Current Psychology (2023) 42:18163–18177
https://doi.org/10.1007/s12144-021-02615-5

Explaining short-term memory phenomena with long-term memory theory: Is a special state involved?

Michael S. Humphreys[1] · Gerald Tehan[2] · Oliver Baumann[3] · Shayne Loft[4]

Accepted: 7 December 2021 / Published online: 25 March 2022

Abstract

The idea that some recently encountered items reside in a special state where they do not have to be retrieved has come to be a critical component of short-term memory theories. In the current work, the existence of such a special state was tested using the probe-recognition paradigm followed by a delayed recognition test. Across two experiments participants received a series of probe recognition trials where list lengths of 1-, 4- and 8-items were intermixed. Delayed recognition performance for non-target probes was poorer for the only item in 1-item lists than for the last item in multi-item lists. At the same time, the delayed recognition of studied-but-not probed items was better for the 1-item list, compared to the last item in a multi-item list, indicating that some form of a retrieval effect was involved and not lower levels of attention/initial learning. An examination of the size of the testing effect as it varied across list lengths and experiments also indicated that residence in a special state was not playing an important role. Overall, the data are not in support of the assumption that items at the focus of attention are in a special state that do not require retrieval. Our conclusions are that special states cannot be used to define STM memory and that the probe recognition paradigm may be useful in determining how testing affects memory.

Keywords short-term memory · probe recognition · long-term memory; episodic memory · testing effect

Humphreys et al. (2020b) asked how much of the results from short-term memory (STM) paradigms could be explained using ideas about long-term memory (LTM), especially the retrieval processes employed in LTM. They addressed the problems of how the information stored in the cortex could contribute to both episodic memory and retrieval over short intervals, and described a modern theory of associative interference that incorporated ideas about distributed representations and a role for context. They then applied those ideas to the areas which seemed most likely to differentiate between STM and LTM. These included the closely linked ideas about immunity to proactive interference (PI) and capacity limitations (Atkinson & Shiffrin, 1968; Cowan, 1995, 2001), the evidence about interference and decay (Nairne, 2002), and the role of articulatory and phonological information at short and long retention intervals (Oberauer et al., 2018). Humphreys et al. (2020b) concluded that many aspects of retention over short intervals could be explained by a theory of LTM. However, there were some areas such as the retention functions for articulatory and phonological information which probably required supplementation with a more specific STM theory. One of the critical theorised distinguishing aspects between STM and LTM required further experimental investigation, which forms the focus in the current paper.

Atkinson and Shiffrin's (1968) conception of a limited-capacity short-term store has largely given way to ideas about a limited number of items residing in the focus of attention (Cowan, 1995, 2000). In Cowan's model items in the focus of attention are immediately accessible and do not need to be retrieved. In his words these items "are, in a sense, already retrieved; they reside in a limited-capacity store, eliminating the retrieval step in which PI arises"

✉ Michael S. Humphreys
mh@psy.uq.edu.au

✉ Shayne Loft
shayne.loft@uwa.edu.au

[1] School of Psychology, The University of Queensland, Brisbane, Australia

[2] The University of Southern Queensland, Toowoomba, Australia

[3] Bond University, Gold Coast, Australia

[4] School of Psychological Science, The University of Western Australia, Crawley, WA 6009, Australia

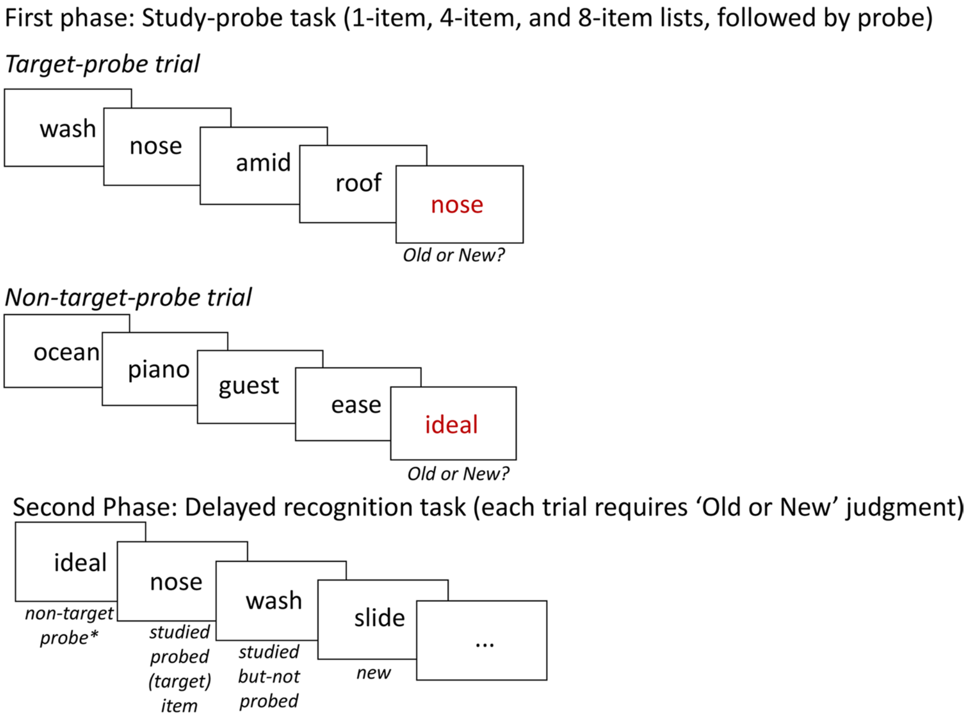

First phase: Study-probe task (1-item, 4-item, and 8-item lists, followed by probe)
Target-probe trial
wash
nose
amid
roof
nose
Old or New?
Non-target-probe trial
ocean
piano
guest
ease
ideal
Old or New?
Second Phase: Delayed recognition task (each trial requires 'Old or New' judgment)
ideal
non-target probe*
nose
studied probed (target) item
wash
studied but-not probed
slide
new
...

Fig. 1 Sequence of events and trial types for both experiments (*Non-target probes were only employed in Experiment 1). In the study-probe task, each study item was presented for 800 ms, with a 200 ms blank screen between successive words. The last studied word from each list was followed (after a further 200 ms blank) by a single probe word (presented in red font). The probe word remained on the display until participant response. In the delayed recognition task words remained on the display until the participant responded

(Cowan, 2001, p. 103). Thus, Cowan argues that immunity to PI is a signature characteristic of a limited capacity store. In line with this, Cowan claims that there is immunity to PI with immediate tests of lists of four items (or fewer) but not with longer lists (see Halford et al., 1988 and Wickens et al., 1981 for evidence consistent with this position; and see Beaudry et al., 2014 review of short-term PI effects which is not entirely consistent with this position).

The goal of the current paper was to re-examine the question of whether an item which would be considered as being in a short-term store, or in the focus of attention, needs to be retrieved. This idea of re-examining whether items regarded as residing in the focus of attention needed to be retrieved originated from prior work on prospective memory (PM). Humphreys et al. (2020a) (also see Loft & Humphreys, 2012) found that a multi-target, or categorical, PM task improved the delayed recognition of non-target items in an ongoing task more than a single-target PM task. Specifically, Humphreys et al. (2020a) examined the effect of a PM task on the delayed recognition of non-target (i.e., non-PM) items in the ongoing task in which PM items were embedded. Participants either completed a lexical decision or a word naming task (ongoing task), and separate groups of participants were additionally required to remember to press an alternative response key if presented on the ongoing task with a single target word, a member of a category (e.g., any *fruit* word), or one of multiple target words. After completion of the ongoing-PM task, delayed recognition of the non-target items from the ongoing task was tested. Performance on delayed recognition of non-target items was better if participants were previously attempting to detect an item from a list of multiple PM targets, or PM targets of a category, compared to a single-item PM target or if they had no prior PM task and had completed the ongoing task only. That is, there were long-term effects on items that had not been studied, but only tested for membership in the PM target set. Further, recognition of non-target items was very similar between the single-target PM and control condition. These outcomes could reflect that at times the single PM target was in a readily available state at the time of the ongoing-PM task because of the number of learning trials and rehearsal during the ongoing task (Strickland et al., 2022). The ready availability of the single PM target might mean that participants were overconfident in their ability to recognize the target which may have caused less complete processing of the non-target words during PM retrieval, resulting in non-target words being less well recognized on the delayed recognition test. This change in some aspect of non-target word processing during the Humphreys et al. (2020a) PM task paradigm may have implications for the role of a special state in STM, as outlined further below.

