Cognition 231 (2023) 105330

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

The contribution of episodic long-term memory to working memory for bindings

Lea M. Bartsch *, Klaus Oberauer

University of Zurich, Switzerland

ARTICLE INFO

Keywords:
Bindings
Working memory
Long-term memory

ABSTRACT

The present experiments support two conclusions about the capacity limit of working memory (WM). First, they provide evidence for the *Binding Hypothesis,* WM capacity is limited by interference between bindings but not items. Second, they show that episodic LTM contributes substantially to binding memory when the capacity of WM is stretched to the limit by larger set sizes. We tested immediate memory for sets of word-picture pairs. With increasing set size, memory for bindings declined more precipitously than memory for items, as predicted from the binding hypothesis. Yet, at higher set sizes performance was more stable than expected from a capacity limited memory, suggesting a contribution of episodic long-term memory (LTM) to circumvent the WM capacity limit. In support of that hypothesis, we show a double dissociation of contributions of WM and episodic LTM to binding memory: Performance at set sizes larger than 3 was specifically affected by proactive interference – but were immune to influences from a distractor-filled delay. In contrast, performance at set size 2 was unaffected by proactive interference but harmed by a distractor-filled delay.

1. Introduction

In theories of human memory, a distinction is often made between working memory (WM) and long-term memory (LTM). WM is understood as a system with limited capacity that holds mental representations available for processing, whereas in LTM information is stored more permanently with presumably unlimited capacity (Cowan, 2008). WM and LTM show considerable structural and functional overlap (Eriksson, Vogel, Lansner, Bergström, & Nyberg, 2015; Lewis-Peacock & Postle, 2008; Ranganath, 2006; Ranganath & Blumenfeld, 2005; Ranganath, Cohen, & Brozinsky, 2005) and the relationship between these systems is the focus of debate between proponents of several memory models, which either portray them as two separable systems of memory (e.g. Atkinson & Shiffrin, 1968; Baddeley, 2012; Barrouillet & Camos, 2012; Cowan, 2008), or as a unitary memory system (e.g., Crowder, 1982; Melton, 1963; Nairne, 1990, 2002). Intermediate theories between these extremes conceptualize WM as a subset of LTM representations that – for a limited time – are in a heightened state of accessibility (Cowan, 1995; Oberauer, 2002).

A close relation between WM and LTM is also empirically well supported. For instance, measures of WM capacity are highly correlated with the ability to remember over the long term (Unsworth, 2010; Wilhelm, Hildebrandt, & Oberauer, 2013) and with associative learning (Tamez, Myerson, & Hale, 2008, 2012). One way in which WM and LTM are interacting is that rapidly formed representations in episodic LTM can contribute to performance in tasks designed to test WM (Cowan, 2019).

A contribution of episodic LTM to performance in WM tests is not always acknowledged by researchers, but its plausibility is undeniable: Surely episodic memory, which maintains information on events we experienced a long time ago, also maintains information about events a few seconds ago. A role of episodic memory is taken for granted in models of free recall (e.g., Davelaar, Goshen-Gottstein, Ashkenazi, Haarmann, & Usher, 2005; Sederberg, Howard, & Kahana, 2008), though not in models of serial recall (e.g., Lewandowsky & Farrell, 2008), which is the most frequently used paradigm for testing WM. The difference between free and serial recall, however, is grounded more in a historical divide of research traditions than a substantial difference (Grenfell-Essam & Ward, 2012; Ward, Tan, & Grenfell-Essam, 2010). For example, one of the main work horses of WM research, the complex-span paradigm (Daneman & Carpenter, 1980), is very similar to the continual-distractor paradigm in research on episodic LTM (Bjork & Whitten, 1974): Presentation of list items is interleaved with periods of distractor processing. In the episodic LTM literature, distractor tasks are

* Corresponding author at: Department of Psychology, Cognitive Psychology Unit, University of Zurich, Binzmuehlestrasse 14/22, 8050 Zurich, Switzerland.
E-mail address: l.bartsch@psychologie.uzh.ch (L.M. Bartsch).

https://doi.org/10.1016/j.cognition.2022.105330
Received 19 May 2022; Received in revised form 9 November 2022; Accepted 15 November 2022
0010-0277/© 2022 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

used to displace information from WM to obtain a purer measure of episodic LTM. Some researchers have argued that the same could happen in complex-span tasks (McCabe, 2008; Unsworth & Engle, 2007), so that performance reflects to a substantial degree recall from episodic LTM.

More generally, Cowan (2019) has argued that episodic traces are created rapidly in the *activated part of LTM* and contribute to performance in WM tests. Beukers, Buschman, Cohen, and Norman (2021) have argued that information successfully maintained in WM tasks while remaining "neurally silent" (i.e., not decodable from neural activity during the retention interval) is not maintained in WM but in episodic LTM. In support of that possibility, research has shown that quick formation of episodic LTM traces is possible within a few seconds (Huebner & Gegenfurtner, 2011; Verhaeghen, Vandenbroucke, & Dierckx, 1998).

The goal of the present study is to understand how LTM contributes to WM based on the assumption that there must be some form of separation between contents of WM and representations in LTM to reflect their different functions. Thereby, the present work will not speak to the debate on whether WM and LTM are structurally separate systems. Rather, we focus on identifying functional differences between them by which we can identify to what extent performance on a memory task relies on WM, and to what extent it relies on episodic LTM.

1.1. WM capacity is a limit on bindings

The question how WM and LTM are related did not originally motivate the present experiments. Rather, the starting point of our investigation was: Why is the capacity of WM limited? One hypothesis about the nature of this capacity limit is that it places a specific limit on the temporary *bindings* remembered (Oberauer, 2019). This *binding hypothesis* states that the limit of WM pertains to the short-term maintenance of bindings but not items.

Memory for items is defined as the ability to remember which individual items (e.g., words, visual objects) have occurred recently. It can be achieved through sustaining a high level of activation of the presented items' representations, which in some theories of WM is the only form of maintenance (e.g., Davelaar et al., 2005; Ma, Husain, & Bays, 2014). Item memory is often sufficient to remember which items were included in the current memory set – the ones with the highest level of activation – but not for remembering the structure of the memory set (e.g., the order of the items). Bindings are understood as temporary links between representations of items, coding the relations between them (e.g., which object has been presented together with which word), or relations between items and their context (e.g., which object has been presented in which location, or which word has been presented in which serial position of a list). A failure of binding memory while item memory is intact manifests itself in characteristic errors in WM tests. For instance, in serial recall, people often recall the list items but reproduce them in the wrong order (Henson, 1996). In a test of WM for spatial arrays of objects, participants are very good at remembering which objects were in the array, and which locations were occupied, but often fail in assigning each object to its correct location (Pertzov, Dong, Peich, & Husain, 2012).

The *binding hypothesis* states that the capacity of WM is a limiting factor for the short-term maintenance of bindings but not items. Memory for items is maintained in the activated part of LTM, which is part of the WM system but not subject to a capacity limit (Cowan, 1995; Oberauer, 2009a) The binding hypothesis can be tested by capitalizing on the primary experimental finding demonstrating the limited capacity of WM: the set-size effect. As the number of items to be held in WM increases, people's performance decreases steeply (see Fig. 1 for a few examples). The binding hypothesis implies that with increasing set size, memory for bindings should decline, whereas memory for items should be (largely) unimpaired. Oberauer (2019) provided evidence for this prediction in two experiments testing immediate memory for word lists of varying set sizes. Memory for probes in particular list positions was probed through multiple-alternative forced-choice tests in which some alternatives represented item errors (i.e., words not in the list), whereas others represented binding errors (i.e., list words from other than the tested position). Participants' error rates increased with the length of the list, and this was nearly exclusively due to binding errors (but see Cowan, 2022). Additional support for the binding hypothesis comes from individual-differences studies, which have shown that the ability to maintain temporary bindings is highly correlated with fluid intelligence, and is likely to drive the high correlations between performance in WM tasks and intelligence tests (Chuderski, 2019; Wilhelm et al., 2013).

To date, the claim that WM capacity is limited by interference between bindings has been experimentally investigated only for bindings of items to their respective serial positions in a list. However, the effect of set size on the ability to remember the relation between list items and their position might not be a pure reflection of the capacity limit on maintaining bindings. This is because set size might be confounded with the distinctiveness of serial positions. Several formal models of serial recall imply such a confound. For instance, in the Start-End model (Henson, 1998) serial positions are represented by a weighted combination of a marker for the start of the list and a marker for the end. The distinctiveness of these position representations is lower in the middle of the list than closer to the start and the end – giving rise to primacy and recency effects. As the list length increases, more items are allocated to positions further away from the start and the end, so that average positional distinctiveness decreases. In another model (Botvinick & Watanabe, 2007), serial positions are assumed to be similar to representations of numerical quantities. Higher quantities are less distinctive than lower ones – for instance, it is easier to distinguish between 1 and 2 objects than between 6 and 7. As the length of a list increases, positions corresponding to higher numbers are recruited, again implying that the average positional distinctiveness in the list decreases.

Therefore, in the present study, we aimed to test the binding hypothesis for another form of bindings in WM: the bindings between words and pictures. This choice of materials has the advantage that there is no known confound between the memory set size and the distinctiveness of the representations that are bound to each other: The pairwise distinctiveness between words or pictures sampled at random from a pool should not be affected by how many words or pictures are included in the memory set. Previous work using a similar probed-recall paradigm for investigating memory for short lists of word pairs (Murdock Jr., 1963) has shown that the probability of recalling the correct pair decreases with increases in set size. Therefore, this task is suited to further investigate the contribution of item and of binding memory to the set-size effect.

According to the binding hypothesis, interference between bindings creates a capacity limit selectively on the ability to maintain them. Therefore, we should still expect that increasing set size strongly impairs people's ability to remember which word was combined with which picture. At the same time, increasing set size should have no, or a much more benign, effect on people's ability to remember individual items (i.e., to remember which words, or pictures, they have seen in the current trial).

The prediction for item memory is not that there is no set-size effect at all, because if item memory is not constrained by WM capacity, it is still limited by factors limiting access to episodic LTM.[1] However, we expect that set-size effect not to show the signature of a capacity limit, as often observed in WM tests: Stable performance at a high leve for set sizes within capacity, followed by a precipitous decline once the set size

[1] Most models of episodic LTM (e.g., SAM, Gillund & Shiffrin, 1984; REM, Shiffrin & Steyvers, 1997; SIMPLE, Brown, Neath, & Chater, 2002) assume that retrieval is partly driven by a temporal context cue. Larger set sizes imply that a larger number of events are associated to similar temporal contexts. This leads to more cue overload for the temporal context that is re-instated at retrieval, and thereby, an effect of set size or list length.

