strict adherents to the dual-process approach would argue that we should have analyzed a measure of corrected recollection (hits labeled R minus false alarms labeled R). These values are summarized as the last column of Table 1. The same interaction [$F(1,118) = 8.89$], with the identical interpretation would have been obtained with that dependent measure rather than the one summarized in panel B. For the sake of completeness, when the know responses are converted to estimates of familiarity based on the independence assumption [$K/(1 - R)$], es-

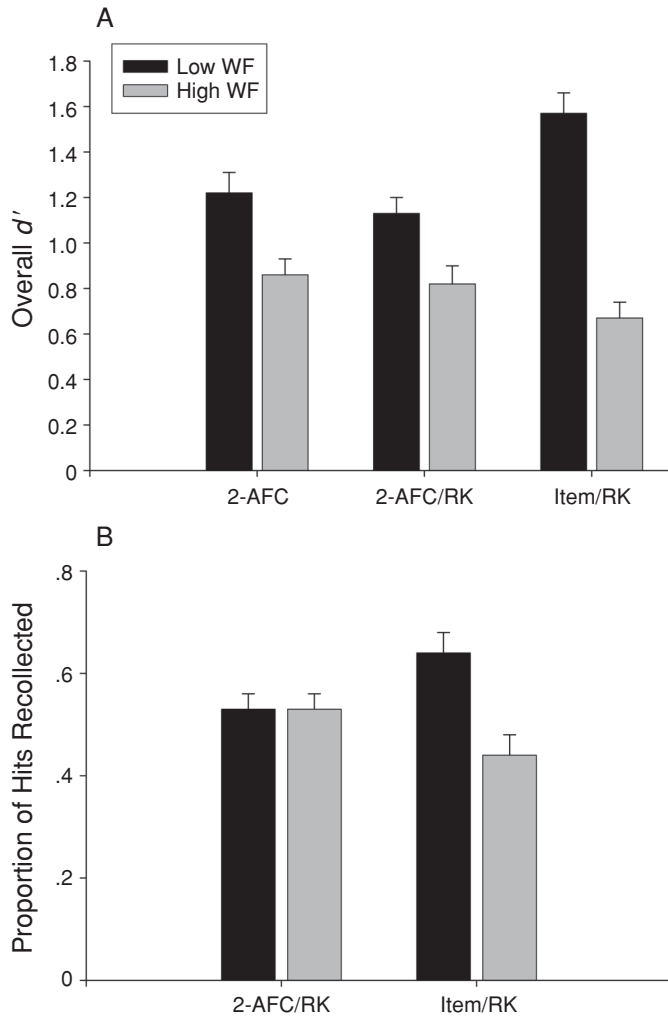


Figure 1. Panel A summarizes corrected recognition memory as d' across the six conditions, and panel B summarizes the proportion of the hit rate that was recollected. WF, word frequency; 2 AFC, two-alternative forced-choice task; RK, tests followed by remember-know judgment.

estimates are higher for low-frequency words (.56) than for high-frequency words (.49) [$F(1,118) = 5.87$], and this is true in both kinds of tests. Therefore, estimates of recollection are not necessarily invariant across the two test formats, but estimates of familiarity appear to be highly similar to those found by Kroll et al. (2002) and Khoe et al. (2000).

CONCLUSIONS

We have replicated the finding that using the remember-know procedure does not change performance on forced-choice tests. More important, we have shown that stimulus characteristics such as word frequency can lead to recollection differences in item tests that straddle the amount of recollection in a forced-choice test. Low word frequency led to estimates of greater recollection in item

than in forced-choice testing, whereas high word frequency lead to less recollection. The current data contravene the theoretical claim that recollection and familiarity are invariant across the two test formats. Rather, the current data suggest that recollection has a variable relationship across the two test formats. We have tested conditions leading to both more and less recollection, and others have found it to be invariant (e.g., Khoe et al., 2000; Kroll et al., 2002). However, we did replicate the invariance of familiarity's contribution in the two different test formats.