To afford a more focussed investigation of the issue of whether an item which would be considered as being in a short-term store or in the focus of attention needs to be retrieved, rather than using a PM task, in the current studies we used a classic probe recognition task (Sternberg, 1966) followed by a delayed recognition test. As illustrated in Fig. 1, in the probe recognition task a serial list of words is followed immediately by a probe consisting of one studied word (target probe) or new word (non-target probe). The current experiments intermixed list lengths of one, four and eight items. After a series of these probe recognition trials there was a delayed recognition test which included old words and new words (Fig. 1). The old words on the delayed recognition task include studied probed words (i.e., target probes) on the probe recognition task, non-target probes, and studied but-not-probed words from the same serial position as the probed words.

Our primary focus in Experiment 1 was on showing that after studying the only item in a one-item list, non-target probes for 1-item lists would not be as well recognized on the delayed recognition task as non-target probes for longer lists. This result would be in line with the findings of Humphreys et al. (2020a). There appear to be two alternatives as to why there might be a reduction in the probability of recognizing a non-target probe on the delayed recognition test (change in retrieval parameters or special state plus diversion). In addition, there is a third alternative (special state) which does not directly address the delayed recognition of the non-target probes but is relevant to the testing effects that we observed in Experiment 1 and then replicated and extended in Experiment 2.

Change in retrieval parameters

The first explanation for the delayed recognition of the non-target probes is that the recognition process stays the same for the different list lengths but one or more of the parameters of the recognition process change following participant identification of the list length (before the retrieval process starts). According to evidence-accumulation models of cognition, as applied to probe recognition, there is a race to respond old or new (Strickland et al., 2022). Evidence accumulates about each response during the race until one of the two responses (old or new) passes a response threshold. Overconfidence in one's ability to perform the task could result in the lowering of the response threshold for both the old and the new response when the participant realizes that only a single word has been studied prior to the presentation of a probe (this is possible in the current studies because the probe is presented in a different color than study list words). Such a shift in response thresholds, corresponding to a more lenient criterion for responding, would result in a less thorough or less complete processing of both the target and non-target probes.[1] In turn this will produce poorer delayed recognition performance for the non-target probes from the 1-item lists than from the multi-item lists. It should also produce poorer probe recognition for the target probes from 1-item lists than target probes from the last position in multi-item lists. However, due to a ceiling effect it may not be possible to observe this effect. Note that in this proposal "availability" can refer either to a property of the individual item or the strength of an association such as a context–to-item association. In contrast residence in a special state or focus of attention is a property of the item not a context-to-item association. Furthermore, there is no special state as items would simply vary in the level of availability.

Items resident in a special state with or without diversion

In deciding how residency in a special state could affect delayed recognition of non-target probes it is necessary to consider just how a probe that matches an item that is in the special state can produce a *yes* response. Here, there seems to be no alternative to assuming that some sort of retrieval process is involved, though it need not be a retrieval process that is subject to PI. That is, the probe in interaction with the item in the special state must return a signal that can be used to decide that the probe had been studied (Humphreys et al., in press). The reason for calling this a retrieval process is because there is a degree of flexibility in producing the response. For example, in immediate serial recall participants tend to respond with the first item from the study list, not the item which is most likely in a special state (the last item). Duncan and Murdock (2000) have also found that when participants are post-cued for either probe recognition or serial recall the serial position curve for probe recognition flattens. Again, the item which is most likely to be in a special state is not necessarily the item which is produced first.[2]

The alternative to our first account of delayed recognition (change in retrieval parameters) is that the retrieval process starts in the same manner, with the same retrieval parameters, regardless of whether the study list is 1-, 4-, or 8-items in length. However, the presence of one or more items in the focus of attention may divert the retrieval process. In this scenario, with a 1-item list the participant says *yes* if the probe is identified as old because it was in the focus of attention, and new otherwise. This strategy will not work if there are also old items which are not in the focus of attention so the participant would have to continue with the

[1] Although the overconfidence would be produced by the ease of recognizing the highly available target it is possible that the ease of recognition would be attributed to the task as a whole not just the ease of recognizing the target.

[2] Duncan and Murdock (2000) argued that post cueing resulted in a change in the memory structure which was stored. However, it is also possible that it is a cueing effect, not a storage effect. Regardless by Cowan's (1995, 2001) account the last item in the list should be in the focus of attention regardless of the memory structure which has been stored.

normal retrieval process with multi-item lists. The strategy for 1-item lists would reduce the probability of delayed recognition of the non-target probes because they would not be processed as thoroughly as if the normal recognition process had run to completion. However, there should be no effect on probe recognition, assuming that probe recognition is nearly perfect whenever an item in the focus of attention is probed. The process of diversion would be a complicated process. It seems unlikely that participants would adopt such a process. Nevertheless, it needs to be kept in mind in evaluating the evidence that residence in a special store is not involved in our current results.

Testing effect

Our design also allowed us to look at the delayed recognition of both studied probed items and studied but-not-probed items from the same serial position. A version of the testing effect is obtained if we subtract the probability of recognizing a studied but non-probed item from the probability of recognizing a studied probed item from the same serial position.[3] Rose et al. (2014) observed that single items show larger testing effects if they are not in the focus of attention (i.e. when rehearsal is prevented by a secondary task) compared to a condition in which they are rehearsed. The initial learning consisted of studying a single item under either deep or shallow levels of processing. Recall of the single item on each list was tested after 10 seconds (initial recall). In one condition participants were free to rehearse the single item during that 10 seconds, but in the other two conditions the retention interval was filled with either easy or hard (rehearsal-preventing) maths problems. Participants then recalled as many items from the lists as they could (final free recall). Initial recall was better in the rehearsal condition than in the two rehearsal suppression conditions. In addition, on initial recall, the levels of processing effect was larger for the hard compared to easy math problem condition, and in turn for the easy math condition compared to the free rehearsal condition. However, on the final free recall test, memory conditional on successful initial recall was substantially better for the condition where rehearsal was made difficult by the math problems, compared to the rehearsal condition, and there were strong levels of processing effects in all three conditions. Rose et al. (2014) argued that these findings indicate that the initial recall of the to-be-remembered word following the math task involved slower, cue-driven search and retrieval from LTM, producing a larger testing effect than in the rehearsal condition where the initial recall involved reporting the to-be-remembered item directly from the focus of attention.

The probe recognition task we use provides a different way to manipulate the likelihood that the tested item will be in the focus of attention. In turn this provides for far more control over the length of the retention interval. That is, more recently studied items are more likely to be in the focus of attention than less recently studied items. There is of course a dispute over the number of items which can reside in the focus of attention. Cowan (1995, 2000) argued for four whereas Jonides et al., 2008 and McElree, 2001 argued for one. Nevertheless, if we probe for the last item in a list that item should be in the focus of attention for all three list lengths (1, 4, and 8) except for minor failures due to fatigue or inattention. The magnitude of fatigue or inattention driven failures can be estimated via inspection of the probe recognition results and the delayed recognition of the studied but not probed items. Furthermore, the testing effect should increase as the number of items following the probed item increases.

If the last item in a list is equally likely to be in the special state regardless of the list length then there should be no difference in the testing effects for the last item in a list as a function of list length. That is, the last item from both 1-item and multi-item lists are in the same state prior to the presentation of the probe so the testing effect should be the same unless there has been some change in the parameters or process. The alternative account of how a retrieval parameter change could differentially affect the delayed recognition of non-target probes makes no prediction about the size of the testing effect. Furthermore, the previous prediction about the size of the testing effect increasing as the number of following items increases depends only on the assumption that the testing effect will increase as it becomes less likely that the item is in a special state, and the Rose et al. (2014) argument that items in a special state will produce a smaller testing effect. It is independent of the assumption that retrieval is diverted by the presence of items in a special state. That assumption seems to be required in order to predict that the delayed recognition of non-target probes is poorer for 1-item lists than for multi-item lists. The differential predictions that follow from our assumptions about items being in a special state (with or without diversion), or changes in retrieval parameters, are presented in Table 1.

