Evidence for a memory threshold in second-choice recognition memory responses

Colleen M. Parks* and Andrew P. Yonelinas†,‡,§

*Psychology Department, University of Nevada, Las Vegas, NV 89154-5030; and †Psychology Department, University of California, Davis, CA 95616

Communicated by Mortimer Mishkin, National Institute of Mental Health, Bethesda, MD, May 19, 2009 (received for review September 30, 2008)

A fundamental question in the study of cognition is whether memory strength varies continuously or whether memories sometimes fall below a threshold and fail completely. Previous studies examining this question have relied exclusively on 1 method—receiver operating characteristics—so in the current study, we addressed this issue by using a completely different approach. We tested memory for single items and for arbitrary associations (e.g., memory for random word pairs) by using a 4-alternative forced-choice test in which subjects either made a single choice or a first and a second choice. In item recognition, single- and second-choice scores were directly related, as expected if a continuous strength signal supported performance. In contrast, in associative recognition, single- and second-choice scores were found to be unrelated, as predicted by high-threshold theories. However, when the word pairs were encoded as single compound words rather than arbitrary associations, associative recognition appeared to rely more on a continuous strength process. The results support memory models that include both a continuous familiarity process and a threshold recollection process.

Does memory sometimes fail, such that the strength of a memory trace falls below one’s memory threshold? If so, then our attempts to discriminate between studied items and new items would be effectively reduced to random guesses. However, another possibility, consistent with signal detection theory (1, 2), is that memory never truly fails, but instead varies in a continuous manner such that the things we have encountered previously are simply more familiar on average than things that are new. In this way, memories never fall below a threshold, and thus even the very weakest of memory signals should be useful in discriminating between old and new items.

Memory theorists have debated the existence of a memory threshold for more than half a century (2, 3), and recent neurocomputational models have postulated that the hippocampus supports a threshold recollection process whereby qualitative information is retrieved with some probability, whereas the surrounding cortex supports a continuous memory strength process that underlies familiarity-based discriminations (4–6). However, the empirical evidence relevant to this debate has come almost exclusively from the analysis of receiver operating characteristics (ROCs), functions that relate correct and incorrect recognition memory responses as response bias varies. These functions are important because threshold theories predict linear ROCs, whereas signal detection theories predict curved ROCs (7). ROC results from humans and rats have provided evidence for a memory threshold when subjects are required to retrieve arbitrary associations from memory (8–11), but other studies have reported evidence more in favor of signal detection theory, particularly in tests of item recognition where subjects must discriminate between old and new items (12, 13). However, ROC shape can be influenced by various confounding factors, such as ceiling and truncation effects, that complicate the interpretation of those results (14, 15). What is needed is an alternative method that can arbitrate between these different theories.

To that end, we made use of a second-choice procedure in the current study. Specifically, we examined forced-choice recognition memory responses in which subjects either made a single choice to indicate which of 4 items was studied, or they made a first (most confident) and a second (next most confident) recognition choice. Of particular interest was the performance on the second choices when the first choice was incorrect. Signal detection models predict that there should be a positive relationship between single- and conditional second-choice accuracy because both responses are based on the same underlying strength distributions (Fig. 1A). That is, studied items are assumed to be more familiar, on average, than new items, and the subject is assumed to select the most familiar item as old. This will often lead to a correct first response, but because of the overlap in the distributions of old and new items, the strongest item will sometimes be a new item, making the first choice incorrect. The likelihood that the second choice will be correct in these cases depends on the average strength of the studied items—that is, the greater the average strength of the studied items, the more likely it is that the second strongest item will be the studied item. Thus, if the signal detection model is correct, then there should be a fairly strong relationship between single-choice and second-choice accuracy, such that subjects with high accuracy scores on their single choices should also have high accuracy scores on their conditional second choices (Fig. 1B).