We concur with Bodner and Lindsay (2003) that remember responses in the remember-know procedure will reflect both the stimulus characteristics and the manner in which encoding takes place. In the present study, we expected greater recollection for low-frequency words than for high-frequency words with an item recog-

nition test because Gardiner and Java (1990) had found the identical result. In order to explain the interaction we obtained for recollection, the discriminability of the alternatives in the forced-choice must be considered. In the low-frequency case, the studied alternative contains many recollective details and quite a bit of familiarity (e.g., Mandler, 1980). The new alternative has neither pre-existing familiarity nor recollective details, and consequently, participants in this condition may discover that they do not need to depend as much on recollective details to make this relative judgment. By contrast, when the alternatives are of high frequency, the studied item does not have as much recollective detail as a low-frequency item would, and the new alternative already has high pre-existing familiarity. Moreover, whatever familiarity the high-frequency alternative had would have dissipated more quickly than it would have for a low-frequency item (Mandler, 1980). Therefore, the discrimination is much more difficult with high-frequency alternatives than with low-frequency alternatives. Thus, participants may discover that they need to rely more on recollective details and search more thoroughly for them. Although this explanation is not directly supported by the data, it remains an open question whether Kroll et al. (2002) would have found the same results with many fewer studied pictures, or whether Khoe et al. (2000) would have found invariance of the test format without an orienting task or if low word frequency had been used. Regardless, the present results argue that memory theorists can no longer claim that item tests usually recruit more recollection than their forced-choice test counterparts (e.g., Nölde et al., 1998). By the same token, a claim of uniform equivalence cannot be unambiguously asserted either. We do believe that remember responses index recollection, but they index recollection in a way that is sensitive to both the encoding and test contexts (Bodner & Lindsay, 2003). Of course, our results are wholly dependent on our use of the remember-know procedure, and before the conclusion can be drawn firmly that a variable relationship exists in recollection across the two test formats, other procedures of assessing recollection must be tested.

Unfortunately, the variable relationship of recollection in the two types of tests may greatly limit our ability to compare them. This situation is unfortunate because the different test formats need to be taken into account (Kroll et al., 2002) in comparisons across laboratories, across human and animal studies, and with different neuropsychological patients. The variable relationship, however, is predicted by the general idea with which we opened our introduction. That is, memory depends not only on how information is studied, but also on the circumstances surrounding the test situation (see, e.g., Guttentag & Carroll, 1997; Hicks & Marsh, 1999). Similar to Tulving's (1985) findings that different types of recall tasks are supported by different amounts of recollection, our findings suggest that different types of recognition memory tasks can recruit different amounts of recollection even when the study episodes are otherwise identical.