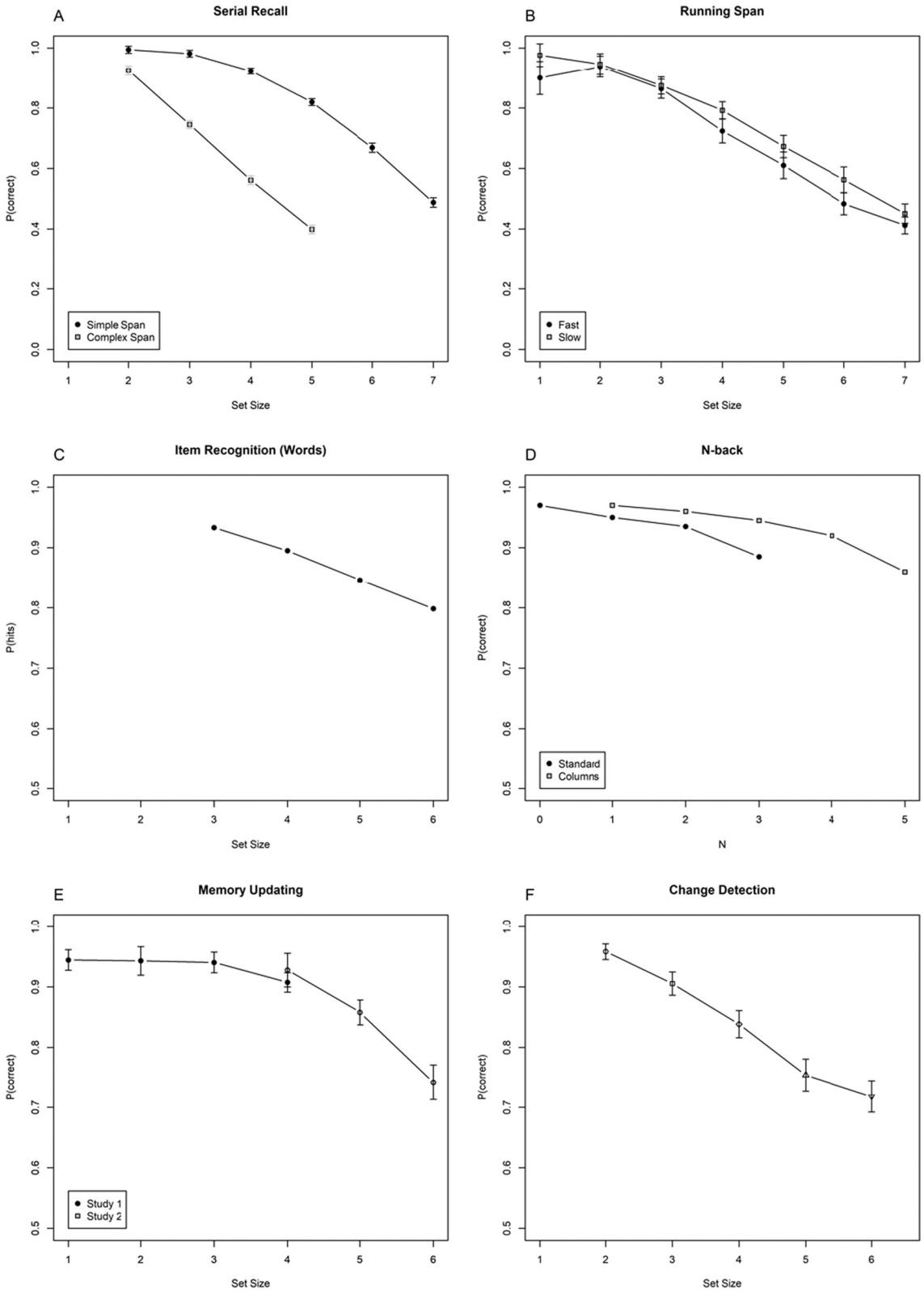

Fig. 1. Representative findings demonstrating the set-size effect on accuracy (taken from Oberauer et al., 2018). (A) Serial recall in simple and complex span tests with verbal materials (Unsworth & Engle, 2006). (B) Running memory span (Bunting, Cowan, & Scott Saults, 2006). (C) Item recognition (McElree & Dosher, 1989). (D) Standard N-back (Jonides et al., 1997), and a version of N-back in which subsequent stimuli are presented across N columns, such that each stimulus appears in the same column as the one N steps back (Verhaeghen & Basak, 2005). (E) WM updating with digits and arithmetic operations (Oberauer & Kliegl, 2001). (F) Change detection with arrays of colored squares (Adam, Mance, Fukuda, & Vogel, 2015).

exceeds capacity.

1.2. The present study

The first goal of the present study was to provide a new test of the binding hypothesis of WM capacity. We achieved this with Experiment 1, but that experiment also revealed a strong hint that episodic LTM contributed substantially to our test of binding memory, especially at larger set sizes. Therefore, our second goal was to investigate this contribution, and to isolate it from the contribution of WM to binding memory. We did so by introducing manipulations that are assumed to specifically disrupt either WM, such as introducing a distractor task prior to test (e.g. Peterson & Peterson, 1959), or episodic LTM, by introducing proactive interference (Cowan, 2005). We aimed to investigate at which set size each of these manipulations influences memory performance in an immediate 4-alternatives forced choice recognition test.

2. Experiment 1

In Experiment 1, we aimed to test the binding hypothesis by testing subjects' memory for varying numbers of word-picture pairs. At test memory was cued by either pictures or words, and subjects were to choose the other element of the tested pair from four response options. These options included the target and three different types of lures: (1) a *within-trial lure:* a stimulus that was part of a pair in the current pair but not the tested pair; (2) an old lure: a stimulus that was part of a pair in a previous trial; and (3) a new item: a stimulus not used in any trial. This test was designed to distinguish item and binding memory: Rejection of the within-trial lure requires memory for the correct binding between the respectively cued picture (or word) and the target word (or picture). If subjects have intact item memory but no binding memory, they would be able to exclude old and new lures but have to choose between the target and within-trial lure with equal probability. If they have neither binding nor item memory in WM, but some familiarity signal for previously presented items in general that enables them to exclude the new lure, then they would choose between the target, the within-trial lure and the old lure. If they have no memory at all, they have to guess between all four response options (see Bartsch, Loaiza, & Oberauer, 2019; Oberauer, 2019; Wilhelm et al., 2013 for similar approaches).

If WM capacity is truly limited by interference between bindings, whereas item memory is unaffected by the WM capacity limit, then people should be able to avoid item errors (i.e., selecting old or new lures) fairly well regardless of set size, but they should show a sharp increase of binding errors (i.e., selecting the within-trial lure) as set size reaches and exceeds WM capacity (i.e., around 3–4 pairs). Hence, performance should decrease with increasing set sizes, and that increase should be due primarily to a higher percentage of falsely chosen within-trial lures at test.

2.1. Method

2.1.1. Participants

We collected data of 20 participants for Experiment 1 (Mean age = 23.67; 15 female). We chose the initial sample size of $N = 20$ for this and the following experiments because it is sufficient to detect medium to large effects in within-subjects designs, and because memory set-size effects are known to be large. The use of Bayesian statistics means that the sample size could be increased in case of ambiguous evidence until the ambiguity was resolved (Rouder, 2014; Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017); this was necessary for Experiments 2, 3, 4 and 5. Only participants whose mother tongue is German, aged between 18 and 35 years, and reporting normal or corrected-to-normal vision took part in this and the following experiments. All were students enrolled at the University of Zurich or the ETH Zurich. Participants signed an informed consent prior to the study and were debriefed at the end. The experimental protocol of this and all following experiments is in accordance with the regulations of the Ethics Committee of the Faculty of Arts and Social Sciences at the University of Zurich. This and all following experiments were run in-person in the laboratory.

2.1.2. Materials and procedure

In order to measure working memory for bindings and varying conditions of WM load, we sequentially presented arbitrary word-picture pairs at set sizes 2, 3, 4, 6, and 8 (see Fig. 2). Words were drawn from a pool of 1198 neutral words from the Berlin Affective Word List (BAWL-R, Vo et al., 2009), with a minimum length of 2 and a maximum length of 10 letters. The words had a mean frequency per million of 59.33. Pictures were drawn from a pool of 2400 photographs of real-world objects (Brady, Konkle, Alvarez, & Oliva, 2008). Words and pictures of each trial were drawn from these pools without replacement, so that each stimulus is used in only one trial.

Each pair was presented for 900 ms with a 100 ms inter-stimulus-interval. After a 1000 ms retention-interval, immediate memory was tested for a random half of the presented pairs in random order with a four-alternative forced choice procedure. In the odd-numbered set size 3 trials only one of the three pairs was tested. In half of the trials, subjects were cued with the word of a pair and were to choose from four pictures displayed around it by clicking on the option they thought to have been presented together with the word at encoding (word-cue condition). In the other half of the trials, the subjects were cued with a picture and were to choose from 4 words (picture-cue condition). The four options always comprised the correct item, a same-trial lure, a lure from a previous trial (old intrusion lure; selected from any randomly chosen prior trial) and a new item (new lure). Again, this test was designed to distinguish item and binding memory: Rejection of the within-trial lure requires memory for the correct binding between the respectively cued picture (or word) and the target word (or picture). If subjects do not have intact binding memory, they would choose between the target and within-trial lure with equal probability. Similarly, rejection of the old-trial lure requires memory for the correct items of the current trial. Only if subjects have no memory at all, then they would guess between all four options with equal probability.

The within-trial lure was always drawn from a pair presented immediately before or after the target pair, in order to control for increasing context changes across the pairs within a trial at larger set sizes. The reason for only testing half of the presented pairs was that we wanted to avoid repeating any response option across tests to minimize dependencies between responses in the same trial. Therefore, the same-trial lures were drawn from the not-tested pairs.

The one-hour experiment comprised 80 trials, with 8 trials per combination of cue-condition (word vs. picture cue) and set size (2,3,4,6, and 8).

