Similarity-Guided Depth of Retrieval: Constraining at the Front End

Yujiro Shimizu and Larry L. Jacoby, Washington University

Abstract Lee Brooks has done important work to show that categorization often reflects reliance on specific instances rather than on an abstract representation. His work on the advantages of using a diagnostic hypothesis to search medical stimuli has demonstrated how *constraining* what one looks for influences clinical reasoning. Similarly, cognitive control can be accomplished by constraining memory retrieval in ways that influence interpretation of a memory probe. Here, we report two experiments in which the similarity of study items constrained how test items were interrogated for an immediate memory test and thereby produced differences in the *depth of retrieval*. A novel procedure that tests *foil memory* was used to diagnose differences in similarity-guided retrieval depth.

Lee Brooks forwarded the idea that judgments such as categorization of a stimulus are sometimes based on the similarity of that stimulus to a specific exemplar stored in memory, rather than to a prototype (Brooks, 1978, 1987; see also Jacoby & Brooks, 1984). More recently, Lee and his colleagues have expanded their investigation of "nonanalytic" cognition toward clinical applications, examining the influence that prior instances and the availability of a diagnostic hypothesis have on the reliability and accuracy of medical diagnoses (e.g., Brooks, LeBlanc, & Norman, 2000; Brooks, Norman, & Allen, 1991; Kulatunga-Moruzi, Brooks, & Norman, 2001; LeBlanc, Brooks, & Norman, 2002; Norman & Brooks, 1997).

From our perspective, this work highlights the importance of having front-end *constraints* on what information comes to mind. As an example, Lee's research has revealed that the generation of a diagnostic hypothesis can be of greater advantage when it is used to *constrain*, up front, the search for supporting features, rather than when it is synthesized later from having gathered unconstrained data (Norman, Brooks, Colle, & Hatala, 1999). Similarly, here we illustrate how "constraining at the front end" operates in answering a query about one's immediate memory, a notion we refer to as *similarity-guided depth of retrieval*.

Our approach contrasts sharply with traditional descriptions of recognition memory, such as global matching models (e.g., Gillund & Shiffrin, 1984), which emphasize the *quantitative* relationship between the strength or familiarity of a memory probe against some decision criterion: By this account, if a probe's familiarity exceeds criterion, it is judged as "old," otherwise, it is judged as "new." This perspective largely neglects the *qualitative* bases used in making recognition memory judgments, bases that, we argue, are critical for constraining retrieval. Specifically, we suggest that imposing constraints on what comes to mind during retrieval influences the bases by which "old" items are accepted and "new" items are rejected. As will be shown, predictions regarding the fate of memory for new items (foils) follow directly from this line of reasoning.

A recent experiment from our lab illustrates the notion of *retrieval depth*. Jacoby, Shimizu, Daniels, and Rhodes, (in press) manipulated levels of processing (Craik & Lockhart, 1972) such that participants made pleasantness (deep) judgments for words in one list and vowel (shallow) judgments for words in another list. During a second phase, participants were given a recognition memory test in which they were told correctly that "old" words were ones for which they had earlier made pleasantness judgments, and another recognition memory test for which they were told correctly that "old" words were ones for which they had earlier made vowel judgments. That is, participants were correctly informed regarding the source of old items for each of the tests. As expected, we found the levels of processing effect for these initial recognition memory tests, with higher recognition memory for pleasantness-judged words.

More important, we found evidence for differences in retrieval depth due to specifying the source of the earlier-presented old words. This evidence came from a third phase in which we tested participants' memory for the *new items (foils)* that had appeared on the previous recognition memory tests. Deep foils, new words

initially encountered during the recognition test for pleasantness-judged old words, were better recognized later than were shallow foils, new words that were initially encountered during the recognition test for vowel-judged old words.