REFERENCES

- AGGLETON, J. P., & SHAW, C. (1996). Amnesia and recognition memory: A re-analysis of psychometric data. *Neuropsychologia*, *34*, 51-62.
- BASTIN, C., & VAN DER LINDEN, M. (2003). The contribution of recollection and familiarity to recognition memory: A study of the effects of test format and aging. *Neuropsychology*, *17*, 14-24.
- BODNER, G. E., & LINDSAY, D. S. (2003). Remembering and knowing in context. *Journal of Memory & Language*, *48*, 563-580.
- ELDRIDGE, L. L., SARFATTI, S., & KNOWLTON, B. J. (2002). The effect of testing procedure on remember-know judgments. *Psychonomic Bulletin & Review*, *9*, 139-145.
- GARDINER, J. M., & JAVA, R. I. (1990). Recollective experience in word and nonword recognition. *Memory & Cognition*, *18*, 23-30.
- GARDINER, J. M., JAVA, R. I., & RICHARDSON-KLAVEHN, A. (1996). How level of processing really influences awareness in recognition memory. *Canadian Journal of Experimental Psychology*, *50*, 114-122.
- GARDINER, J. M., RAMPONI, C., & RICHARDSON-KLAVEHN, A. (1998). Experiences of remembering, knowing, and guessing. *Consciousness & Cognition*, *7*, 1-26.
- GLANZER, M., & ADAMS, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *16*, 5-16.
- GLANZER, M., & BOWLES, N. (1976). Analysis of the word-frequency effect in recognition memory. *Journal of Experimental Psychology: Human Learning & Memory*, *2*, 21-31.
- GUTTENTAG, R., & CARROLL, D. (1997). Recollection-based recognition: Word frequency effects. *Journal of Memory & Language*, *37*, 502-516.
- HICKS, J. L., & MARSH, R. L. (1999). Remember-know judgments can depend on how memory is tested. *Psychonomic Bulletin & Review*, *6*, 117-122.
- HICKS, J. L., MARSH, R. L., & RITSCHER, L. (2002). The role of recollection and partial information in source monitoring. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *28*, 503-508.
- JACOBY, L. L., YONELINAS, A. P., & JENNINGS, J. M. (1997). The relationship between conscious and unconscious (automatic) influences: A declaration of independence. In J. D. Cohen & J. W. Schooler (Eds.), *Scientific approaches to consciousness* (pp. 13-47). Mahwah, NJ: Erlbaum.
- JOORDENS, S., & HOCKLEY, W. E. (2000). Recollection and familiarity through the looking glass: When old does not mirror new. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *26*, 1534-1555.
- KHOE, W., KROLL, N. E. A., YONELINAS, A. P., DOBBINS, I. G., & KNIGHT, R. T. (2000). The contribution of recollection and familiarity to yes-no and forced-choice recognition tests in healthy subjects and amnesics. *Neuropsychologia*, *38*, 1333-1341.
- KISHIYAMA, M. M., & YONELINAS, A. P. (2003). Novelty effects on recollection and familiarity in recognition memory. *Memory & Cognition*, *31*, 1045-1051.
- KROLL, N. E. A., YONELINAS, A. P., DOBBINS, I. G., & FREDERICK, C. M. (2002). Separating sensitivity from response bias: Implications of comparisons of yes-no and forced-choice tests for models and measures of recognition memory. *Journal of Experimental Psychology: General*, *131*, 241-254.
- KUČERA, H., & FRANCIS, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.
- LINDSAY, D. S., & KELLEY, C. M. (1996). Creating illusions of familiarity in a cued recall remember-know paradigm. *Journal of Memory & Language*, *35*, 197-211.
- MACMILLAN, N. A., & CREELMAN, C. D. (1991). *Detection theory: A user's guide*. New York: Cambridge University Press.
- MANDLER, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review*, *87*, 252-271.
- NOLDE, S. F., JOHNSON, M. K., & D'ESPOSITO, M. (1998). Left prefrontal activation during episodic remembering: An event-related fMRI study. *NeuroReport*, *9*, 3509-3514.
- PARKIN, A. J., YEOMANS, J., & BINDSCHAEDLER, C. (1994). Further characterization of the executive memory impairment following frontal lobe lesions. *Brain & Cognition*, *26*, 23-42.

- RUBIN, D. C., SCHRAUF, R. W., & GREENBERG, D. L. (2003). Belief and recollection of autobiographical memories. *Memory & Cognition*, **31**, 887-901.
- TULVING, E. (1985). Memory and consciousness. *Canadian Psychologist*, **40**, 1-12.
- YONELINAS, A. P., DOBBINS, I., SZYMANSKI, M. D., DHALIWAL, H. S., & KING, L. (1996). Signal-detection, threshold, and dual-process models of recognition memory: ROCs and conscious recollection. *Consciousness & Cognition*, **5**, 418-441.

NOTES

1. The two-alternative tests used a sequential judgment procedure of first identifying which alternative was old and then providing a remember-know judgment, whereas the item tests used a simultaneous three-alternative

judgment. Because the two-alternative test is a criterion-free measure, this difference should not matter. Moreover, Hicks and Marsh (1999) found that the difference in simultaneous versus sequential remember-know judgments was on criterion C , not on d' , and the latter is more important in this study.

2. The false alarm rates to high-frequency words in this study were sizable, thereby resulting in a strong mirror effect. However, we are confident that participants understood the remember-know instructions because they read detailed descriptions and the experimenter reviewed the instructions and queried the participants about them before allowing the test phase to be administered.

(Manuscript received May 28, 2004;
revision accepted for publication December 6, 2004.)