2.1.3. Data analysis

We analyzed the data of Experiment 1 using a Bayesian hierarchical logistic regression predicting correct or incorrect responses in the 4-AFC test by memory set size. Correct responses were defined as choosing the target item from the alternatives (i.e., within-trial lure, old-intrusion lure, new lure). We included a by-participant random intercept and a

Fig. 2. Illustration of the paradigm of Experiment 1. Subjects were sequentially shown a list of word-picture pairs at varying set sizes. Immediate memory was tested for half of the pairs in random order, by either cueing with the picture or word. Subjects had to choose between the four response options, comprising the correct item, a same-trial lure, an old-trial lure and a new item.

random slope for the fixed effect (i.e., set size). The fixed effect was set size and was included as a continuous predictor in the model.

In case we found any set size effect, we tested whether it was driven by reduced memory for bindings, which would become manifest in an increase in probability to choose the same-trial lure above all the remaining options. We implemented this in a Bayesian hierarchical logistic regression predicting the probability of errors in the 4-AFC test separately for each type of error (same-trial intrusion, old intrusion lure, and new lure) by memory set size.

It is to note that the four dependent variables are not independent, because the number of responses for each of the 4 categories (correct, same-trial intrusion, old intrusion lure and new lure) sum up to one. Therefore, one of these analyses is redundant. We nevertheless report all four to present the evidence for the set-size effect on each category of responses in the most transparent way.

The model was implemented in the R package *brms* (Bürkner, 2017, 2018). The regression coefficients were given Cauchy priors with a scale of 0.353. These scales were chosen because these priors were recently proposed as default priors for model comparisons with logistic models (Oberauer, 2019).We used completely non-informative priors for the correlation matrices, so-called LKJ priors with shape parameter 1.

We calculated Bayes Factors to estimate the strength of evidence for the effect of each predictor, by comparing the model including that predictor to a model excluding it. Similarly, we tested for evidence including the interaction term as a random effect – where applicable – and maintained it for subsequent model comparisons if there was evidence for it. Models excluding a main effect only excluded its fixed effect but kept its random effect. We used the Bayes factor for model comparison, calculated through the *bridge sampler* (Gronau, Singmann, & Wagenmakers, 2017) included in the *brms* package.

A BF_{10} larger than 1 gives evidence for an effect; a BF_{10} lower than 1 yields evidence against an effect, and hence evidence for the null hypothesis. The strength of evidence for the null hypothesis is $BF_{01} = 1/BF_{10}$. A BF_{10} of 10 indicates that the data are 10 times more likely under the alternative hypothesis than under the null hypothesis. Usually, BFs > 3 are regarded as providing substantial evidence for one hypothesis over the other.

We used an MCMC algorithm (implemented in Stan; Carpenter et al., 2017) that estimated the posteriors by sampling parameter values proportional to the product of prior and likelihood. These samples are generated through 4 independent Markov chains, with 1000 warmup samples each, followed by 50,000 samples drawn from the posterior distribution which were retained for analysis. Following Gelman et al. (2013), we confirmed that the 4 chains converged to the same posterior distribution by verifying that the $\widehat{R}$ statistic – reflecting the ratio of between-chain variance to within-chain variance – was <1.02 for all parameters, and we visually inspected the chains for convergence.

Further, we calculated the common estimate of the capacity of WM across our variation of set size and additional manipulations in each experiment. Based on the assumptions of discrete memory states, this estimate K is derived from the number of items remembered (Adam, Vogel, & Awh, 2017; Cowan, 2001; Zhang & Luck, 2008). For n-alternative forced choice tests of a single item, this estimate is obtained from

$$K = N\frac{P(correct) - g}{1 - g}$$

with K for the number of remembered items; N for the memory set size, P (correct) for proportion of correct responses, and g for the chance of guessing the correct response. The chance of guessing depends on assumptions about the capacity limit and on whether the difficulty of the lures affect the probability of choosing them. If the capacity limit is a limit on all information remembered about an item, then we distinguish a state of remembering the tested item (leading to a correct response) from a state of no information about the item, in which case the person guesses with equal probability from the entire response set: Here the chance of guessing is determined by the number of response alternatives, resulting in $g = 1/4$. If we assume that WM capacity places a limit on binding memory but not item memory, then we can distinguish a state of remembering the item-position binding (leading to a correct response) from a state of not remembering the binding, but still having item memory available to restrict the response set to the candidates from the *current* list. On this assumption, guessing chooses each candidate from the current list with equal chance, leading to guessing being set to ½. Here we report *K* estimates based on $g = ¼$. because it reflects the most common assumption about capacity limits; K estimates based on $g = ½$ can be found in the Supplementary Material.

Finally, we also computed the input-output distance of the pairs as a predictor of the probability of choosing the correct response. The input-output distance represents the number of intervening events – encoding or testing of other pairs – between encoding of a given pair and the test of that pair. For example, if a pair in the set-size 8 condition is presented in the serial position 6, and tested first, the input-output distance would be 2. If the same pair would be tested 4th, the input-output distance would be 5. This analysis was exploratory, as we had no predictions for it, but it turned out to be informative about the division of labor between WM and episodic LTM.

The data and the analysis scripts can be accessed in the Open Science Framework (https://osf.io/ymgkq).

2.2. Results

Fig. 3 shows the mean probability of choosing each of the four response options across the set sizes. The interested reader can also find the mean proportion of correct responses broken down by cue-type, serial position, as well as output position in the Supplementary (Fig. S2A, Fig. S 3A, Fig. S 4A). Table 2 lists the formulae, intra-class correlations as well as number of observations of the full models of the hierarchical logistic regression models.

2.2.1. Accuracy

Our first question was whether memory load (i.e., set size) affected the probability of choosing the correct option. The Bayesian hierarchical logistic regression revealed a main effect of set size ($BF_{10} = 5.2 \times 10^{10}$), compared to a model excluding its fixed effect. Yet, as shown in Fig. 2, the mean probability of choosing the correct item only decreased from set sizes 2 to 4, with performance leveling-off between set size 4 to 8. In order to estimate the evidence for that deceleration of the set size effect, we contrasted models with different set-size codings. We replaced the linearly dropping set size predictor with alternative predictors, in which set-size-coding reflected the different possible plateaus (Table 1). The model that predicted the correct response probability with set size leveling-off at values 4 and higher described the data best; compared to the model assuming a linear drop across all set sizes, it was preferred with a BF_{10} of 8.12×10^4 (see Table 1) for the comparisons of the linear drop model to all possible plateaus).

2.2.2. Error types

Our second question was whether this set size effect was driven by reduced memory for bindings, which would become manifest in an increase in probability to choose the same-trial lure above all the remaining options. We therefore used a Bayesian hierarchical logistic regression predicting the probability of errors in the 4-AFC test separately for each type of error (same-trial intrusion, old intrusion lure, and new lure) by memory set size. We found that the set size effect was fully driven by choosing the same-trial lure, ($BF_{10} = 9.18 \times 10^{10}$), as there was evidence against a set size effect for both other error types, choosing old trial intrusions ($BF_{10} = 12.37$), as well as new items ($BF_{01} = 8.97$).

2.2.3. Input-output distance

Finally, we inspected the effect of input-output distance on the probability of choosing the correct response. As shown in Fig. 4, recall

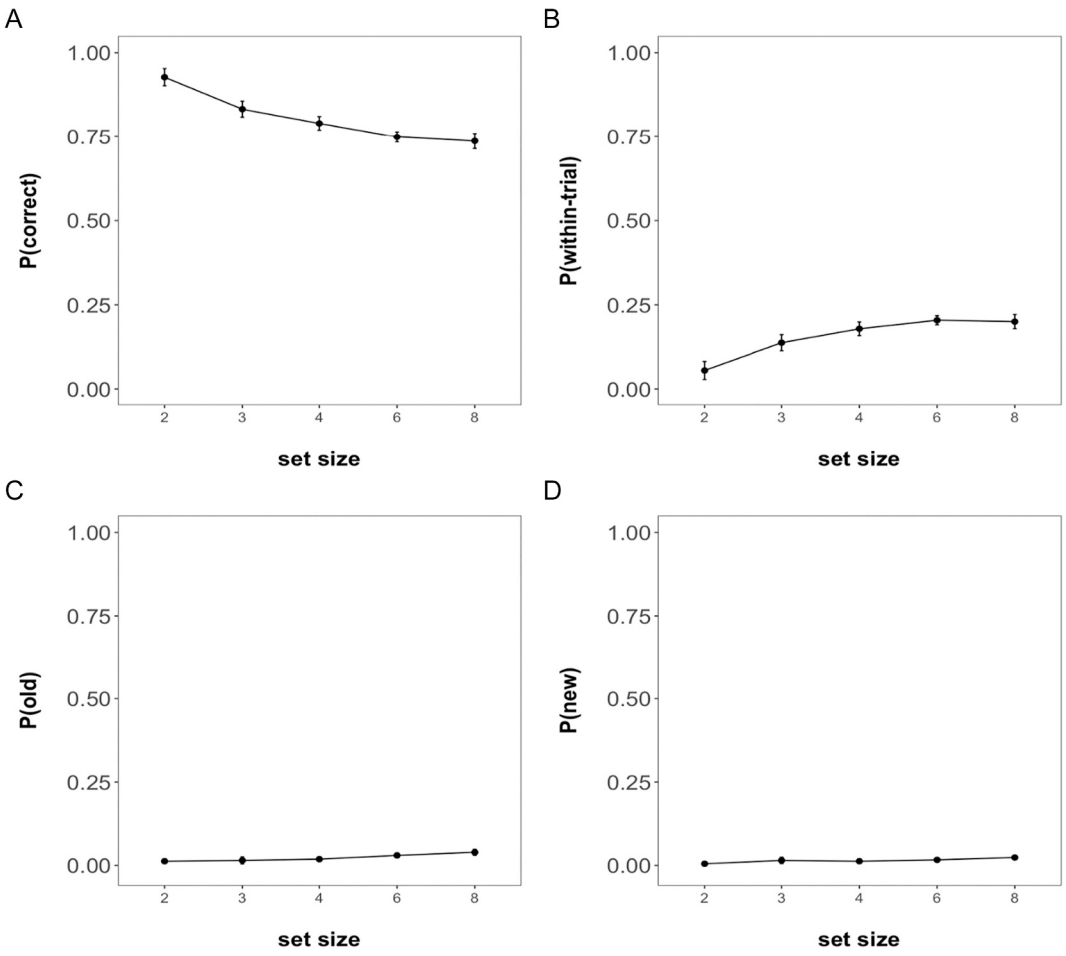

Fig. 3. Immediate Memory performance in Experiment 1. Panel A shows the probability of choosing the correct item across set sizes. Panels B, C, and D show the probability of choosing the within-trial lure, the the old lure, and the new lure, respectively. Error bars represent the within subject 95% confidence interval.