We predicted this result on the rationale that the earlier recognition had been accomplished by constraining retrieval processing in a way that recapitulated the original study processing of items. When attempting to recognize old words whose pleasantness was earlier judged, participants likely processed the meaning of *both* target words and foils, perhaps judging pleasantness to help decide whether pleasantness of the test word was earlier judged. In contrast, attempting to recognize old words whose vowels were earlier judged was likely to be less reliant on meaning-based processing. The goal of recognizing old words whose source had been specified guided the encoding of test items for both targets and foils. Subsequent recognition memory of the foils reflected source-constrained retrieval.

Experiment 1

The current experiments also used a subsequent recognition memory test for foils to measure retrieval depth. However, instead of varying the nature of the orienting task to produce differences in depth of retrieval, we varied the form of similarity shared by the old items. In the first phase of Experiment 1, participants received an immediate memory test consisting of a series of study/test trials. Each trial consisted of a "study set" of four words to remember followed by a single test probe. The words comprising the study sets were either semantically related (e.g., "BED, REST, WAKE, DREAM"), or orthographically similar (e.g., "TRUCK, TRAIN, TREND, TRAMP"). Following each study set, participants received a test probe that was either an old word (one of the four words that they just saw) or a foil. Importantly, the foils were always *unrelated* in meaning and appearance to the words in the study set to rule out alternative explanations based on the relatedness of foils to the study sets.

We predicted that the nature of the study sets would influence how the foils were interrogated. We expected the foils following the semantically related sets to be rejected on the basis of their meaning whereas we expected the foils following the orthographically related sets to be rejected on the basis of their appearance. Because meaning-based (deep) processing generally leads to better recognition memory performance, we expected that foils that followed a meaning-related set would be later better recognized than would be foils that followed an orthographically related set. This result would provide evidence for differences in retrieval depth.

Method

Participants. Sixteen Washington University undergraduates participated for course credit.

Materials. Foils were 84 words (80 critical, 4 buffers), four to seven letters in length, and were semantically and orthographically unrelated to the study sets. Critical items were rotated through three conditions: foils in a semantic context (20), foils in an orthographic context (20), and final test lures (40). The assignment of words to conditions was fully counterbalanced.

There were two types of study sets: semantic and orthographic. Each semantic set contained four semantically related words selected from McDermott and Watson (2001). Each orthographic set contained four visually similar words, with each word having an equal number of letters and beginning with the same two letters.

Procedure. In the first phase, participants received 68 study/test trials (60 critical and 8 buffer lists). For each test, participants indicated by response key whether a test probe was in the immediately preceding study set. An old probe was presented on 1/3 of the tests and a new probe was presented on 2/3 of the tests. The presentation order of study sets was intermixed and random. In the final phase, foil recognition was tested by intermixing 40 brand new foils with 20 foils encountered in a semantic context and 20 foils encountered in an orthographic context. Participants were told to judge an item as old if it was presented previously at any point during the study. In all phases of the experiment, responding was self-paced and order of presentation of words within a phase was random.

Results

All significance tests used a criterion of $p < .05$.

Recognition memory, corrected for guessing by subtracting false alarms from hits, on the immediate test was near perfect for both the semantic study sets ($M = .99$) and the orthographic study sets ($M = .98$), $F < 1$. Consequently, subsequent differences in foil recognition cannot be attributed to differences in general performance or false alarms during the initial tests.

Mean foil hit and false-alarm recognition rates for both experiments are shown in Table 1. As predicted, foils following a semantic set were better recognized than were foils following an orthographic set, $t(15) = 3.15$, providing evidence of differences in retrieval depth.

Our argument is that the shared similarity of the study sets dictated the manner in which both the tar-

TABLE 1
Probability of Responding "Old" for the Foil Recognition Test in Experiments 1 and 2

	Initial Test Context			
Experiment 1	Semantic	Orthographic		New
	.75	.68		.21
Experiment 2	Synonyms	Rhymes	Unrelated	New
	.75	.66	.69	. 29

gets and foils were interrogated. Foils encountered following a semantic set were interrogated for their meaning, and were thus better recognized on a later test relative to foils encountered following an orthographic set. Importantly, if subjects simply assessed the strength of each memory probe during the study/test trials, there would be no reason to expect differential processing of the foils depending on the shared similarity of the study sets, and consequently, no reason to expect differences in encoding and subsequent memory for foils.