As evident in Table 1, probe recognition performance is the only measure that can directly distinguish between the change in retrieval parameters versus special state (with or without diversion) accounts. However as highlighted in the General Discussion, a theory that makes no prediction for a given effect (e.g., special state account makes no prediction for recognition of non-target probes, and the change in retrieval parameters account makes no prediction for the

[3] The testing effect is traditionally calculated as the difference between recognizing an item that had been tested and an item which had received an additional study trial (but see Rose et al., 2014).

Table 1 Predictions about probe recognition of the last item, delayed recognition of non-target probes, and the size of the testing effect on the last item assuming only residence in a special state, residence in a special state plus a diversion of the retrieval process, and a common retrieval process with different parameter settings.

Theoretical Position	Predicted result			
	Poorer target probe recognition of the only item in 1-item lists compared to last studied item in multi-lists	Poor delayed recognition of non-target probes from 1-item lists compared to from multi-lists	Testing effect increases with the number of following items	Testing effect is different for the last item across list lengths
Special State	No	No prediction	Yes	No
Special state plus diversion	No	Yes	Yes	No
Change in Retrieval Parameters	Yes	Yes	No prediction	No prediction

testing effect), are by definition more incomplete. In addition, while testing effects alone cannot be used to directly compare the change in retrieval parameters versus special state (with or without diversion) accounts, the testing effects are informative from a theoretical perspective in that they provide a further level of support (or not) for the special state theory assumption embedded in two of the three theoretical accounts in Table 1.

Experiment 1

A primary focus of Experiment 1 was to compare probe recognition for the only item in 1-item lists and the last item in multi-item lists. In doing so we subtracted the false-alarm rate from the hit rate in an effort to avoid a ceiling effect. We did not use *d*’ because of the (as expected) relatively large number of participants who had perfect scores for either hits or false alarms (see Table 3 in Results section). Another primary focus of Experiment 1 was the prediction of the change in retrieval parameters and special state plus diversion accounts that delayed recognition of non-target probes would be poorer following a 1-item list than following a multi-item list, indicating that following a 1-item list participants might overestimate the ease of recognition and employ a recognition process that is less effective than required.

As a check on our ability to control the retention interval we examined the delayed recognition of studied but-not-probed targets as a function of the number of following items in the list. The assumption here was that each following item provided an opportunity for rehearsal which, if it occurred, would increase the probability of delayed recognition.

A second aim of Experiment 2 was to examine whether the testing effect decreased when the lag between the study item and the test item decreased. Finding a difference in the testing effect for items in the same study position as a function of list length would provide partial support for the hypothesis that residence in the focus of attention controlled the testing effect. That is, an item studied in a given position from the beginning of the list would more likely be in the focus of attention the fewer number of items which followed it in the list.

We also wanted to see whether the testing effect was the same when the last item in a list was probed, regardless of the length of the list. If the testing effect on the last list item is invariant across list lengths it would indicate that residence in the focus of attention was the dominant or primary explanation for testing effects in probe recognition. Finally we wanted to see if our version of the testing effect was comparable to the standard version where the effect of a test is compared to the effect of an additional study trial. As a partial test of this equivalence we compared the delayed recognition of non-target probes, which are non-studied but tested items, with the delayed recognition of a studied but-not-probed item, which are studied but untested items.

Method

Participants Sixty undergraduates from the University of Western Australia participated in return for course credit. In Experiment 1 and Experiment 2, we required 60 participants to achieve 80% power to detect small-medium effect sizes. One participant was excluded because they did not follow task instructions, leaving a total of 59 participants for analyses. Study 1 and 2 were approved by the Human Research Ethics Office at the University of Western Australia. Informed consent was obtained from each participant.

Materials and procedure Four-hundred and eight medium frequency words (occurring 20-50 times per million, length 4-8 letters) were randomly selected from the 1994 issues of the SMHWD (Dennis, 1995). For each participant, 48 words were randomly selected to be used for the 1-item study lists, 96 for the 4-item study lists, 192 for the 8-item study lists, 48 words were used as non-target probes (24 to be presented

Table 2 Summary of the design of the study-probe trials and delayed recognition test for Experiments 1 and 2.

	Experiment 1		
Test	1-Item-List	4-Item List	8-Item List
Study-Probe Total = 96 study- probe trials	48 studied (target) items 24 target probes position 1 24 non-target probes	24 studies (target) items 6 target probes position 2 6 target probes position 4 12 non-target probes	24 studied (target) items 4 target probes position 2 4 target probes position 4 4 target probes position 8 12 non-target probes
Delayed Recognition Total= 120 trials (including 24 new items not presented during study- probe trials).	24 non-targets	6 studied probed position 2 6 studied probed position 4 6 studied probed position 2 6 studied probed position 4 12 non-target	4 studied probed position 2 4 studied probed position 4 4 studied probed position 8 4 studied but-not probed position 2 4 studied but-not probed position 4 4 studied but-not probed position 8 12 non-target
Experiment 2			
Test	1-Item-List	4-Item List	8-Item List
Study-Probe Total = 72 study-probe trials	12 studied (target) items 6 target probes position 1 6 non-target probes	24 studied (target) items 6 target probes position 1 6 target probes position 4 12 non-target probes	36 studied (target) items 6 target probes position 1 6 target probes position 4 6 target probes position 8 18 non-target probes
Delayed Recognition Total= 96 words (including 24 new items not presented during study- probe trials)	6 studied probed position 1 6 studied but-not probed position 1	6 studied probed position 1 6 studied probed position 4 6 studied but-not probed position 1 6 studied but-not probed position 4	6 studied probed position 1 6 studied probed position 4 6 studied probed position 8 6 studied but-not probed position 1 6 studied but-not probed position 4 6 studied but-not probed position 8

after 1-item lists, 12 after 4-items lists, and 12 after 8-item lists), and 24 words were presented as new items on the delayed recognition test. The number of different item types for the study-probe trials and the delayed recognition test for Experiment 1 and 2 are summarised in Table 2.

Participants were first presented ninety-six study-probe trials. Each word from the list was presented one at a time for 800 ms, with a 200 ms blank screen between successive words. In total, participants were presented 48 1-item study lists, 24 4-item study list, and 24 8-item study lists. The last studied word (studied words presented in black font) from each list was followed (after a further 200 ms blank) by a single probe word (presented in red font). The probe word remained on the display until participant response (the next study list was presented 200ms following the participant response). Participants were instructed to press the "O" key if they believed the probe word was presented in the study list, and the "N" key if they believed the probe word was not presented in the study list, as quickly and accurately as possible. The probe items either matched one of the items in that studied list (target probe) or did not match any item in the studied list (non-target probe). For the 1-item lists, 24 non-target and 24 target probes were presented. For the 4-item lists, 12 non-target probes, six target position 2 probes, and six target position 4 probes were presented. For the 8-item lists 12 non-target probes were presented, four target position 2 probes, four target position 4 probes, and four target position 8 probes[4].

Before the delayed recognition test participants completed a three minute Sudoku puzzle. On the delayed recognition test participants were presented a single word, which remained on the display until the participant responded. Participants were instructed to press the "O" key if they believed the word was presented during the study-probe phase, and the "N" key if they believed the word was not presented during the study-probe phase, as quickly and accurately as possible. One-hundred and twenty words were presented on the delayed recognition test. This comprised of 24 non-target probes from 1-item lists, 12 non-target probes from each of the 4-item and 8-item lists, six studied probed items (i.e, target probes) from each of position 2 and position 4 from the 4-item list, six studied but- not-probed items from each

[4] A reviewer pointed out that in Experiment 1 there were more probes of position 1 (24 target probes position 1) than there were for position 2 (10 target probes of position 2 across the 4-item and 8-item lists), position 4 (10 target probes of position 4 across the 4-item and 8-item lists), and position 8 (4 target probes of position 8 in the 8-item lists). If anything, this weighting likely increased the perceived value of studying1-items lists, providing a stronger test of the prediction of the change in retrieval parameters theoretical account that probe recognition of item lists would be poorer than probe recognition on the last item in multi-item lists.