Table 1
Results of the Bayesian hierarchical logistic regression predicting the number of correct responses in Experiment 1.

Model	Set size Coding					Comparison	BF_{10}
I	0	1	2	4	6		
II	0	1	2	4	4	vs. linear [I]	2011
						vs. Best [III]	0.025
III	**0**	**1**	**2**	**2**	**2**	**vs. linear [I]**	8.12×10^{4}
IV	0	1	1	1	1	vs. linear [I]	252
						vs. Best [III]	1.22×10^{-5}

Note: All models include participant as a random effect. For tests of each set size predictor (1–4), the first row presents the Bayes factor (BF) of each model (shown in different columns) over the linear model [1]. The second row shows the BF of the model using the respective constrained predictor over the best model. The model printed in bold is the best model compared to the Null model.

was best for short input-output distances, and the set-size effect arose primarily from the fact that larger set sizes imply longer input-output distances on average.

2.3. Discussion

The goal of Experiment 1 was to test the binding hypothesis for the bindings between items. The results support the binding hypothesis, as the set size effect was driven exclusively by reduced memory for bindings, which became manifest in an increase in probability to choose the same-trial lure above all the remaining options.

Unexpectedly, the detrimental effect of memory load on WM

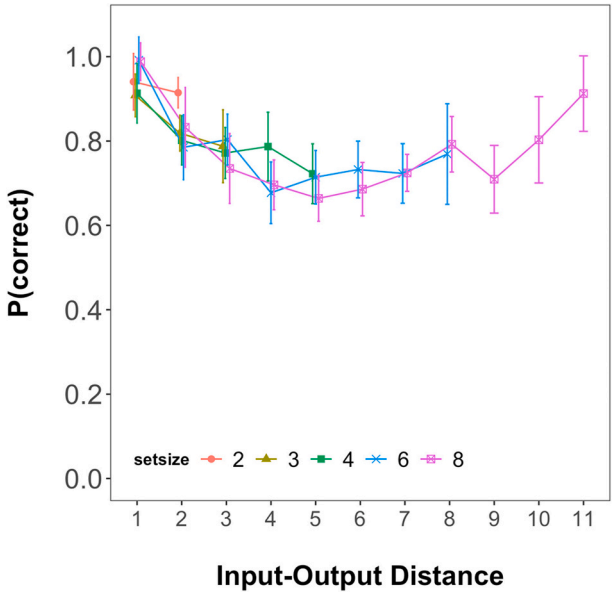

Fig. 4. Probability of choosing the correct response across Input-Output Distance and set sizes in Experiment 1. The error bars show between-subject confidence intervals.

decelerated after 4 pairs, with participants performing comparably in the immediate memory task across set sizes 4, 6, and 8. This was surprising, as previous research has consistently shown that across various WM paradigms, accuracy for item-context bindings drops continuously – often in an accelerated fashion – over increasing set sizes (for an overview: Oberauer et al., 2018: Benchmark 1.1 and Fig. 1). A decelerating effect of set size is not what we should expect from a capacity-limited memory, because it implies that as more information is presented, more information is maintained in memory. To illustrate, Fig. 5A shows the K values reflecting the number of pairs remembered across set sizes, indicating that the capacity estimates increased continuously with set size (see Supplementary Fig. S 1 for K estimates based on 50% guessing probability). One possible explanation for the decelerating set-size effect is that participants rely at least in part on their episodic LTM for maintaining word-picture bindings, and the influence of LTM increases with larger set sizes in such a way that performance levels-off at set size 4 to 8 at a level of accuracy that can be sustained by using LTM alone. We therefore hypothesize that at larger set sizes, episodic LTM contributes strongly to performance in our WM task – in particular, it contributes not only to item memory but also to binding memory.

To test this hypothesis, in Experiments 2 and 3, we manipulated the usefulness of relying on episodic LTM representations by inducing proactive interference in one of the conditions. If episodic LTM specifically contributes to performance at larger set sizes, the effect of proactive interference (PI) should be present only at larger set sizes. In contrast, at smaller set sizes, where performance relies on WM representations, PI should not have any effect, because WM protects its contents against PI (Cowan, 2005; Cowan, Johnson, & Scott Saults, 2005; Halford, Maybery, & Bain, 1988; Oberauer, Awh, & Sutterer, 2017; Wickens, Born, & Allen, 1963; Wickens, Moody, & Dow, 1981). In Experiments 4 and 5 we tested the converse hypothesis that a distractor task during the retention interval, which should selectively impair WM, affects performance only at short set sizes.

A further finding was the relatively high performance at short input-output distances. This observation is compatible with our assumption that WM capacity is limited by interference between bindings, if we make one additional assumption: With larger set sizes, the last-presented pairs are the ones maintained preferentially in WM. There could be two reasons for this. One is that interference in WM is mostly retroactive, so that earlier presented pairs are most corrupted by interference. Alternatively, earlier presented pairs are intentionally outsourced to episodic LTM because they have the longest time for being established in episodic memory. On these assumptions, the last-presented pairs are most accessible in WM, especially when they are tested early in the output sequence. At later tests, they are increasingly corrupted through output interference from the preceding tests, caused by retrieving other pairs from episodic memory into WM.

When testing the hypothesis that episodic LTM contributes specifically to performance at larger set sizes in the following Experiments, we will therefore also investigate whether the performance for pairs with short input-output distances are selectively protected against PI, and are selectively vulnerable to secondary-task interference, also at larger set sizes.

3. Experiments 2 and 3

In Experiments 2 and 3 we aimed to investigate the contribution of episodic LTM to our WM binding task at varying set sizes. We tested the hypothesis that LTM more strongly contributes to performance at larger set sizes. In Experiment 1 we used trial-unique stimuli, that is, each word and each picture was presented only once across all trials. In Experiments 2 and 3 we induced proactive interference by repeating the same words and pictures from a closed pool of items, yet in new combinations, and compared this condition to a condition with trial-unique stimuli. In addition, we investigated set sizes over a larger range (1, 2, 3, 6, 8, and 30 pairs in Experiment 2; 2, 3, 8 and 16 pairs in Experiment 3). If the decelerating ing set-size effect observed in Experiment 1 reflects an asymptote of performance determined by episodic LTM, then it should extend to even higher set sizes. In that case, memory for 16 or for 30 pairs should be on the same level as memory for 6 or 8 pairs.

3.1. Method

3.1.1. Participants

We collected data of 30 participants for Experiment 2 (mean age = 23.21; 22 female), and data of 40 participants for Experiment 3 (mean age = 23.0; 29 females).

3.1.2. Materials and procedure

We test memory for bindings in Experiment 2 at set size 2, 3, 6, 8, and set size 30, and in Experiment 3 at set size 2, 3, 8 and set size 16 in conditions with and without proactive interference (PI).

Experiment 2 involved two-sessions à 1-h comprising 120 trials, with 20 trials per condition for each small set size (2,3,6,8) and five trials for set size 30. The conditions (PI vs. no PI) were varied between sessions. The order of sessions was counterbalanced between participants. In the PI session, the stimuli were drawn from sets of twelve words and twelve pictures for the small set sizes (2,3,6,8), and from sets of 45 words and 45 pictures for set size 30. These 45 stimuli included the twelve from the closed set of the small set sizes. In the no-PI session, stimuli were again drawn without replacement from the large sets of Experiment 1. Apart from these changes, the paradigm followed the one from Experiment 1.

Experiment 3 was also a two-session à 1-h experiment comprising 208 trials, with 13 trials per condition for each set size. The conditions (PI vs. no PI) were varied between sessions, and the cue (word vs. picture cue) was held constant within participants (half the participants were always cued with words, and the other half with pictures).[2] In the PI session, the stimuli were drawn from sets of 24 words and 24 pictures.[3] Apart from these changes, the paradigm followed the one from Experiment 2.

3.1.3. Data analysis

We analyzed Experiment 2 and 3 using a Bayesian hierarchical logistic regression predicting correct or incorrect responses in the 4-AFC test by set size and proactive interference. Again, correct responses were defined as choosing the target item from the alternatives (i.e., within-list intrusion, old-intrusion, new). The fixed effects were *proactive interference* and *set size* and the latter was included in the model as a continuous predictor, increasing over small set sizes and decelerating at set sizes 4 and larger (i.e., 6,8, and 30 for Exp2, and 8 and 16 for Exp. 3) as established in Experiment 1 (i.e., Coding III in Table 1). The model included a random intercept, and random slopes of set size and proactive interference.

3.2. Results

Fig. 6 and Fig. 7 show the mean probability of choosing each of the four response options as a function of set size and proactive interference condition, for Experiment 2 and 3, respectively. The interested reader can also find the mean proportion of correct responses broken down by cue-type, serial position, as well as output position in the Supplementary (Fig. S2B and C, Fig. S 3B and C, Fig. S 4B and C). Table 2 lists the formulae, intra-class correlations as well as number of observations of

[2] In order to induce proactive interference more efficiently, we decided to hold the cue constant in Exp 3. In case the cue can change from trial to trial, it takes longer to build up proactive interference for each of them.

[3] The choice of set size 16 rather than 30 for the very large sets made it possible to use the same small pool for all set sizes, thereby circumventing the need to introduce further stimuli for the largest set size in Experiment 2, which might have weakened the effect of PI for that set size.