Experiment 2

In Experiment 1, the similarity of study items along a salient dimension, such as meaning or appearance, influenced what was encoded about a test item and subsequent retrieval depth. In Experiment 2, we sought to extend this finding. The changes included using study sets comprising synonyms, rhymes, or unrelated words. Set size was increased to 5, and following each study set, 10 test probes were presented (5 old and 5 new) instead of just 1. We also included a prompt before each study set that indicated the dimension on which the words would be similar. This latter addition was made to reduce between-participant variability in the time taken to realize the nature of the similarity dimension. Despite these changes, we expected to find evidence of differences in retrieval depth as indicated by the subsequent recognition memory test for foils. Specifically, we predicted that foils following the synonym set (predicted to be rejected on the basis of meaning) would be better recognized later than foils following the rhyme set (predicted to be rejected on the basis of sound). No predictions were made for later recognition of foils following the unrelated set.

Method

Participants. Eighteen Washington University undergraduates participated for course credit.

Materials. Foils were 120 words, four to six letters in length, and were semantically, phonemically, and orthographically *unrelated* to the study sets. These words were rotated through four conditions: foils in a synonym context (20), foils in a rhyme context (20), foils in an unrelated context (20), and final test lures (60). The assignment of words to conditions was fully counterbalanced.

There were three types of study sets: synonym, rhyme, and unrelated. The synonym sets were words similar in meaning (e.g., GIRL, WOMAN, LADY, FEMALE, MISS). The rhyme sets were words similar in sound (e.g., CHAIR, PAIR, STAIR, SHARE, CARE). The unrelated sets were words unrelated in meaning or sound (e.g., LAMP, BALL, SHIRT, KNIFE, KING).

Procedure. In the first phase, participants were presented with a series of study/test trials. Each trial began with the presentation of a prompt in the centre of the screen for 1.5 seconds. This prompt was the word "synonyms," "rhymes," or "unrelated." Participants were told that what the five words had in common might help them be quicker and more accurate for the test. Following the prompt, five study words (the study set) were presented one at a time immediately followed by 10 test probes (the five study words and five unrelated foils in random order). The presentation order of study sets was intermixed and random. All the foils were unrelated in meaning, sound, and orthography to the study sets. As in Experiment 1, recognition memory for the foils was tested in a final phase. In all phases of the experiment, order of presentation of words within a phase was random.

Results

As in Experiment 1, recognition memory, corrected for guessing by subtracting false alarms from hits, on the immediate memory task was at ceiling and was equivalent for all conditions (M = .99 for all conditions), $F < 1$.

More important, differences in foil memory emerged. There was an overall main effect of study set, indicating that foil recognition differed across conditions, $F(2, 34) = 3.415$. As predicted, foils following a meaning-based set were better recognized than foils following a nonmeaning-based rhyme set, $t(17) = 2.716$. This conceptual replication of Experiment 1 with different stimuli, experimental parameters, and forms of similarity provides further support for the notion of retrieval depth driven by the nature of the similarity of the study items. Recognition for the foils following an unrelated set did not differ reliably from the other conditions (ps > .13). We speculate that this result likely reflects earlier rejection based on a mixture of meaning and nonmeaning-based processes.

General Discussion

Two experiments provided support for similarity-guided depth of retrieval. The shared similarity of the study sets constrained, up front, the subsequent interpretation of test probes during the immediate memory test. Testing subsequent foil memory provides a means of gauging differences in depth of retrieval during the initial test. This method revealed that meaning-based constraints afforded greater depth of retrieval, compared to constraints based on orthography (Experiment 1) or phonology (Experiment 2).

Our results are consistent with theories holding that performance in a variety of memory tasks, including recognition memory, rely on a match between reinstated context and the context that is retrieved using the probe as a cue (e.g., Dennis & Humphreys, 2001). However, we go beyond prior experiments in showing that a match in processing context is gained by means of constraining the processing of test items in a way that matches the dimension made salient during study. In this regard, our findings of effects on memory for foils converge with results reported by Rugg, Allan, and Birch (2000). They manipulated levels of processing during study and found differences in event-related potentials elicited at retrieval, a result consistent with our notion of differences in depth of retrieval.