Table 3 Hit and false alarms rates (probe recognition) for the Study-Probe blocks in Experiment 1. 95% within-subject confidence intervals are in parentheses (Cousineau, 2005). The % value in each cell indicates the percentage of participants with perfect hit (=1) or false alarm (=0) rates as a function of list length.

	FA Rate	Hit Pos 1	Hit Pos 2	Hit Pos 4	Hit Pos 8
1-item list	0.06 [0.05, 0.07] 33.9%	0.96 [0.93, 0.98] 45.2%	—	—	—
4-item list	0.04 [0.03, 0.05] 64.5%	—	0.85 [0.82, 0.89] 50%	0.98 [0.96, 1.0] 85.5%	—
8-item list	0.06 [0.04, 0.07] 51.6%	—	0.69 [0.64, 0.74] 22.6%	0.73 [0.68, 0.78] 32.3%	0.98 [0.96, 1.0] 88.7%

of position 2 and position 4 from the 4-item list, four studied probed items from each of position 2 position 4, and position 8 from the 8-item list, and four studied but-not- probed items from each of position 2, position 4, and position 8 from the 8-item list, and 24 new items that were not presented during the study-probes phase. The studied but- not- probed items presented on the delayed recognition test were from lists in which a non-target probe had previously been presented on the probe recognition task.

As describe above, 24 non-target probes presented for 1-item lists and 12 non-target probes each for 4–item and 8-item lists, as we wanted to ensure that as many non-target probes were presented for 1-item lists as there were for the multi-items lists when combined. In addition, we did not include studied probed (target) items from 1-item lists on the delayed recognition test. The rationale for these design choices was that our focus in Experiment 1 was on probe recognition and the recognition of non-target probes on the delayed recognition test and we wanted to maximize the power achieved for non-target probe delayed recognition differences as a function of list length (as we had maximum number of items we could present in the time allowed for the experiment sessions and for ensuring not too high task difficulty), whereas in Experiment 2 we focused on replicating the probe recognition findings and replicating and extending the testing effects.

We approximately equated the average retention interval across items (Humphreys et al., 2010) by presenting the study-probes in six blocks. Each block of six study-probe trials presented eight 1-item study lists, four 4-item study lists, and four 8-item study lists, and the associated probes for these studied lists. On the delayed recognition test, there were also six blocks. In Block 1 of the recognition test, four new items were presented along with the combination of eight non-target probes, four studied probed (target) items, and four studied but- not-probed items from Block 1 of the six study-probe phases. In Block 2 of the delayed recognition test, four new items were presented along with the combination of non-target, studied probed items, and studied but- not-probed items (16 in total) from Block 2 of the six study-probe phase, and so on.[5]

Results and discussion

Positive probe recognition on study-probe trials As would be expected, during the Study-Probe blocks (Table 3) participants were highly accurate at recognising probes for the 1-item list, and probe recognition for the 4-item lists and 8-item lists was highly accurate when the last item was being probed. Table 3 also presents the proportion of participants with perfect hit rates or perfect false alarm rates for each list length. We calculated the hit-false alarm rates for the study-probe items as a function of item position.

The change in retrieval parameters account, but not the two special state accounts, predict poorer probe recognition of the only item in 1-items lists than the probe recognition of the last item presented in multi-item lists. There was higher probe recognition of the 8th studied item from 8-item lists [M= 0.93 (95% CI, 0.90, 0.96)], compared to the probe recognition for the only item in a 1-item list [M= 0.90 (0.87, 0.93)], $t(58) = 2.11$, $p = .04$, $d = 0.28$ (within-subjects effect size; Morris & DeShon, 2002). Finding that probe recognition on the 8th item in an 8-item list is better than probe recognition of the only item in a 1-item list could indicate

[5] With four 8 item lists presented in each study-probe block it was not possible for each block to contain two non-target probe trials, and one studied probed (target) trial for each of position 2, 4 and 8 on the delayed recognition test. Instead, each study-probe block contained two non-target, and one studied probed position 2, 4 or 8, and one studied but-not-probed position 2, 4 or 8. We balanced it such that across the 6 blocks of study probe trials each target position for list length 8 occurred an equal (four) times. The studied but-not-probed items presented on the delayed recognition test were from lists in which a non-target probe had previously been presented and they were presented across the six blocks of the recognition test in the manner; such that across the 6 blocks of delayed recognition each studied-but-not-probed position for list length 8 occurred an equal (four) times.

Table 4 Hit rates for non-target probes in Experiment 1 for each list length. 95% within-subject confidence intervals are in parentheses (Cousineau, 2005).

List Length	Hit Rate
1 item	0.46 [0.44, 0.48]
4 items	0.55 [0.52, 0.57]
8 items	0.55 [0.53, 0.58]

that participants were anticipating the end of the 8-item list (Niemi & Näätänen, 1981). From the participant's perspective this is the only study item presentation that was never followed by another study item. However, this interpretation is weakened by the finding that probe recognition for the 4th item in a 4-item list [*M*= 0.94 (0.92, 0.97)] was better than probe recognition for the only item in a 1-item list, $t(58) = 3.12$, $p < .01$, $d = 0.41$.

Another explanation for the observed probe recognition performance is that participants were not as prepared to study the first item of any study list. That is, on some occasions the first item studied may not have entered into the focus of attention or into a special store. The test for this is to examine if delayed recognition for studied but not probed items from a 1-item list was poorer than delayed recognition for studied but-not-probed items from the last position from multi-item lists, but Experiment 1 did not include studied but-not -probed items from the 1-item list (we do include these items for study to test this explanation in Experiment 2). The final possibilities, and the ones that we favour, come from the idea that either some of the retrieval parameters are changing or that the presence of one or more items are in a special state are diverting the retrieval process. That is, there is a sub-optimal change in parameter settings or in the retrieval process.

Final delayed recognition On the delayed recognition test the mean false alarm rate was 0.31 [95% CI (0.27, 0.34]. The hit rates for non-target probes are presented in Table 4. The special state account makes no prediction regarding delayed recognition. In contrast the special state plus diversion, and the change in retrieval parameters account, predict poorer delayed recognition of non-target probes presented following 1-items lists compared to presented following multi-item lists. In line with this, the hit rate was higher for the non-target probes from the multi-item lists compared to non-target probes from the 1-item list, $t(58) = 5.57$, $p < .001$, $d = .53$. There was no difference in hit rate for non-target probes from the 4-item list compared to the 8-item list, $t<1$.

It appears that participants are overconfident in their ability to recognize after studying a 1-item list, and neglect some aspect of the memory retrieval process. This neglect of some aspect of the memory retrieval process was also confirmed in the probe recognition results. That is, the only item in a 1-item list was more poorly recognized on the probe recognition task than was the last item in the 4- and 8-item lists. However, this poorer probe recognition could have been due to inattention at the time of study. This will be tested in Experiment 2 where we will compare the delayed recognition of the only studied but-not-probed item in a 1-item list to the recognition of the last studied but-not-probed item in the multi-item lists.

In order to examine the possibility of displaced rehearsals enhancing delayed recognition when a study item was presented earlier in the probe-recognition study list we looked at final delayed recognition of the studied but-not-probed items that occupied the same serial position across the different list lengths (see Table 5). The probabilities of recognizing the studied but-not-probed words from position 2 in the 4-item and 8-item lists were not significantly different, $t<1$. The probabilities of recognizing the studied but-not-probed items from position 4 in 4-item and 8-item lists were also not significantly different, $t(58) = 1.19$, $p = .24$. Thus, there is no evidence that delayed recognition differs for items in the same serial position in different length lists, and thus we can assume that displaced rehearsals were not playing an important role in this task and that the retention interval is primarily determined by the position within the study list.

Testing effect We calculated the testing effect on the final recognition test by subtracting the probability of recognizing a studied but not-probed word from the probability of recognizing a probed studied word for the same list length and list position (see Table 6). The change in retrieval parameters account makes no prediction regarding test effects, whereas

Table 5 Hit rates for the studied but not-probed items on the delayed recognition test in Experiment 1. 95% within-subject confidence intervals are in parentheses (Cousineau, 2005).