In contrast, high-threshold models predict that there will be no relationship between single-choice and second-choice accuracy. These models assume that there is some probability that a studied item will exceed a memory threshold. If this occurs, then the subject is expected to select that item and make a correct recognition response. If, however, memory fails and the studied item does not exceed the memory threshold, then the subject must effectively guess which of the test items was studied. Thus, when the first choice is incorrect, it means that the studied item did not exceed the memory threshold and the first response was based on a guess; that entails that the second choice will also be based on a guess. Thus, performance on second choices should be near chance, regardless of how good (or bad) single-choice accuracy is. As a result, second-choice performance should not be correlated to single-choice performance, because single-choice performance reflects memory strength above the threshold, whereas second-choice performance reflects memory strength below the threshold. Note that second-choice performance need not be perfectly at chance if there is a second (low) threshold below which some new items can fall (i.e., a 2-high-threshold model). The influence of a low threshold is reflected

Author contributions: C.M.P. and A.P.Y. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

1To whom correspondence may be addressed. E-mail: colleen.parks@unlv.edu or apyonelinas@ucdavis.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0905050106/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.0905050106
in recall-to-reject strategies in which subjects reject lures based on memory of original pairings from study (e.g., D is rejected as a lure for the cue A because the subject remembers that D had been paired with C). We designed the following experiments to minimize the usefulness of such a strategy by pairing some of the items in the study list with multiple different items (see SI for details).

Based on prior ROC studies, we expected that a signal detection-based familiarity process would be useful in standard old/new item recognition, and thus we expected a strong relation between single- and second-choice performance across subjects. In contrast, in recognition tests in which subjects are required to remember arbitrary associations, performance is expected to rely more heavily on recollection because both items are familiar. Thus, if recollection is a threshold process, then there should be little or no relationship between single- and second-choice performance. In a second experiment, we examined only associative recognition performance but manipulated whether the pairs of items were encoded as 2 separate words or as a single compound word. If a signal detection process is useful in supporting memory for items, then it should contribute more to associative recognition when the pairs of words are treated as single compound words (i.e., as single items). Thus, single- and second-choice performance should be more strongly related for pairs encoded as compound words than for those encoded as separate words.

We tested these predictions by presenting subjects with word pairs to study under elaborate encoding conditions. Each word was paired with 4 different targets over the course of the study phase to reduce recall-to-reject strategies. After the study phase, subjects select the strongest item as having been studied. In this example, the word “far” was studied and then appeared in a 4-alternative, forced-choice test with 3 lures: “cat,” “see,” and “pit.” On average, the studied items will be stronger than the nonstudied items, but because the strength distributions overlap, a nonstudied item will sometimes have higher memory strength than the studied item, and thus the subject will incorrectly select a nonstudied item first, and in this case select the studied item as a second choice. (8) For signal-detection theory, subjects with better recognition memory (i.e., greater distance between the studied and nonstudied distributions) will exhibit better performance on single-choice and second-choice responses, so single- and second-choice performances should be directly related (see SI for additional details of the model predictions).

Results

Performance was measured as the proportion of correct single- and first-choice responses, and the proportion of correct second-choice responses given that the first choice was not correct. Chance performance was 0.25 (1 out of 4) for first and single choices, and 0.33 (1 out of the remaining 3) for second-choice performance. For the associative test, the average proportions of correct scores were 0.51 (SD = 0.15), 0.53 (SD = 0.17), and 0.38 (SD = 0.10) for the single-, first-, and second-choice responses, respectively; for the item test, the scores were 0.53 (SD = 0.15), 0.51 (SD = 0.17), and 0.42 (SD = 0.10). All scores were greater than their respective chance values (all P < 0.05). A comparison of the single- and first-choice scores indicated that there was no difference between the item and associative scores, no difference between the single- and first-choice scores, and no interaction between these 2 factors (P > 0.05). Thus, performance was comparable in the item and associative recognition tests, and there were no differences between single- and first-choice responses.