The effect of study-list context on memory for foils is of the same sort as effects found by Brooks and his colleagues in their investigations of the importance of constraining possible interpretations in clinical settings. When interpreting a stimulus such as an electrocardiogram, it is sometimes advantageous to constrain the search for corroborative features at the outset (Norman et al., 1999). Similarly, when preparing to recognize a word whose meaning has been made salient by list context, it is advantageous to process meaning of test items from the outset, which necessarily involves processing the meaning of foils as well as that of targets, and enhances subsequent memory for the foils.

Individual differences in depth of retrieval may be an important source of variation in memory performance. Deficits in using more cognitively controlled bases of memory (e.g., recollection) may prevent older adults from constraining their retrieval to the same extent as done by young adults, resulting in age-related deficits in memory. We are currently using tests of foil memory to investigate the possibility of age-related deficits in retrieval depth. Tests of foil memory might also be useful for better specifying the basis for judgments in categorization and decision-making tasks by allowing one to diagnose the salient dimensions involved in rejection.

Let us end by trumpeting the general importance of memory for foils, which can be thought of as memory for the not chosen. When pondering back upon "life" decisions, what does one remember about alternatives that were not chosen en route to one's current lover, one's current job, etc.? Memory for the not chosen is telling with regard to the deciding factors of difficult decisions. What contributions does memory for alternatives that were not chosen make to the specific instances used to make future decisions? We end on this speculative note to remind Brooks of conversations with Jacoby, greatly enjoyed by Jacoby, in which they progressed (?) from discussion of a few findings to the building of "sky castles" regarding the broad implications of those findings for theory and, ultimately, for the meaning of life.

Please direct correspondence to Yujiro Shimizu or Larry L. Jacoby, Department of Psychology, Washington University, One Brookings Drive, Saint Louis, Missouri 63130 (E-mail: yshimizu@artsci.wustl.edu or lljacoby@artsci.wustl.edu).

References

Brooks, L. R. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 169-211). Hillsdale, NJ: Erlbaum.

Brooks, L. R. (1987). Decentralized control of categorization: The role of prior processing episodes. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 141-174). New York: USsity Press.

Brooks, L. R., LeBlanc, V. R., & Norman, G. R. (2000). On the difficulty of noticing features in patient appearance. *Psychological Science, 11,* 112-117.

Brooks, L. R., Norman, G. R., & Allen, S. W. (1991). Role of specific similarity in a medical diagnostic task. *Journal of Experimental Psychology: General, 120,* 278-287.

Craik, F. I. M.., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11,* 671-684.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108,* 452-478.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91,* 1-67.

Jacoby, L. L., & Brooks, L. R. (1984). Nonanalytic cognition: Memory, perception and concept learning. In G. H. Bower (Ed.), *The psychology of learning and motivation, Vol. 18* (pp. 1-47). New York: Academic Press Inc.

Jacoby, L. L., Shimizu, Y., Daniels, K., & Rhodes, M. G. (in press). Recognition and source memory and modes of cognitive control: Depth of retrieval. *Psychonomic Bulletin & Review.*

Kulatunga-Moruzi, C., Brooks, L. R., & Norman, G. R. (2001). Coordination of analytic and similarity-based processing strategies and expertise in dermatological diagnosis. *Teaching & Learning in Medicine, 13,* 110-116.

LeBlanc, V. R., Brooks, L. R., & Norman, G. R. (2002). Believing is seeing: The influence of a diagnostic hypothesis on the interpretation of clinical features. *Academic Medicine, 77,* S67-S69.

McDermott, K. B., & Watson, J. M. (2001). The rise and fall of false recall: The impact of presentation duration. *Journal of Memory and Language, 45,* 160-176.

Norman, G. R., & Brooks, L. R. (1997). The non-analytical basis of clinical reasoning. *Advances in Health Sciences Education, 2,* 173-184.