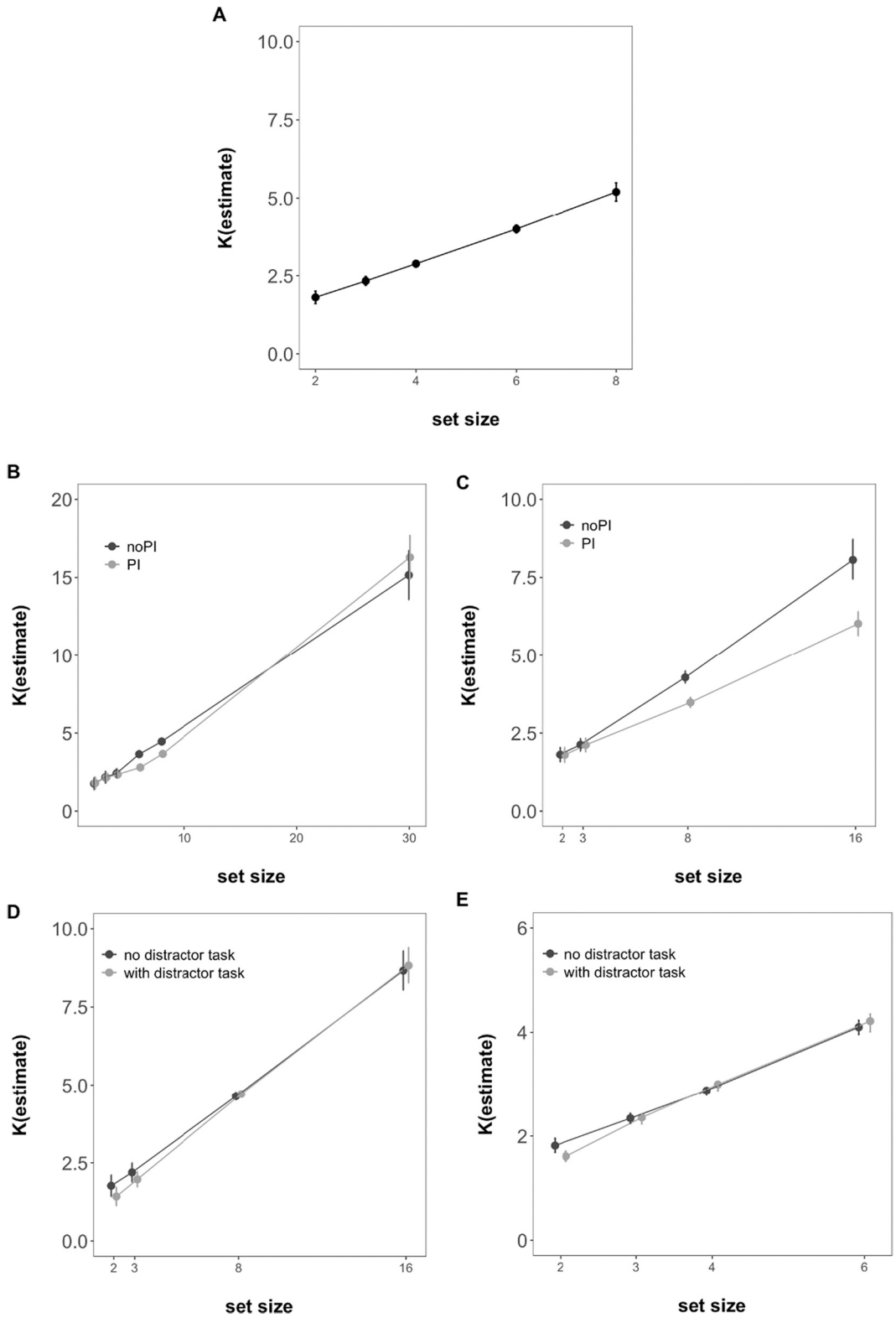

Fig. 5. Capacity estimates K across set sizes and conditions of (A) Experiment 1, (B) Experiment 2, (C) Experiment 3, (D) Experiment 4, and (E) Experiment 5.

full models of the hierarchical logistic regression models.

3.2.1. Accuracy

Our first question was whether proactive interference affected the set size effect on the probability of choosing the correct option, limiting the analysis to the set-size range of Experiment 1 (i.e., 2 to 8). The Bayesian hierarchical logistic regression revealed substantial evidence for an interaction effect both, in Experiment 2 and 3 ($BF_{10} = 7.81$ and $BF_{10} = 10.19$, respectively) as well as strong evidence for the main effects of set size ($BF_{10} = 1.19 \times 10^{12}$ and $BF_{10} = 2.25 \times 10^{20}$), and proactive interference ($BF_{10} = 16.10$ and $BF_{10} = 406.78$) across both experiments. There was a higher probability of choosing the correct item at smaller compared to larger set sizes, and without PI than with PI. Follow-up analyses on data of Experiment 2 revealed that proactive interference

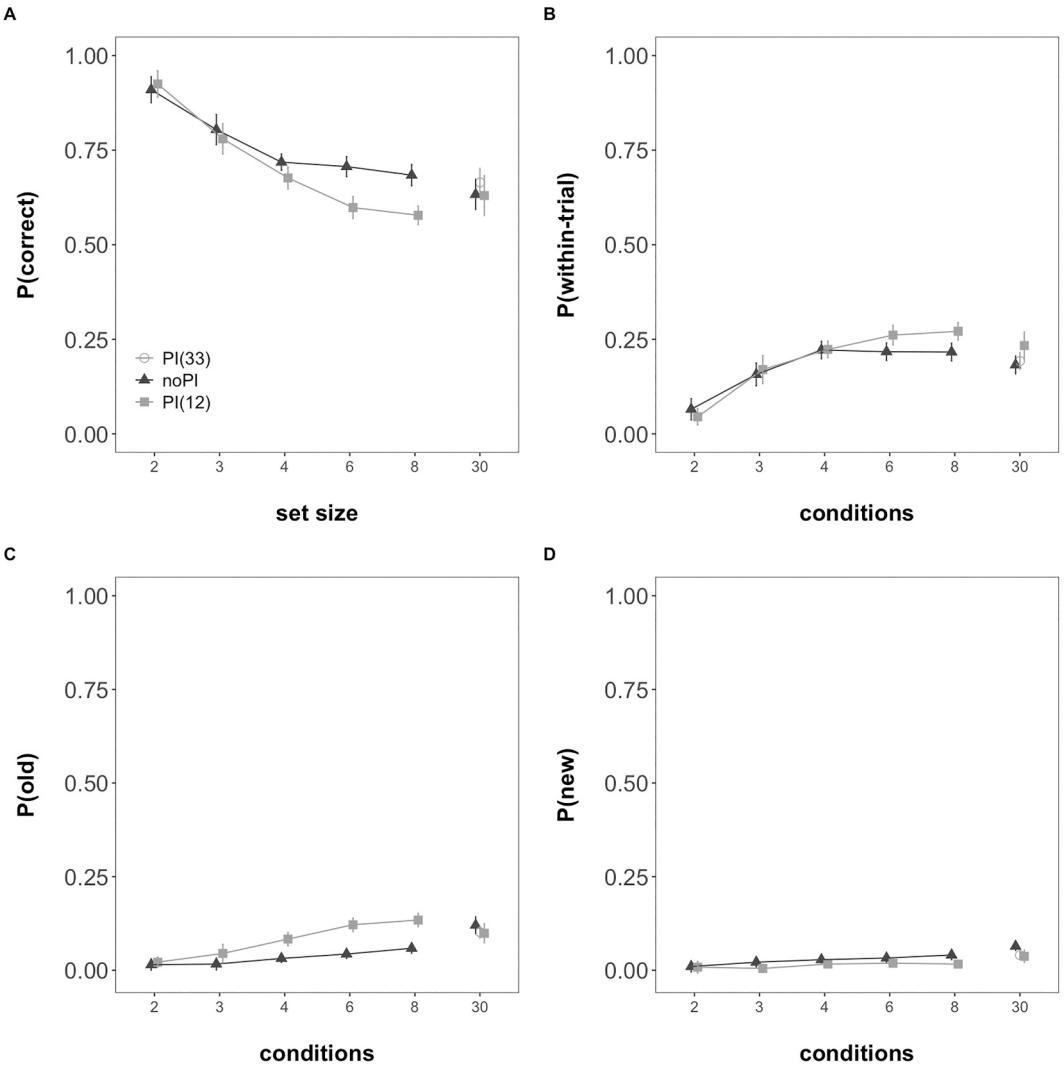

Fig. 6. Immediate memory performance in Experiment 2. Panel A shows the probability of choosing the correct item across the set sizes and under conditions with proactive interference realized by a small pool of 12 items [PI (12)], and without (noPI). For set-size 30 in the proactive-interference condition, performance is also shown for stimuli from an additional pool of 33 items [PI(33)]. Panels B, C, and D show the probability of choosing the within-trial lure, the old lure, and the new lure, respectively. Error bars represent the within subject 95% confidence interval.

only (negatively) affected the probability of choosing the correct option for set sizes 6 and 8 ($BF_{10} = 5890$, and $BF_{10} = 17.37$, respectively), but had no detrimental effect at set sizes 2, 3, and 4 ($BF_{01} = 5.46$, $BF_{01} = 9.25$, and $BF_{01} = 13.27$, respectively). Equivalently, follow-up analyses of Experiment 3 revealed that proactive interference only (negatively) affected the probability of choosing the correct option for set size 8 ($BF_{10} = 20.14$) but had no detrimental effect at set sizes 2 ($BF_{01} = 13.44$) and 3 ($BF_{01} = 20.61$).

3.2.2. Error types

Our second question was whether the set size effect was driven by erroneously choosing the same-trial lure, reflecting an effect specifically on binding memory, and how this effect is affected by proactive interference. We therefore used a Bayesian hierarchical logistic regression predicting the probability of errors in the 4-AFC test separately for each type of error (same-trial lure, old intrusion lure, and new) by set size and proactive interference condition.

The probability of choosing the same-trial lure increased with set size in both Experiments 2 and 3 ($BF_{10} = 6.96 \times 10^{7}$and $BF_{10} = 5.06 \times 10^{14}$), and there was either no evidence for an interaction with PI (Exp 2: $BF_{10} = 0.67$), or evidence against it (Exp. 3: $BF_{10} = 0.13$). In Experiment 2, there was evidence *against* a main effect of PI (BF01 = 76.92), whereas in Experiment 3, where the cue was held constant within the participants, there was evidence for a main effect of PI on the probability of choosing the same-trial lure ($BF_{10} = 24.39$).

For the probability of choosing an old intrusion lure there was evidence for both main effects across both Experiments (set size: Exp. 2: $BF_{10} = 2.34 \times 10^{6}$; Exp. 3: $BF_{10} = 1.11 \times 10^{9}$, and PI: Exp. 2: $BF_{10} = 45.96$; and Exp. 3: $BF_{10} = 5.54 \times 10^{5}$). Evidence on the interaction was ambiguous in Exp 2 ($BF_{10} = 0.88$), yet credible in Exp 3 ($BF_{10} = 4.66$). Follow-up analyses revealed that proactive interference increased the probability of choosing the old trial lure for set size 8 ($BF_{10} = 5.41 \times 10^{6}$) but had no effect at set size 2 and at set size 3 ($BF_{01} = 5.79$, and $BF_{01} = 5.73$, respectively).