The relationship between the single-choice and second-choice trials was examined by plotting these scores for each subject (Fig. 2A). (A subsequent analysis directly contrasted first- and second-choice scores, and it led to the same pattern of results.) In the item-recognition test, there was a positive correlation between single- and second-choice accuracy (r = 0.56, P < 0.001, slope = 0.339, P < 0.001), consistent with signal detection theory. In contrast, in associative recognition, no correlation was found between single- and second-choice accuracy (r = −0.07, P = 0.68, slope = −0.047, P < 0.70). Most importantly, the correlation in the item test was significantly greater than that seen in the associative test (z = 2.44, P < 0.008). The results are consistent with the claim that item recognition is strongly influenced by a signal detection process, whereas associative recognition is more dependent on a threshold process.

Several alternative accounts of the results can be ruled out. First, it is possible that, despite our instructions, subjects may have distributed their most confident responses among their first and second choices, thus increasing their second-choice scores and decreasing their first-choice scores. However, the scores in the single-choice trials did not differ from those in the first-choice trials, indicating that the requirement to make a second choice did not depress subjects’ performance on their first choices. Second, the low correlation observed in the associative recognition test could arise if performance was more variable in that test than in item recognition. However, the standard deviations of the accuracy scores were the same in the item and associative tests, indicating that the differences in correlations could not be due to differences in variability. Third, the associative test had fewer test trials than the item test, and this difference may have influenced the observed correlations. However, in a subsequent analysis, we controlled for number of trials by scoring only the first 90 item-recognition test trials (the same as the number of trials for the associative test), and the results were unaffected, indicating that the observed...
Fig. 2. Scatterplots of single-choice versus second-choice accuracy. (A) The relationship between single- and second-choice accuracy for item and associative recognition. (B) The relationship between single- and second-choice accuracy after separate or compound word encoding in associative recognition.

differences could not have been due to differences in the number of trials. Finally, differences in the ratios of new to old item variance could potentially contribute to differences in correlations found for the 2 tests; however, simulations with the signal detection model using different strengths and variance ratios indicated that variance differences had little effect on the expected correlation (see SI).

The results suggest that when subjects retrieve arbitrary associations, a threshold process is involved, whereas retrieval of single items relies more on a signal detection process. To test this further, we conducted another experiment in which we examined only associative recognition but manipulated whether the pairs of items were encoded as 2 separate words or as a single compound word. If a signal detection process is useful in supporting memory for items, then it should also begin to contribute to associative recognition, as long as the pairs of words are treated as single compound words (i.e., as single items). To manipulate processing of the words as either arbitrary pairs of items or as meaningful single units, 40 subjects encoded word pairs by rating how well each word fit into a sentence frame (e.g., Feeling full of ___ , he untied the ___ . for the pair shame ribbon), and another 40 subjects encoding the pairs by rating how well a novel definition defined the new compound word created by the pair (e.g., an embarrassing ornament worn as punishment: shame ribbon). If a signal detection process can support memory for items, then single- and second-choice performance on associative recognition should be strongly related when the pairs are encoded as single compound words, but they should not be when the pairs are encoded as separate words.

Accuracy scores in the sentence condition were 0.57 (SD = 0.19), 0.55 (SD = 0.17), and 0.36 (SD = 0.19) for the single-, first-, and second-choice responses, respectively; in the compound condition, the scores were 0.53 (SD = 0.17), 0.55 (SD = 0.16), and 0.39 (SD = 0.10). First- and single-choice scores were greater than chance (P < 0.05); second-choice scores were greater than chance in the unitize condition but only marginally greater than chance in the sentence condition (P < 0.001 and P < 0.08, respectively). Neither of the main effects (study condition, response type) nor the interaction was reliable (P > 0.20), indicating that performance was comparable in the sentence and unitize conditions, and that there were no differences between single- and first-choice responses.