	Studied but-not-Probed		
	Pos 2	Pos 4	Pos 8
1-item list	—	—	—
4-item list	0.47 [0.42, 0.52]	0.46 [0.41, 0.51]	—
8-item list	0.44 [0.38, 0.50]	0.51 [0.45, 0.56]	0.49 [0.45, 0.54]

Table 6 The testing effect (the probability of recognizing a studied but-not-probed word subtracted from the probability of recognizing a studied probed word as a function of list length and position within the list).

	Pos 2	Pos 4	Pos 8
4-item list	0.14 [0.08, 0.21]	0.10 [0.04, 0.16]	—
8-item list	0.26 [0.19, 0.34]	0.20 [0.13, 0.27]	0.04 [-0.03, 0.10]

the two special state accounts predict the testing effect will increase as the number of words studied following the probed studied word in the list increases. In line with this, the testing effect was significantly greater for position 2 in 8-item lists than for position 2 in 4-item lists, $t(58) = 2.22$, $p =.03$, $d = .39$. In turn, the testing effect was significantly greater for position 4 in 8-item lists than for position 4 in 4-item lists, $t(58) = 2.06$, $p = .04$, $d = .35$. Thus the testing effect is increased as the number of words studied following the probed studied word in the list increased. Overall, these results provide support for the hypothesis that residence in the focus of attention is driving the size of the testing effect because items in the same serial position from the beginning of the list are more likely to be in the focus of attention when the probe is presented if there are fewer subsequent studied items. However, as we will see later there are alternative explanations for this effect of the number of subsequent items.

In addition there was no significant difference in the size of the testing effect between the last studied items in 4-and 8-item lists, $t(58) = 1.04$, $p = .30$, supporting the prediction of the two special state accounts that the size of the testing effect for the last studied item should be invariant across list length.

The testing effect has been most commonly defined as a comparison between a delayed test on a tested item and a delayed test on a repeated item. In contrast we have compared delayed test performance on studied probed items and studied but non-probed items. However, there is one comparison we can make which is conceptually similar to the comparison between tested and restudied items. We compared the delayed recognition of non-target probes, which are non-studied but tested items, with the delayed recognition of a studied but-not-probed item, which are studied but untested items. Non-target probes were better recognised than the average recognition of studied but-not-probed items from 4-item lists (positions 2 and 4), $t(58) = 3.16$, $p = .002$, $d = .45$ The same comparison was made and same result found for 8-item lists (averaged across testing positions 2, 4 and 8), $t(58) = 3.34$, $p = .001$, $d = .35$ Thus a non-studied item which has been tested was better recognized on the delayed recognition test that an item studied one time.

Experiment 2

Experiment 2 was conducted to replicate Experiment 1 with a more thorough examination of the testing effect by increasing the number of observations in the 8-item lists and examining the testing effect in 1-item lists. Experiment 2 also served to establish the reliability of the Experiment 1 finding that target-probe recognition for the only item in a 1-item list was poorer than target-probe recognition for the last item in multi-item lists. Experiment 2 also served to determine whether delayed recognition of a studied but-not-probed item from a 1-item list is any different than delayed recognition of the studied but-not- probed items from the last position of the multi-item lists.

To achieve these aims, we equated the number of observations for the studied probed items and the studied but-not-probed items across the different list lengths. We had participants study 12 1-item lists (6 target probes and 6 non-target probes), 24 4-item lists (6 target probes for each of positions 1 and 4, and 12 non-target probes), and 36 8-item lists (6 target probes for each of positions 1, 4, and 8, and 18 non-target probes). In contrast to Experiment 1, the delayed recognition test did not include non-target probes. Instead, we tested studied but-not-probed words as well as the studied probed (i.e., target) words. This resulted in the same number of observations at each list position for each list length.

Failure to pay full attention to the first item presented would be indicated by two findings. First, probe recognition of the only item in a 1-item list would be poorer than probe recognition of the last item in multi-item lists (replicating Experiment 1). Second, final delayed recognition of a studied but-not-probed item from a 1-item list would be poorer than for the studied but-not probed-items from the last position of multi-item lists.

However, although we expected to replicate Experiment 1 with respect to the first above hypothesis, we did not expect to find support for the second hypothesis. Instead, we expected that there would either be no difference in delayed recognition for a studied but-not-probed item from a 1-item list compared to studied but-not-probed items from the last position of multi-item lists, or that the delayed recognition of the only but-not-probed item from a 1-item list would be slightly better than the recognition of the last but-not-probed item in multi-item lists. The same or better performance in delayed recognition, coupled with poorer performance in probe recognition, would indicate that participants adequately paid attention and studied the first presented study item, but overestimated the ease of probe recognition when only a single item had been previously presented. This combination of effects, poor probe recognition for 1-item lists coupled with comparable delayed recognition of the studied but-not-probed item from those lists, would uniquely support the hypothesis that some of the probe recognition retrieval parameters were changed when the participant realized that a 1-item list was being probed.

Method

Participants Sixty undergraduates from the University of Western Australia participated in return for course credit.

Materials and procedure Four-hundred and fifty six medium frequency words (occurring 20-50 times per million, length 4-8 letters) were randomly selected from the 1994 issues of the SMHWD (Dennis, 1995). For each participant, 12 words were randomly selected to be used for the 1-item study lists, 96 for the 24 4-item study lists, 288 for the 36 8-item study lists, 36 words were used as non-target probes (6 to be presented after 1-item lists, 12 after 4-items lists, and 18 after 8-item lists), and 24 words to be presented as new items on the surprise final delayed recognition test.

The Experiment 2 design is summarised in Table 2. Participants were presented 72 study-probe trials. Participants were presented 12 1-item study lists, 24 4-item study lists, and 36 8-item study lists. For the 1-item lists, 6 non-target and 6 target probes were presented. For the 4-item lists, 12 non-target probes, six target-position 1 probes, and six target-position 4 probes were presented. For the 8-item lists 18 non-target probes were presented, six target-position 1 probes, six target-position 4 probes, and six target-position 8 probes.

Ninety-six words were presented on the delayed recognition test. This comprised of six studied probed items from the 1-item list, six studied but-not-probed items from the 1-item list, six studied probed items probes from each of position 1 and position 4 from the 4-item list, six studied but-not-probed items from each of position 1 and position 4 from the 4-item list, six studied probed items from each of position 1, position 4, and position 8 from the 8-item list, and six studied but-not-probed items from each of position 1, position 4, and position 8 from the 8-item list, and 24 new items that were not presented during the study-probe phase.

We approximately equated the average retention interval across items by presenting the study-probes in 6 blocks. Each block contained two 1-item, four 4-item, and six 8-item study lists, and the associated probes for those studied lists. On the subsequent recognition test, there were also six blocks. In Block 1 of the recognition test, four new items were presented along with the combination of probed studied probed items and studied but-not- probed items (12 in total) from Block 1 of the six study-probe phase. In Block 2 of the recognition test, four new items were presented along with the combination of studied probed items and studied but-not-probed items (12 in total) from Block 2 of the six study-probe phase, and so on.

Results and discussion

Positive probe recognition on study-probe trials Performance on the Study-Probe blocks is presented in Table 7, including indication of the proportion of participants with perfect hit rates or perfect false alarm rates for each list length. During the Study-Probe blocks participants were highly accurate at recognising probes for the 1-item list, and probe recognition for the 4-item lists and 8-items lists was highly accurate when the last item studied was probed. As in Experiment 1, we calculated the hit-false alarm rates for the Study-Probe items as a function of item position.

Table 7 Hit and false alarms rates for the Study-Probe blocks in Experiment 2. 95% within-subject confidence intervals are in parentheses (Cousineau, 2005). The % value in each cell indicates the percentage of participants with perfect hit (=1) or false alarm (=0) rates as a function of list length.