The analysis of probability of choosing a new item showed that subjects chose new items more often in case there was no PI (main effect $BF_{10} = 48.19$, $BF_{10} = 2.44 \times 10^{100}$). There was no evidence for an effect of set size in Exp. 2 (main effect $BF_{10} = 0.73$), yet credible evidence in Exp. 3 ($BF_{10} = 7.12 \times 10^{12}$).

3.2.3. Very large sets

Third, we were interested in how the set size effect in memory for bindings manifests at a memory load that clearly exceeds WM capacity, namely at set size 30 (Exp. 2) and 16 (Exp. 3), and in how proactive

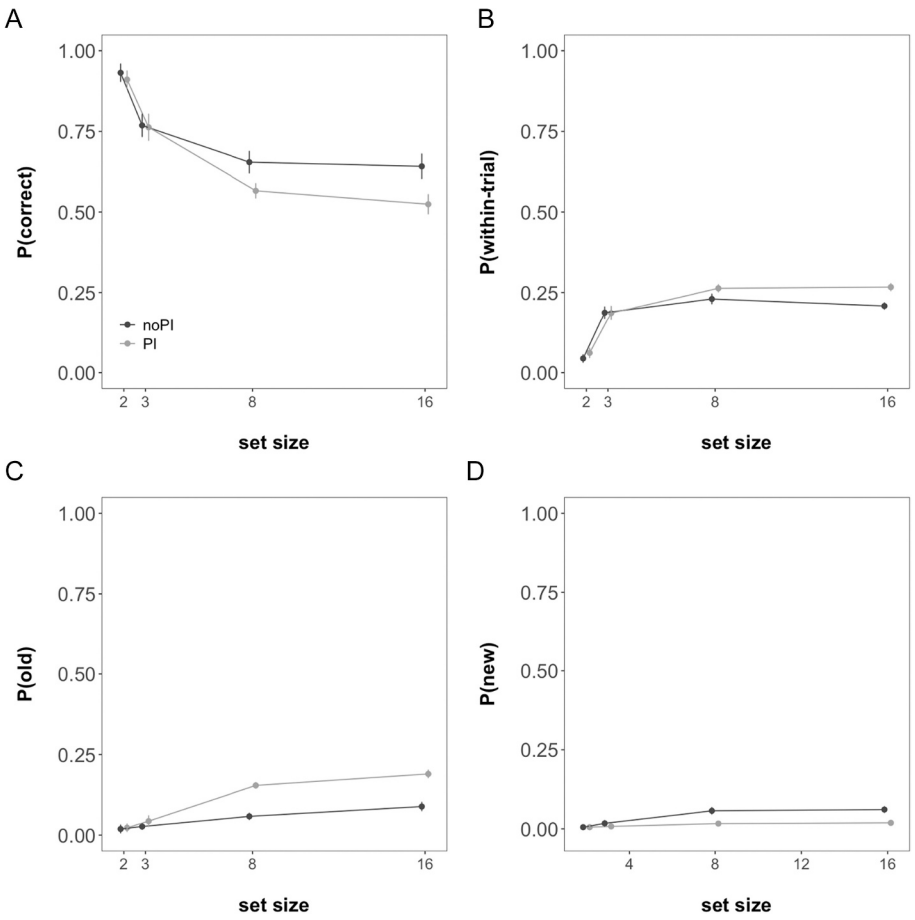

Fig. 7. Immediate memory performance in Experiment 3. Panel A shows the probability of choosing the correct item across set sizes and under conditions with proactive interference realized by a small pool of items (PI) and without (noPI). Panels B, C, and D show the probability of choosing the within-trial lure, the old lure, and the new lure, respectively. Error bars represent the within subject confidence interval.

Table 2
Details of the hierarchical logistic regression models of Experiments 1–5.

Exp.	Formulae of the models	Intra-class correlation coefficient (ICC)				Number of observations
		correct	same trial lure	old trial lure	new lure	
1	response~setsize +(1 + setsize ||subject)	0.15	0.10	0.19	0.35	4400
2	response~setsize :	0.15	0.07	0.17	0.47	13,188
3	PI +(1 + setsize + PI ||subject	0.13	0.05	0.20	0.40	21,715
4	response~setsize :	0.20	0.14	0.17	0.16	21,680
5	distractor +(1 + setsize + distractor ||subject)	0.21	0.17	0.25	0.46	11,480

Note: the response stands for each of the four DVs: correct, same trial lure, old trial lure and new lure responses.

interference affects this pattern. We therefore used a Bayesian hierarchical logistic regression predicting the probability of correct responses for set sizes 6 and 8 compared to set size 30 (Exp. 2) and for set size 8 compared to set size 16 (Exp. 3).

Experiment 2. We pooled the data of set size 6 and 8 as our previous analysis revealed evidence against a difference in performance between these two set sizes ($BF_{01} = 9.18$). Immediate memory performance did not differ between the set sizes 6, 8 (pooled) and 30 ($B_{01} = 132.95$). Surprisingly, there was evidence against an effect of PI at set size 30 ($BF_{01} = 9711$). One reason for this was the larger pool used for this set size: Of the 45 stimuli from which set-size 30 trials were sampled, 33 were used only for set-size 30 trials, and therefore occurred only rarely even in the PI condition, so that the overall level of PI was relatively benign for that set size. We therefore conducted an analysis comparing the performance restricted to the twelve stimuli that were also included in the closed pool of the small set sizes to performance in the open pool at set size 30 and found evidence against a difference ($BF_{01} = 11.08$).

Experiment 3. The analysis revealed that immediate memory performance did not differ between the set sizes 8 and 16 ($B_{01} = 2.69$). There was evidence for a main effect of PI ($BF_{10} = 1026.91$), but evidence against an interaction ($BF_{01} = 40.38$).

Further, we evaluated the estimates of WM capacity based on the number of items remembered across set sizes (Fig. 5B and C and Supplementary Fig. S 1 B and C). These drastically exceeded common estimates of WM capacity – around 4 items – reaching about 16 at the highest set size. Proactive interference affected the K estimates at larger set sizes only, but not at set sizes 2 and 3.

3.2.4. *Input-output distance*

Finally, we analyzed the effect of proactive interference on selecting the correct response across the input-output distance of the pairs presented at set size 6 and 8 (Exp. 2) and 8 and 16 (Exp. 3) – which overall showed credible effects of PI. As can be seen in Fig. 8 and Fig. 9, proactive interference did not affect the probability of choosing the correct response for input-output distances of 1 and 2 at these set sizes (see

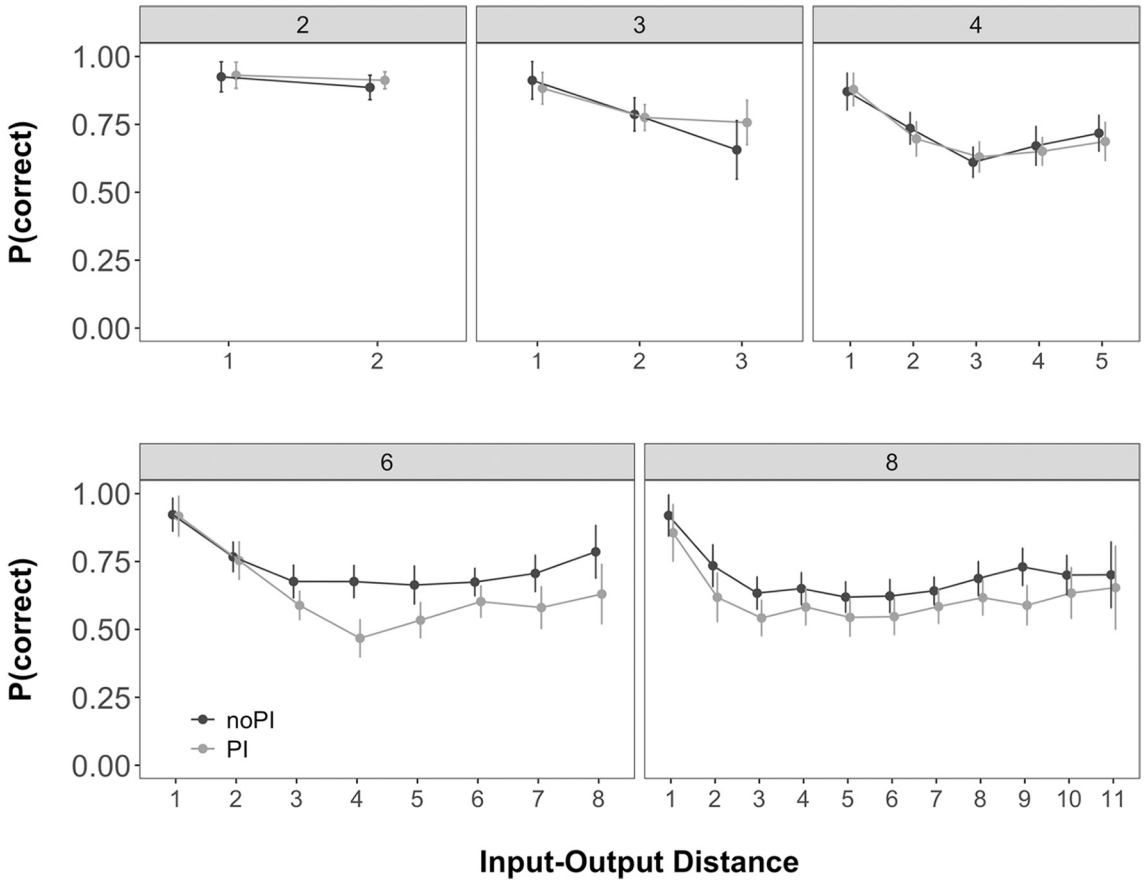

Fig. 8. Probability of choosing the correct item in Experiment 3 under conditions with proactive interference and without, across input-output distances and set sizes. Error bars represent the between subject confidence interval.

Table 3 for BFs).