Norman, G. R., Brooks, L. R., Colle, C. L., & Hatala, R. M. (1999). The benefit of diagnostic hypotheses in clinical reasoning: Experimental study of an instructional intervention for forward and backward reasoning. *Cognition and Instruction, 17,* 433-448.

Rugg, M. D., Allan, K., & Birch, C. S. (2000). Electrophysical evidence for the modulation of retrieval orientation by depth of study processing. *Journal of Cognitive Neuroscience, 12,* 664-678.

Sommaire

Les interprétations classiques de la mémoire de reconnaissance mettent l'accent sur le rapport quantitatif entre la force ou la familiarité d'une sonde de mémoire en regard de quelque critère décisionnel. Elles font fi, en grande partie, des fondements qualitatifs sur lesquels la mémoire de reconnaissance prend appui pour porter des jugements. Nous soutenons, par contraste, que l'imposition de contraintes particulières à ce qui vient à l'esprit au cours de la récupération est susceptible d'infléchir les fondements qualitatifs qui font que d'« anciens » éléments sont acceptés et que de « nouveaux » éléments sont rejetés. Nous désignons par l'expression « profondeur de la récupération » l'application de divers fondements qualitatifs au cours de la récupération.

Nous rendons compte de deux expériences au cours desquelles la similitude des articles à l'étape de l'étude a imposé des contraintes à la façon d'évaluer des articles lors d'un test de mémoire à court terme, ce qui a donné lieu à des différences dans la profondeur de la récupération. Les deux expériences ont fait appel à une méthode novatrice qui éprouve la mémoire des exceptions afin d'évaluer les écarts dans la profondeur de récupération guidée par la similitude.

Au cours de la première expérience, 16 participants ont passé une série d'essais qui ont consisté en l'assimilation d'ensembles de quatre mots, suivie d'une seule question de sondage. Les ensembles à assimiler étaient composés de mots comparables soit par la sémantique soit par l'orthographe. Les exceptions (nouveaux mots ajoutés aux ensembles) n'avaient pas de rapport avec les ensembles à assimiler. Nous croyions que la nature des ensembles à assimiler influencerait l'évaluation des mots d'exception. À notre avis, ceux qui suivaient des ensembles comparables sémantiquement seraient rejetés en raison de leur sens, tandis que ceux qui suivaient des ensembles à orthographe comparable le seraient à cause de leur aspect. Au cours d'une étape ultérieure, la mémoire de reconnaissance des mots d'exception a été évaluée en présence de mots nouveaux à titre de leurres. Comme le traitement (profond) fondé sur le sens donne lieu le plus souvent à de bons résultats de reconnaissance, nous nous attendions à ce que les mots d'exception présentés pour la première fois à la suite d'un ensemble comparable par la sémantique soient reconnus plus facilement que ne le seraient ceux qui suivaient pour la première fois des ensembles comparables par l'orthographe. Pareille constatation fournirait la preuve d'écarts de profondeur de récupération. Comme nous l'avions prévu, la mémoire des mots d'exception était le plus fidèle là où les mots d'exception apparaissaient pour la première fois à la suite d'un ensemble de mots comparables par la sémantique. Les résultats précités ont été approfondis lors de la deuxième expérience, au cours de laquelle 18 participants se sont penchés sur des ensembles à assimiler composés de synonymes et de rimes. Conformément aux prévisions, la mémoire de reconnaissance était le plus fidèle là où les exceptions étaient perçues pour la première fois à la suite d'un ensemble de synonymes comparables par la sémantique plutôt que suivant un ensemble de rimes non comparables par la sémantique. Fait important, si les participants avaient simplement évalué la force de chaque sonde de mémoire au cours des essais d'assimilation-test, nous n'aurions eu aucune raison de prévoir un traitement différentiel des exceptions fondé sur la similitude des ensembles à assimiler ni, par conséquent, des différences dans l'encodage et la mémoire ultérieure des exceptions. Au contraire, les deux expériences ont fourni la preuve de différences dans les fondements qualitatifs mis au service de la récupération.