	FA Rate	Hit Pos 1	Hit Pos 4	Hit Pos 8
1-item list	0.08 [0.06, 0.10] 65%	0.96 [0.93, 0.98] 80%	—	—
4-item list	0.04 [0.03, 0.05] 63.3%	0.83 [0.79, 0.87] 43.3%	0.97 [0.95, 0.99] 83.3%	—
8-item list	0.05 [0.03, 0.06] 46.6%	0.75 [0.71, 0.79] 28.3%	0.84 [0.81, 0.87] 36.7%	0.96 [0.94, 0.99] 80%

The change in retrieval parameters account, but not the two special state accounts, predict poorer probe recognition of the only item in 1-items lists than the probe recognition of the last item presented in multi-item lists (as was found in Experiment 1). The hit-false alarm rate for probes that matched (target probes) the 8th studied item from the 8-item list [M = 0.91, 95% CI [0.88, 0.94]) was not significantly higher than for probes that matched the 1-item list [M = 0.87, [0.83, 0.92]), t(59) = 1.59, p =.12, but importantly the effect was in the same direction as Experiment 1. Furthermore, probe recognition of the 4th studied item from the 4-item list [M = 0.93, [0.90, 0.96]) was reliably higher than the recognition of probes that matched the only item in a 1-item list, t(59) = 2.90, p < .01, d = 0.41.

Final delayed recognition On the delayed recognition test the mean false alarm rate was 0.34 [95% CI (0.30, 0.38]. Failure to pay full attention to the first item presented wold result in poorer delayed recognition of a studied but-not-probed item from a 1-item list than for the studied but-not probed-items from the last position of multi-item lists, but on the basis of the change in retrieval parameters account that the delayed recognition of the only but-not-probed item from a 1-item list should be slightly better than the recognition of the last but-not-probed item in multi-item lists (indicating that participants adequately paid attention and studied the first presented study item, but overestimated the ease of probe recognition when only a single item had been previously presented). As shown in Table 8, the studied but not-probed item from the 1-item list was better recognised than the studied but-not-probed last studied item from the 8-item

Table 8 Hit rates for the delayed recognition test in Experiment 2. 95% within-subject confidence intervals are in parentheses (Cousineau, 2005).

	Studied but-not-probed		
	Pos 1	Pos 4	Pos 8
1-item list	0.56 [0.52, 0.60]	—	—
4-item list	0.56 [0.51, 0.60]	0.48 [0.43, 0.53]	—
8-item list	0.52 [0.47, 0.57]	0.51 [0.47, 0.55]	0.47 [0.42, 0.52]

list, $t(59) = 2.53$, $p = .01$, $d = 0.33$, and the studied but-not-probed last studied item from the 4-item list, $t(59) = 2.23$, p =. 03, $d = 0.29$. In addition, the studied but-not-probed item at position 1 from the 4-item list was better recognised than the studied but-not-probed item at position 4 from the 4-item list, $t(59) = 2.23$, p =.03, $d = 0.29$. Comparison of the studied but-not-probed item at position 1 from the 8-item list to the studied but-not-probed item at position 8 from the 8-item list showed the same pattern, although this effect did not reach significance, $t(59) = 1.50$, p =.14.

Three out of the four comparisons across Experiments 1 and 2 showed that probe recognition for the only item in a 1-item list was poorer than probe recognition for the last item from multi-item lists, and the fourth comparison trended in the same direction. This is the opposite pattern than that observed for the delayed recognition test (only available in Experiment 2); three out of the four comparisons indicated that studied-but-not probed items presented in the first position across list lengths were better recognised than studied-but-not probed items from the last position in multi-item lists, and the fourth trended in the same direction. Overall, the results from Experiments 1 and 2 are consistent with the notion that participants adequately paid attention and studied the first presented items, but overestimated the ease of probe recognition when only a single item had been previously presented.

As in Experiment 1 we also looked at whether the delayed recognition of studied but- not-probed items differed as a function of the number of additional items presented in that list. A one way ANOVA on position 1 in the 1, 4, and 8 item lists produced a non-significant result, $F<1$. A similar result was found with position 4 items from the 4 and 8 item lists, $F<1$. This is the same non-significant pattern we found in Experiment 1, thus we conclude there was no evidence for a significant role for displaced rehearsals.

Testing effect We calculated the testing effect by subtracting the probability of recognizing a studied but-not-probed word from the probability of recognizing a studied probed (target) item for the same list length and list position (see Table 9). As in Experiment 1 we confined our analyses to the words which had occurred in the same list position across the different list lengths.

Table 9 The Testing effect (the probability of recognizing a studied but-not-probed word subtracted from the probability of recognizing a studied probed word) as a function of list length and position within the list in Experiment 2. The false alarm rate was .34.

	Pos 1	Pos 4	Pos 8
1-item list	0.02 [-0.05, 0.097]		
4-item list	.10 [0.03, 0.16]	.13 [0.05, 0.20]	
8-item list	.20 [0.14, 0.27]	.22 [0.16, 0.28]	.16 [0.09. 0.22]

The change in retrieval parameters account makes no prediction regarding testing effects, whereas the two special state accounts predict the testing effect will increase as the number of words studied following the probed studied word in the list increases. A one way ANOVA on the position 1 results was significant, $F(2,118) = 8.07$, $p < .001$. Additional comparisons showed that the testing effect for position 1 was not different for 1-item and 4-item lists but trended in the direction expected, $t(59) = 1.71$, $p = .09$, but was significantly different for position 1 between 4-item and 8-item lists, $t(59) = 2.30$, $p = .03$, $d = .40$. The difference in the testing effect for position 4 for 4-item and 8-item lists was nearly significant, $t(59) = 1.98$, $p = .05$.

The testing effect increased as the number of subsequent items in the list increased. Although only one out of the three comparisons was significant in Experiment 2, the other two comparisons trended in the same direction. In addition we had also observed in Experiment 1 that the testing effect increased with the number of subsequently studied items, so this finding seems reliable. These findings support the idea that the testing effect is reduced if the tested item is in focal attention (i.e., special state) when it is tested.

However, the lack of stability of the testing effect when it is calculated across the across the last items in list positions contradicts this conclusion. A one-way ANOVA (repeated measures) comparing the testing effect on the last item at each list length was significant, $F(2,59) = 3.95$, p =.02. The difference between list lengths 1 and 4 was significant, $t(59) = 2.08$, $p = .04$, $d = .35$, as was the difference between lengths 1 and 8, $t(59) = 2.68$, $p = .01$, $d = .49$. The difference between 4 and 8 was not significant, $t< 1$. The 4-item to 8-item list length difference was also the only difference that could be tested in Experiment 1 and it was not significant in that experiment either. It might be supposed that the observed testing effect for the 1-item list in Experiment 2 was spuriously low. However, Rose et al. (2014) had also found a very small testing effect in the rehearsal condition which was the one condition where they assumed that the tested item would be in the focus of attention. Thus from the Rose et al. perspective it is not surprising that the testing effect is small for the 1-item list. However, it is surprising that it was large for the last item in the 4- and 8-item lists. In addition, a between subject comparison of the testing effect on the last item in 8-item lists across Experiments 1 and 2 also found a significant difference, with a larger testing effect in Experiment 2, $t(117) = 2.4$, $p = .018$. Residence in the focus of attention cannot explain this variability in the size of the testing effect when the last item in the list is tested.