3.3. Discussion

Results of Experiments 2 and 3 provide evidence for the greater involvement of episodic LTM at larger set sizes, which explains the decreased set size effect at larger set sizes in these experiments, and in Experiment 1. Specifically, proactive interference only negatively affected performance at set sizes 6, 8, and 16, but not at smaller set sizes. This effect became manifest primarily as an increase in erroneous selections of lures from preceding trials, as is expected from proactive interference between trials: Proactive interference arises from the difficulty of discriminating between the relevant current trial and no longer relevant information from previous trials (Gardiner, Craik, & Birtwistle, 1972). In conditions without proactive interference, the set size effect plateaued at larger set sizes (between 4 and 30), replicating Experiment 1. This plateau probably reflects the level of performance that can be reached by relying on episodic LTM alone. Taken together, the results of Experiment 2 and 3 support the notion that the set size effect on memory for bindings plateaus due to the involvement of episodic LTM at larger set sizes.

Further, in the special case that the last presented one or two pairs are retrieved first, participants showed very good performance, which is not affected by PI in the larger set sizes (6, 8 and 16). This is compatible with these items being retrieved from WM, with no contribution from episodic LTM to performance. Whereas for input-output distance 1 the lack of proactive interference could be attributed to a ceiling effect, for distance 2 it cannot. In conclusion, it appears that at larger set sizes performance is largely driven by episodic LTM. This is true in particular for pairs encoded early, or retrieved late, because interference from intervening events corrupts their WM representations to a degree that renders them practically useless, so that drawing on episodic LTM is the best the person can do.

4. Experiment 4 and 5

Now that we have shown the contributions of LTM to larger set sizes, we aimed to gauge the contribution of WM to the task across the set sizes. WM performance has been shown to be impaired by the processing of distractors during the retention interval (see Brown, 1958; Lewandowsky, Geiger, & Oberauer, 2008; Peterson & Peterson, 1959; and Oberauer et al., 2018 Benchmark 2.1 for an overview). Therefore, in Experiments 4 and 5, we tested WM for bindings across a broad range of set sizes under conditions of either no distraction or distraction between encoding and test. The goal was to test the prediction that a distractor-filled delay impairs performance selectively at the small set sizes, where it relies predominantly on WM.

4.1. Method

4.1.1. Participants

We collected data of 30 participants for Experiment 4 (Mean age = 23.9, 27 females) and 30 participants for Experiment 5 (Mean age = 24.05, 13 females).

4.1.2. Materials and procedure

The paradigm followed the one from Experiment 1, apart from including a 15 s distractor task after the presentation of the pairs in half of the trials. In the remaining trials, memory was tested after a 1 s unfilled retention interval as in Experiment 1. In Experiment 4, the

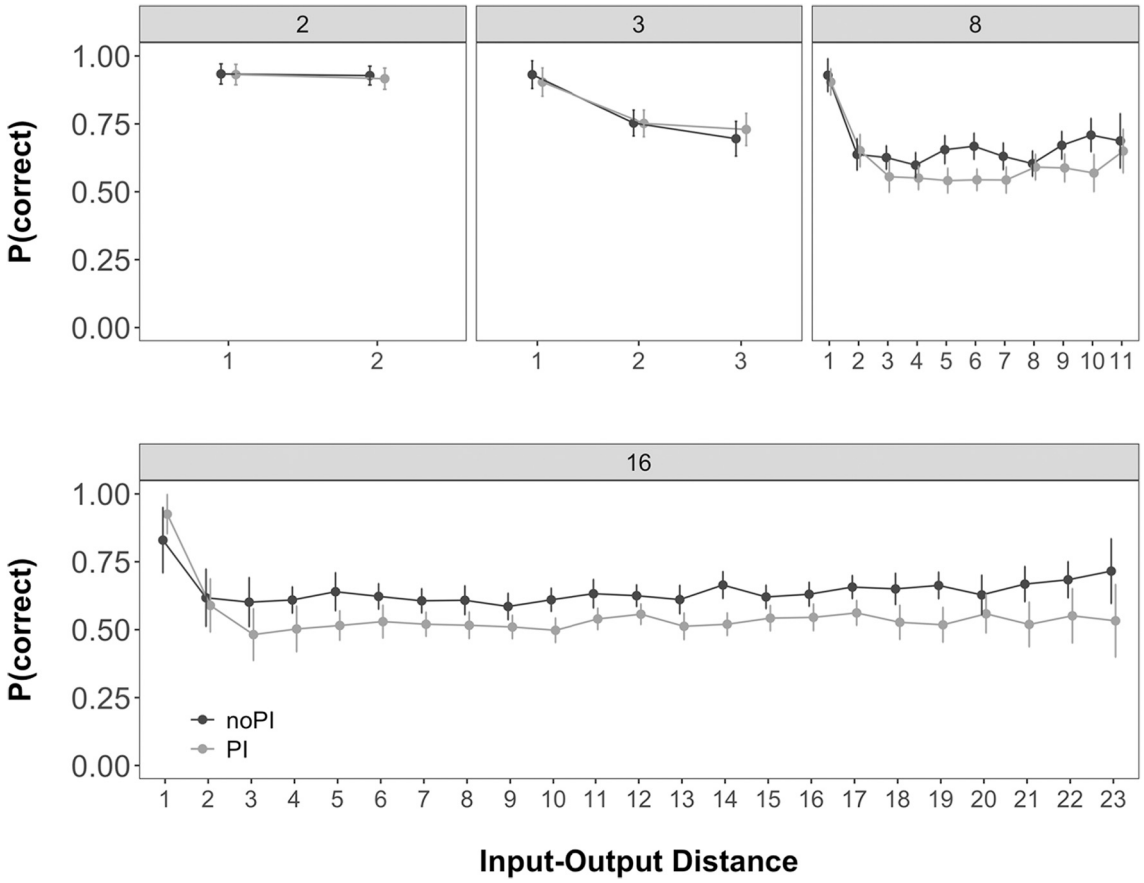

Fig. 9. Probability of choosing the correct item in Experiment 3 under conditions with proactive interference and without across input-output distances and set sizes. Error bars represent the between subject confidence interval.

Table 3
Bayes Factors of the effect of proactive interference at the shortest input-output distances of set sizes 8 and 16, Experiment 2 and 3.

Experiment	Set size	Input-output distance	Effect of PI
2	6	1	$BF_{01} = 6.28$
		2	$BF_{01} = 11.89$
2	8	1	$BF_{01} = 4.79$
		2	$BF_{01} = 2.04$
3	8	1	$BF_{01} = 4.72$
		2	$BF_{01} = 14.92$
3	16	1	$BF_{01} = 5.07$
		2	$BF_{01} = 7.49$

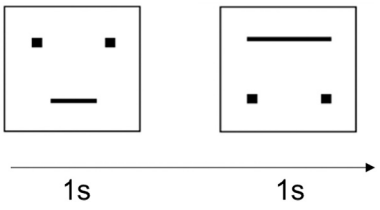

Fig. 10. Example stimuli of the distractor task of Experiment 4.

distractor task consisted of a series of spatial judgments: Participants had to judge with a button press whether a horizontal bar fitted into the gap between two squares (Vergauwe, Barrouillet, & Camos, 2010) (see Fig. 10 for an example stimulus). In Experiment 5, the distractor task comprised the judgement of whether arithmetic equations were correct or not (e.g., $5 \times 4 = 20$?). Within the 15 s retention interval, 15 judgments had to be made, resulting in 1 s presentation time per distractor.[4] Experiment 4 included set sizes 2, 3, 8, and 16. Experiment 5 included set sizes 2, 3, 4, and 6.

Both experiments involved two sessions à one-hour and comprised 208 trials, with 13 trials per condition for each set size. Within each of the 16 blocks of the experiment, each combination of set size and distraction condition (with distraction vs. no distraction) was realized twice; the order of conditions was randomized within each block. Apart from these changes, the paradigm followed the one from Experiment 1.

4.2. Data analysis

We analyzed Experiment 4 and 5 with a Bayesian hierarchical logistic regression predicting correct or incorrect responses in the 4-AFC test by set size and distraction (no distraction vs. distraction). We included a by-participant random intercept and a random slope for *set size* as well as *distraction*. In addition, we estimated the correlation among the random-effects parameters. The fixed effects were *distraction* and *set size* and the latter was included as a continuous predictor plateauing at set sizes 4 and larger (i.e., Exp 4: 8 and 16; Exp 5: 4 and 6) as established in Experiment 1 (i.e., Coding III in Table 1).

[4] We chose to compare a distractor-filled retention interval to a condition with minimal retention interval, as this is a common and generally accepted method for selectively disrupting WM (see Benchmark 2.1 in (Oberauer et al., 2018) Brown, 1958; Lewandowsky et al., 2008; Peterson & Peterson, 1959). For the present purposes, there was no need to disentangle the effects of distractor processing from those of delay.

4.3. Results

Fig. 12 and Fig. 11 show the mean probability of choosing each of the four response options across the set sizes for conditions of *distraction* and *no distraction*, in Experiments 4 and 5, respectively. The interested reader can also find the mean proportion of correct responses broken down by cue-type, serial position as well as output position in the Supplementary (Fig. S2D and E, Fig. S 3D and E, Fig. S 4D and E). Table 2 lists the formulae, intra-class correlations, and number of observations of the full hierarchical logistic regression models.

4.3.1. Accuracy

Our first question was whether distraction affected memory performance differentially across set sizes. Model comparison revealed evidence for an interaction between *set size* and *distraction* in both Experiments 4 and 5 ($BF_{01} = 8.73 \times 10^8$, and $BF_{01} = 1.38 \times 10^4$, respectively) as well as evidence for a main effect of *set size* ($BF_{10} = 3.92 \times 10^{11}$ and $BF_{10} = 1.52 \times 10^8$), yet evidence *against* a main effect of *distraction* ($BF_{01} = 15.33$, and $BF_{01} = 8.94$). Follow-up analyses of Experiment 4 revealed that distraction only (negatively) affected the probability of choosing the correct option for set size 2, ($BF_{10} = 7138$), but had no detrimental effect at set sizes 3, and (pooled) 8 and 16 ($BF_{01} = 4.73$, $BF_{01} = 51.68$, respectively). Equivalently, follow-up analyses of Experiment 5 revealed that distraction only (negatively) affected the probability of choosing the correct option for set size 2, ($BF_{10} = 12.00$), but had no detrimental effect at set sizes 3, and (pooled) 4 and 6 ($BF_{10} = 18.52$, $BF_{10} = 5.18$). Also, the WM capacity estimate K across Experiments 4 and 5 were only negatively affected by the distractor filled RI at set size 2 (see Fig. 5D and E).