More importantly, a significantly positive correlation between single- and second-choice performance was observed in the compound condition (r = 0.39, P < 0.02, slope = 0.215, P < 0.02) but not in the sentence condition (r = 0.13, P = 0.44, slope = 0.072, P < 0.45). Similar to the pattern found in the first experiment, the correlation was significantly greater in the compound than in the sentence condition (z = 1.89, P < 0.03). Thus, in a test of associative memory for pairs encoded as single items, performance was consistent with signal detection theory. However, when pairs were encoded as separate items, performance was more consistent with threshold theory.

Discussion

We examined single and second choices in forced-choice recognition tests to determine whether memory judgments reflected the operation of a memory threshold or a continuous signal-detection process. The second-choice method we used here has been used in perceptual discrimination tasks and has indicated that there is no evidence for a threshold in visual perception (1), but as far as we know it has never been applied to the study of memory. Our results provide evidence that memory for arbitrary associations relies heavily on a threshold process, whereas memory for single items relies more on a signal-detection process. In a recognition test for single items or for pairs that were encoded as single items, single-choice performance was strongly related to second-choice performance, indicating that memory strength information contributed to accurate performance for both types of responses. This is consistent with predictions of models that include a continuous signal-detection process. In contrast, in recognition tests for arbitrary associations, second-choice scores were not related to single-choice scores, as expected if single choices reflected memories that exceeded a memory threshold and second choices reflected memories that fell below that threshold.
The current results are consistent with several previous ROC studies that have suggested that item recognition relies heavily on a signal-detection-based familiarity process, whereas associative recognition relies more on a threshold recollection process (15). The current results, however, are not susceptible to the same measurement concerns related to quantifying the shape of ROC functions. Together, these second-choice results and the previous ROC results provide strong support for current models of the medial temporal lobe that assume that the cortex supports a signal-detection familiarity process that is useful in item recognition, whereas the hippocampus supports a threshold-recollection process that is particularly useful in supporting associative recognition (4–6). By these models, the threshold nature of recollection arises because of the sparse coding and strong inhibitory connections observed within the hippocampus. These models are further supported by lesion and neuroimaging studies indicating that the hippocampus plays a particularly important role in supporting recollection, whereas the perirhinal cortex supports familiarity (8, 16–18).

The current results indicate that human recognition memory cannot be characterized as a pure signal-detection process or as a pure threshold process. Rather, both types of processes contribute to performance, and how they contribute is governed in part by the kind of information tested and how that information was originally processed. These results argue against single- and dual-process models that are based solely on signal-detection processes, because those models assume that memory strength varies continuously and never falls below a threshold (19, 20)—thus, memory can never fail in such models, which it was shown to do here. One way to bring those models in line with the current results is to introduce a second process that has a threshold and thus does categorically fail or succeed, such as recollection or attention (3, 6, 21). Another possible approach is to assume that associative recognition relies on a threshold process that plays a role only in associative tests, whereas item recognition relies on a signal-detection process that plays a role only in item-recognition tests. However, there are several problems with this approach. For example, it is now well established that both recollection and familiarity contribute to item-recognition tests (22, 23), and the second experiment in the current study provided evidence for a threshold process in one associative test condition and a signal-detection process in another. In addition, in the current associative conditions, second-choice performance was slightly above chance, which is problematic for a pure high-threshold model. Slightly above chance scores could arise if some proportion of the studied pairs were encoded as single words, thus allowing familiarity to contribute to performance. In addition, simulations of the pure signal detection model (see SI) produced single vs. second choice slopes of 0.62–0.74, which were much higher than those observed in the current experiments (i.e., 0.05–0.34), whereas a hybrid model that included both a threshold and signal detection process was capable of producing results more in line with the observed slopes.