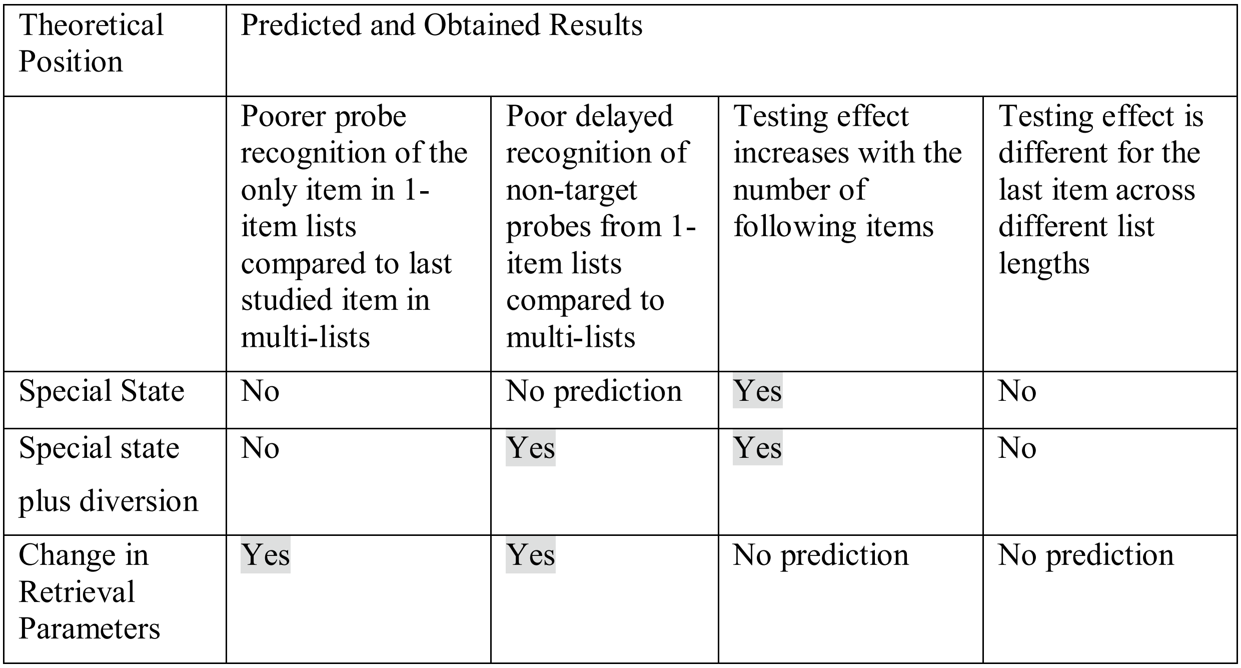
General discussion

The idea that some recently experienced items can reside in a special state, a short-term store or the focus of attention, has been fundamental to the belief that the memory for recently experienced items is primarily subsumed by a STM system that is distinct from the memory system that provides for the retention of items over longer intervals (Humphreys et al., 2020b). Recently this idea has been most commonly expressed by the idea that an item in the focus of attention does not have to be retrieved (Cowan, 1995, 2000; Jonides et al., 2008; McElree, 2001). We now have several lines of converging evidence that contradicts this key tenant of STM theory (Table 10), and some evidence for an alternative theory that assumes a change in retrieval parameters.

Does residence in a special state play a dominant role in probe recognition, delayed recognition and the testing effect?

Rose et al. (2014) argued that items which were resident in a special state would produce a small testing effect. On this basis, the size of the testing effect should increase as a function of the number of subsequently presented studies items as each additional study item should reduce likelihood of the item being in a special state. The overall logic of items residing in a special state also suggests that in probe recognition all items in a special state should be equally well recognized. However, without additional assumptions, such as an increase in fatigue across a list, the special state account does not allow to make predictions about delayed recognition. In support of the special state theoretical position the size of the testing effect did increase as the number of the following items increased. This occurred in both experiments. However, the testing effect was far from being the same across the only item in 1-item lists and the last item in multi-item lists, and residency in a special state does not predict the finding that probe recognition is poorer for the only item in 1-item lists as compared to the last item in multi-item lists. Finally, assumptions about the effects of residency in a special state cannot, on their own, predict the finding that delayed

Table 10 Predictions, along with the obtained results, regarding probe recognition of the last item, delayed recognition of non-target probes, and the size of the testing effect on the last item. The predictions assume only residence in a special state, residence in a spatial state plus a diversion of the retrieval process, and a common retrieval process with different parameter settings. Gray shading = observed result matches predicted result.



Theoretical Position	Predicted and Obtained Results			
	Poorer probe recognition of the only item in 1-item lists compared to last studied item in multi-lists	Poor delayed recognition of non-target probes from 1-item lists compared to multi-lists	Testing effect increases with the number of following items	Testing effect is different for the last item across different list lengths
Special State	No	No prediction	Yes	No
Special state plus diversion	No	Yes	Yes	No
Change in Retrieval Parameters	Yes	Yes	No prediction	No prediction

recognition for non-target probes will be poorer for the only item in a 1-item list than for the last item in a multi-item list, which additionally limits the explanatory power of this theoretical account.

The assumption that the probe retrieval process starts off the same regardless of the list length but that the process can be diverted by the presence of items in a special state can explain why delayed recognition of non-target probes is poorer following a 1-item list than following a multi-item list. The critical assumptions here are that when probed the participant knows that they are recognizing from a special state and that the study list contained only a single item. Under these conditions the failure to recognize from a special state allows the participant to conclude that the probe was new (non-target probe). This inference cannot be made with the multi-item lists because the probe could be old, but just not from a recent item in the focus of attention. Because the inference that the probe was new would terminate retrieval processing, the non-target probes from a 1-item list would be more poorly recognized on the delayed recognition test than non-target probes from multi-item lists. This example shows that while it is possible that residence in a special state is compatible with poor delayed recognition of non-target probes, the hypothetical process is complicated. It may even lead to interference due to the memory problems caused by the need to rapidly switch between tasks. In addition, this hypothesis also shares the other problems with the assumption that residency in a special state plays an important role in probe recognition and the size of the testing effect, which have not been supported by the current results.

The hypothesis that participants modify retrieval parameter settings when they realize that a 1-item list is being probed seems simpler than the hypothesis that when participants realize that not only a 1-item list is being probed, and that the retrieval process involving a special state has failed, they should respond no. For this reason we believe that a change in retrieval parameter settings is a more plausible explanation for poor recognition of non-target probes following the study of a 1-item list. According to this account, the recognition process is the same for the different list lengths but the retrieval parameters change following identification of the list length. More specifically, if participants were overconfident in their ability to recognise probes from 1-item lists this could have resulted in the lowering of their response threshold) for both the old and the new probe recognition responses when the participant realized that only a single word has been studied prior to the presentation of a probe, resulting in less complete processing of both the target and non-target probes, subsequently resulting in poorer probe recognition and poorer delayed recognition for non-targets (Strickland et al., 2022). Taken together with the delayed recognition findings, participants adequately paid attention and studied the first presented items (as indicated by improved delayed recognition of studied but-not-probed items from 1-item lists than from the last position in multi-item lists), but overestimated the ease of probe recognition when only a single item had been previously presented, resulting in changed retrieval parameters at probe recognition.

In addition, the assumption that the last item in our 1-,4-, and 8-item lists should all be in a special state when the probe is presented, and hence all of these items should produce an equal sized testing effect in delayed recognition seems to be an essential component of ideas about special states. However, the size of the testing effect is in the current study was not determined, or not largely determined, by what should be residence in a special state. That is, there was a significant difference in the size of the testing effect on the last item in an 8-item list between Experiments 1 and 2. In addition, in Experiment 2 there is a significant difference in the size of the testing effect between the 1-item lists and the multi-item lists. For these reasons we think that residency in a special state cannot explain, or not uniquely explain, the poor recognition of non-target items from 1-item lists and in the size of the testing effect.

What produces the testing effect?

The change in the retrieval parameters account predicted both the observed poorer probe recognition of the only item is 1-item lists, and poorer recognition of non-target probes from 1-item lists, compared to the last item of multi-item lists. The change in the retrieval parameters account however does not make any predictions about the testing effect so that account must be considered incomplete, limiting the explanatory power of this account. We have no definite answers about what produces the testing effect. However, there are some hints about how we might use the probe recognition task with a delayed recognition test to look for answers. The first hint comes from our current results. We are not certain why the testing effect on the 8th item of an 8-item list was significantly larger in Experiment 2 than it was in Experiment 1. However, it may have something to do with participants' ability to predict that the 8th item in an 8-item list would be the last item presented. The 8-item list was the longest list in both experiments and clearly distinguishable from the 4-item list which was the next longest list. However, it was relatively rare in Experiment 1 and far more common in Experiment 2. The greater exposure to 8-item lists would have made it easier to predict when the longest list would end. Just how this would result in a change in the size of the testing effect is not apparent but the answer may lie in the processes suggested by the third of our three hints about the size of the testing effect (to be discussed shortly).

Experiments by Jacoby and his colleagues (Jacoby et al., 2005a; Jacoby et al., 2005b) suggests that the recognition process recapitulates important aspects of the storage process. In these experiments participants studied lists of words under either deep or shallow processing instructions. They were then given a recognition test on the studied words that were intermixed with new words. A second recognition test on the new words used on the first recognition test was then administered. The new words which the participants had tried to recognize on the first test were better recognized if the old words on the first test had been processed in a deep manner than if they had been processed in a shallow manner. This assumption can explain why with the multi-item lists performance on the delayed recognition task was better following an attempt to recognize a non-studied probe than an item which was studied as part of the probe recognition task. That is, we assume that both the non-studied probe and the studied item went through the same processing stages, but that the limited time available for study, and possibly inattention resulting from the repeated performance of the same actions, slightly depressed the recognition of the studied item. Note that the studied item was presented at a rapid rate (800 ms per item with a 200 ms break between items) whereas the probe recognition test was un-paced.

The internal state of the participant, such as how surprised they are or how alert they are, may also play a role in the size of the testing effect as such states also appear to have a role in the size of repetition effects. Humphreys et al. (2010) had participants study short lists of digits before reading and pronouncing pairs of words. Participants were informed that digit recall was the primary task and that the rehearsal of the word pairs was included to make digit recall more difficult. They found that the rehearsal of low frequency word pairs produced more interference with digit recall than did the rehearsal of high frequency pairs and it also produced better recognition on a delayed recognition task of both individual low frequency words and pairs consisting of low frequency words. Humphreys et al. (2010) argued that the relative novelty of low frequency words captured attention resulting in the interference with digit recall. In turn this higher level of attention produced the better recognition of the low frequency words and word pairs. As Humphreys et al. (2010) acknowledged the weakness of this argument was that participants were motivated to rehearse the digits while they were reading and pronouncing the word pairs.

McFarlane and Humphreys (2012) addressed this problem by looking at the interfering effects of reading one pair on the memory for an immediately preceding pair. In their study participants either rehearsed the same pair twice (AB AB repeat trials) or two different pairs (AB CD switch trials). In addition, either 25% of the trials were repeat and 75% switch or the reverse. In this design participants had no incentive to rehearse the first pair member when the second pair was presented. Switch trials produced more interference with digit recall and with the delayed recognition of the first pair rehearsed when switch trials were uncommon than when they were more common. In addition memory for the second pair on a switch trial was better if switch trials were uncommon. If novelty manipulations affect learning via repetition they might also affect learning via testing. If the anticipation of the end of the list has an alerting function (Niemi & Näätänen, 1981) this could explain why the testing effect on the 8th item in an 8-item list produced a larger testing effect in Experiment 2 (the end of list was more likely to be anticipated) than in Experiment 1.

Conclusions

Most strikingly, the present results show that probe recognition for the only item in a 1-item list was worse than probe recognition for the last item in a multi-item list, even though the reverse effect was found for delayed recognition of studied but-not-probed items. Furthermore, the delayed recognition of non-target probes was worse for the 1-item list than for the multi-item lists, and items which should have been in the focus of attention were not equally well recognized. When all of the results are considered together it seems highly unlikely that residence in a special state such as a short-term store or the focus of attention is playing a prominent role in probe recognition or in producing the size of the testing effect. When the probe recognition task is followed by a delayed recognition task it provides an abundance of information about the status of individual study and testing events. This information has allowed us to conclude that residency in a special state is not playing a prominent role in the paradigm. It may also lead to an enhanced understanding of why testing effects occur.

Availability of data and materials The materials and datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Author contributions All four authors were involved in all aspects of the project, including theory development and conceptualization, data analysis, and paper writing. The order of authors reflects the relative contributions of the authors on these.

Funding information Open Access funding enabled and organized by CAUL and its Member Institutions. This research was supported by Discovery Grant DP12010311 from the Australian Research Council awarded to Loft.

Declarations

Conflicts of interest/Competing interests There are no conflicts of interest or competing interests

Ethics approval This study was approved by the Human Research Ethics Office at the University of Western Australia (RA/4/20/5313).

Consent to participate Informed consent to participate was obtained from each participant

Consent for publication Informed consent for publication was obtained from each participant

Code availability N/A

Preregistration No.

References

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 2, pp. 89–195). Academic Press.

Beaudry, O., Neath, I., Surprenant, A. M., & Tehan, G. (2014). The focus of attention is similar to other memory systems rather than uniquely different. *Frontiers in* Human Neuroscience, 8

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology, 1*, 42–45.

Cowan, N. (1995). *Attention and memory: An integrated framework*. Oxford University Press.

Cowan, N. (2000). Processing limits of selective attention and working memory: Potential implications for interpreting. *Interpreting, 5*, 117–146.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*, 87–185.

Dennis, S. (1995). The Sydney Morning Herald word database.

Duncan, M., & Murdock, B. (2000). Recognition and recall with precuing and postcuing. *Journal of Memory and Language, 42*, 301–313.

Halford, G. S., Maybery, M. T., & Bain, J. D. (1988). Set-size effects in primary memory: An age-related capacity limitation? *Memory & Cognition, 16*, 480–487.

Humphreys, M. S., Hockley, W., & Chalmers, K. A. (in press). Recognition memory: The probe, the returned signal, and the decision. *Psychonomic Bulletin & Review.*

Humphreys, M. S., Li, Y. R., Burt, J. S., & Loft, S. (2020a). How semantic processing affects memory. *Journal of Memory and Language, 113*, 104109.

Humphreys, M. S., Tehan, G., Bauman, O., & Loft, S. (2020b). Explaining short-term memory phenomena with an integrated episodic/semantic theory of long-term memory. *Cognitive Psychology, 123*, 101346.

Humphreys, M. S., Maguire, A. M., McFarlane, K. A., Burt, J. S., Bolland, S. W., Murray, K. L., & Dunn, R. (2010). Using maintenance rehearsal to explore recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 147–159.

Jacoby, L. L., Shimizu, Y., Daniels, K. A., & Rhodes, M. G. (2005a). Modes of cognitive control in recognition and source memory: Depth of retrieval. *Psychonomic Bulletin & Review, 12*, 852–857.

Jacoby, L. L., Shimizu, Y., Velanova, K., & Rhodes, M. G. (2005b). Age differences in depth of retrieval: Memory for foils. *Journal of Memory and Language, 52*, 493–504.

Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., & Moore, K. S. (2008). The mind and brain of short-term memory. *Annual Review of Psychology, 59*, 193–224.

Loft, S., & Humphreys, M. S. (2012). Enhanced recognition of words previously presented in a task with non-focal prospective memory requirements. *Psychonomic Bulletin & Review, 19*, 1142–1147.

McFarlane, K. M., & Humphreys, M. S. (2012). Maintenance-rehearsal: The key to the role attention plays in storage and forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 1001–1018.

McElree, B. (2001). Working memory and focal attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 817–835.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7*, 105–125.

Nairne, J. S. (2002). Remembering over the short-term: The case against the standard model. *Annual Review of Psychology, 53*, 53–81.

Niemi, P., & Näätänen, R. (1981). Foreperiod and simple reaction time. *Psychological Bulletin, 89*, 133–162.

Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D. A., Conway, A., Cowan, N., Donkin, C., Farrell, S., Hitch, G. J., Hurlstone, M. J., Ma, W. J., Morey, C. C., Nee, D., Schweppe, E. J., Vergauwe, E., & Ward, G. (2018). Benchmarks for models of short-term and working memory. *Psychological Bulletin, 144*, 885–958.

Rose, N. S., Buchsbaum, B. R., & Craik, F. I. M. (2014). Short-term retention of a single word relies on retrieval from long-term memory when both rehearsal and refreshing are disrupted. *Memory & Cognition, 42*, 689–700.

Sternberg, S. (1966). High speed scanning in human memory. *Science, 153*, 652–654.

Strickland, L., Heathcote, A., Humphreys, M. S., & Loft, S. (2022). Target learning in event-based prospective memory. *Journal of Experimental Psychology: Learning, Memory and Cognition.*

Wickens, D. D., Moody, M. J., & Dow, R. (1981). The nature and timing of the retrieval process and of interference effects. *Journal of Experimental Psychology: General, 110*, 1–20.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.