The current results are important in highlighting the fact that memory tests do not provide pure measures of memory retrieval processes, in the sense that the associative tests were sometimes found to rely on the same processes supporting item recognition. These results may help to resolve current controversies about whether patients with hippocampal damage exhibit comparable deficits in item- and associative-recognition tests (24, 25). That is, although hippocampal lesions may often disrupt associative recognition more so than item recognition, the extent to which these tests are differentially sensitive to hippocampal damage will depend on various factors, such as how the stimuli are initially encoded. Thus, item-recognition tasks may be disrupted by hippocampal damage if the task relies heavily on recollection and, conversely, associative recognition may not be profoundly disrupted if it relies heavily on familiarity. The second-choice procedure examined here may prove useful in future lesion studies by providing a method for assessing the contributions of different underlying processes to these different types of memory tests.

The second-choice procedure provides a nice alternative to ROCs for testing memory models, but of course it is subject to methodological limitations that may render the procedure less appropriate in some experimental designs. First, to obtain a sufficient number of second choices that follow an incorrect first choice, it is necessary to collect a large number of responses from each subject. Second, to accurately characterize the relationship between single- and second-choice scores, it is critical that overall level of single-choice accuracy is neither too high nor too low, and that there is a broad range of scores across subjects. In the current experiments, the subjects’ scores ranged from 0.30 to close to 0.90—a more restricted range may have limited the ability to detect correlations. Finally, to use the procedure to test for a memory threshold, some critical assumptions must be met, and it may be difficult to satisfy these in some experimental contexts. For example, the predictions of the high-threshold model rely on the assumption that subjects are not using recall-to-reject strategies to eliminate lures from the response options. We minimized the effect of such a strategy by pairing study items with multiple other items. The procedure also assumes that the experimental design will lead to little to no false recollection. If subjects often falsely recollect a large number of lure items, then their single choices may reflect false recollections, and thus their second choices may become related to first choices.

In sum, the results suggest that memory for single items never completely fails, such that even weak feelings of familiarity can serve to help make accurate judgments about the past, whereas the recollection of arbitrary associations often categorically fails. These results indicate that we may have memory traces for all of the individual items or stimuli that we encounter, but that our ability to retrieve the associations that make up detailed memories of specific events can completely fail. Future studies will now need to determine whether these memory failures are due to interruptions of encoding or retrieval. Some models have suggested that memory failures arise as a result of inattention at encoding (26), whereas others have suggested that failures can also arise because of forgetting or interference at the time of retrieval (4, 5). In addition, it will be critical in future studies to examine how these results generalize to other, more real-world, contexts because they have implications for areas of courtroom eyewitness testimony and related fields of false memories and recovered/repressed memories by showing that for some types of information, individuals’ memory guesses can hold important information about past events, whereas for other types of information, weak memory traces hold little accurate information about the past.

Methods

In experiment 1, 80 subjects studied 360 arbitrary cue–target word pairs (e.g., rose theorem) for 6 seconds and then rated the difficulty of creating an image or sentence including those words on a 4-point scale. Each cue word was paired with 4 different targets across the study list (e.g., rose theorem, rose death, rose duration, rose challenge). After study, half of the subjects were given a 4-alternative, forced-choice item-recognition test; each trial consisted of 1 studied target word and 3 lure words on each trial. A total of 120 trials were single-choice trials, and 240 were second-choice trials. The other 40 subjects were given a 4-alternative, forced-choice associative recognition test in which 1 cue word was presented along with 4 target words; all targets had been studied, but only one had been paired with the cue word. Each cue was tested only once in the associative test so as to reduce recall-to-reject strategies. A total of 30 trials on the associative test were single-choice trials, and 60 were second-choice trials. In experiment 2, 80 new subjects encoded words in the sentence or unitize condition. All subjects then took an associative-recognition test designed in the same manner as in experiment 1. Details on materials and methods are presented in the SI.

Acknowledgments. This work was funded by National Institute of Mental Health Grant MH59352.
2. Egan JP (1958). Recognition memory and the operating characteristic. USAF Operational Applications Laboratory Technical Note (Indiana University Hearing and Communication Laboratory, Bloomington), No. 58-51, ii.