4.3.2. Error types

Our second question was whether the set size effect was driven by erroneously choosing the same-trial lure, as predicted from the binding hypothesis, and whether there was an interaction with the presence of the distractor task. Across both Experiments 4 and 5, the probability of choosing the same-trial lure was the only error type showing an interaction between distraction and set size ($BF_{10} = 1.63 \times 10^7$ and $BF_{10} = 3.00 \times 10^6$). Follow-up analyses revealed that only for set size 2 there was an effect of distraction on the probability of choosing the same-trial lure ($BF_{10} = 879.52$ and $BF_{10} = 1452$), whereas there was evidence against this effect at larger set sizes 3 ($BF_{01} = 9.38$ and $BF_{01} = 333.33$) and (pooled) 8 and 16 ($BF_{01} = 49.24$, Exp 4). In Experiment 5, there was evidence for *better* performance in the presence of a distractor task at (pooled) set sizes 4 and 6 ($BF_{10} = 256$).

For all three error types there was evidence for a main effect of set size in both experiments (see Table 4 for BFs). Further, there was evidence *against* the main effect of distraction for same-trial lures and new lures, and anecdotal evidence for an effect of distraction in case of old trial intrusion.

4.3.3. Input-output distance

Lastly, we were interested in how the distractor task affected performance across the input-output distances at all set sizes. Specifically, the results of the previous experiments suggest that the last one or two pairs, when tested early, are retrieved from WM regardless of set size. If that is the case, the distractor task should selectively impair

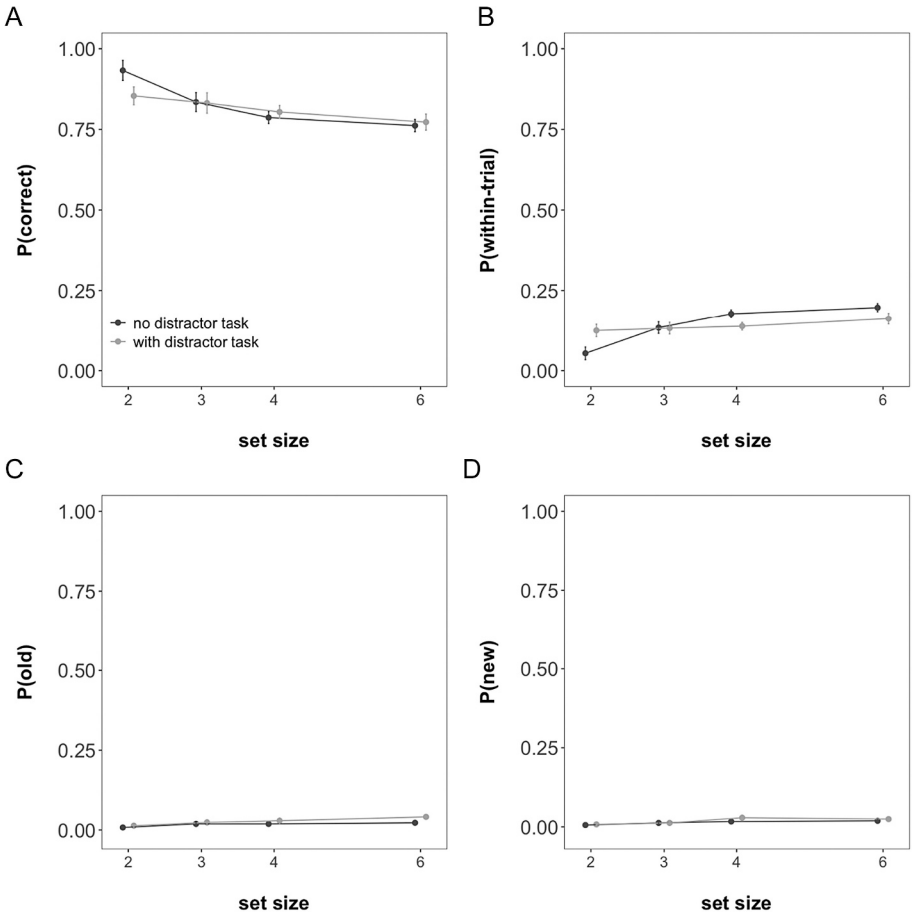

Fig. 11. Immediate memory performance in Experiment 5. Panel A shows the probability of choosing the correct item across set sizes. Panels B, C, and D show the probability of choosing the within-trial lure, the old lure, and the new lure, respectively. The grey line represents the condition with a distraction task following encoding. Error bars represent the within subject confidence interval.

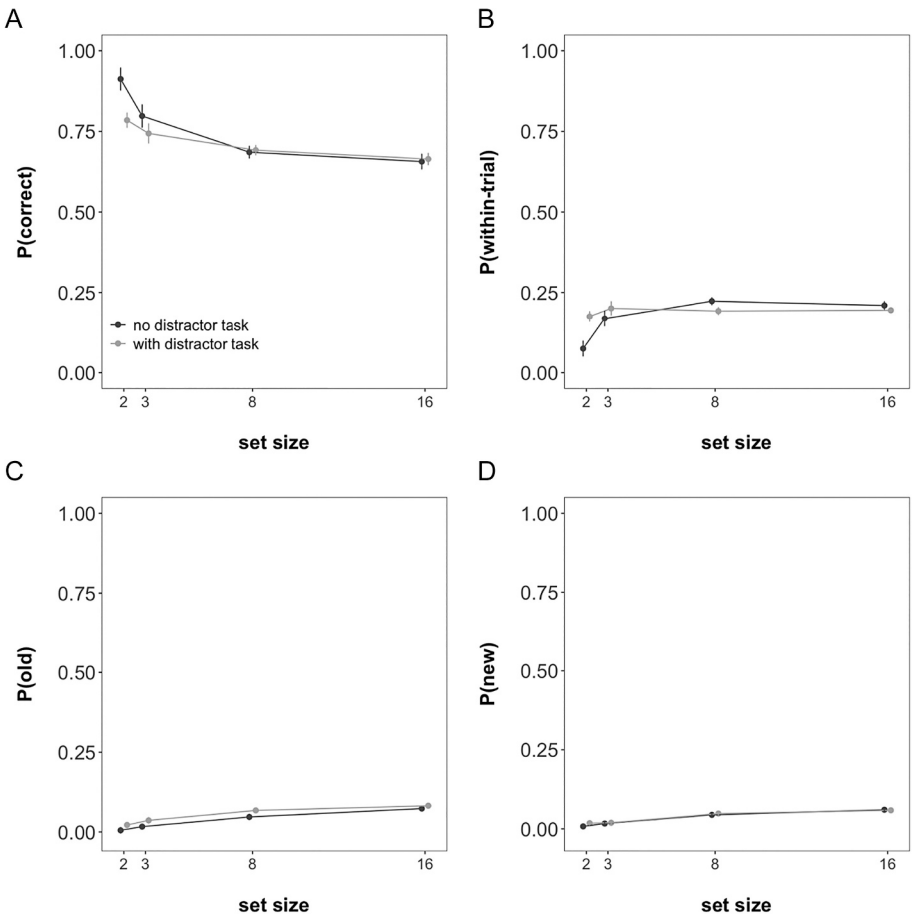

Fig. 12. Immediate memory performance in Experiment 4. Panel A shows the probability of choosing the correct item across set sizes. Panels B, C, and D show the probability of choosing the within-trial lure, the old lure, and the new lure, respectively. The grey line represents the condition with a distraction task following encoding. Error bars represent the within subject confidence interval.

performance at short input- output distances, even for larger set sizes. As shown in Fig. 13 and Fig. 14, the distractor task affected the probability of choosing the correct response for small input-output distances (1, and sometimes 2) at set sizes 2 and sometimes 3, but not at set size 4, 6, 8 and 16 (see Table 5 and Table 6 for BFs). This suggests that in the present experiment, at larger set sizes even the pairs with minimal input-output distance were often retrieved from episodic LTM rather than WM.

4.4. Discussion

In Experiment 4 and 5 we showed that a distraction-filled delay following encoding of word-picture pairs selectively impairs performance for set size 2 but has no detrimental impact at larger set sizes. This indicates that WM contributes more strongly to performance at the smallest set size (and to some extent, still at set size 3 in Exp. 4), yet at larger set sizes, performance is already largely or exclusively driven by LTM, and therefore it is rather unaffected by distractor processing during the retention interval. Together with Experiments 2 and 3, in which proactive interference selectively harmed performance at larger set sizes, the results of Experiments 4 and 5 constitute a double dissociation of the contributions of WM and LTM across set sizes of word-pairs.

Under most circumstances, a double dissociation is compelling evidence for the distinction of two mechanisms or processes, but there are conditions in which that conclusion is not valid. State-trace analysis offers a more solid foundation for assessing the evidence for a distinction (see Newell & Dunn, 2008, for an introduction). The rationale is the following: We need two dependent variables measuring the two hypothetical mechanisms or processes, respectively. In our case, we assume that memory performance at small set sizes reflects predominantly WM, whereas performance at larger set sizes reflects predominantly episodic LTM. Further, we need two independent variables for which we assess their effect on both dependent variables. Here, these are the PI manipulation, and a distractor-filled retention interval. The state trace represents the relation of the two dependent variables across all experimental conditions. In the present case, we plot performance at small set sizes on the x-axis, and performance at larger set sizes in the corresponding condition on the y-axis (see Fig. 15).

We consider two alternative theoretical scenarios. In the 1-dimensional scenario there is only a single latent dimension of memory strength, or memory accessibility, that determines performance for all set sizes. This dimension is affected more or less strongly by the two independent variables. The latent dimension is translated into performance by an unknown, but monotonically increasing function. If that is the case, then all points in the state trace should lie on a single monotonically increasing function. In the alternative two-dimensional scenario, there are two latent memory dimensions, one for WM – predominantly affecting performance at small set sizes – and one for episodic LTM – primarily affecting performance at large set sizes. One independent variable could affect one dimension, and the other independent variable the other dimension. As a result, the data points in the state trace don't need to fall on a single monotonic function. The state trace in Fig. 15 clearly supports the 2-dimensional scenario.

5. General discussion

The present results make two contributions to our understanding of

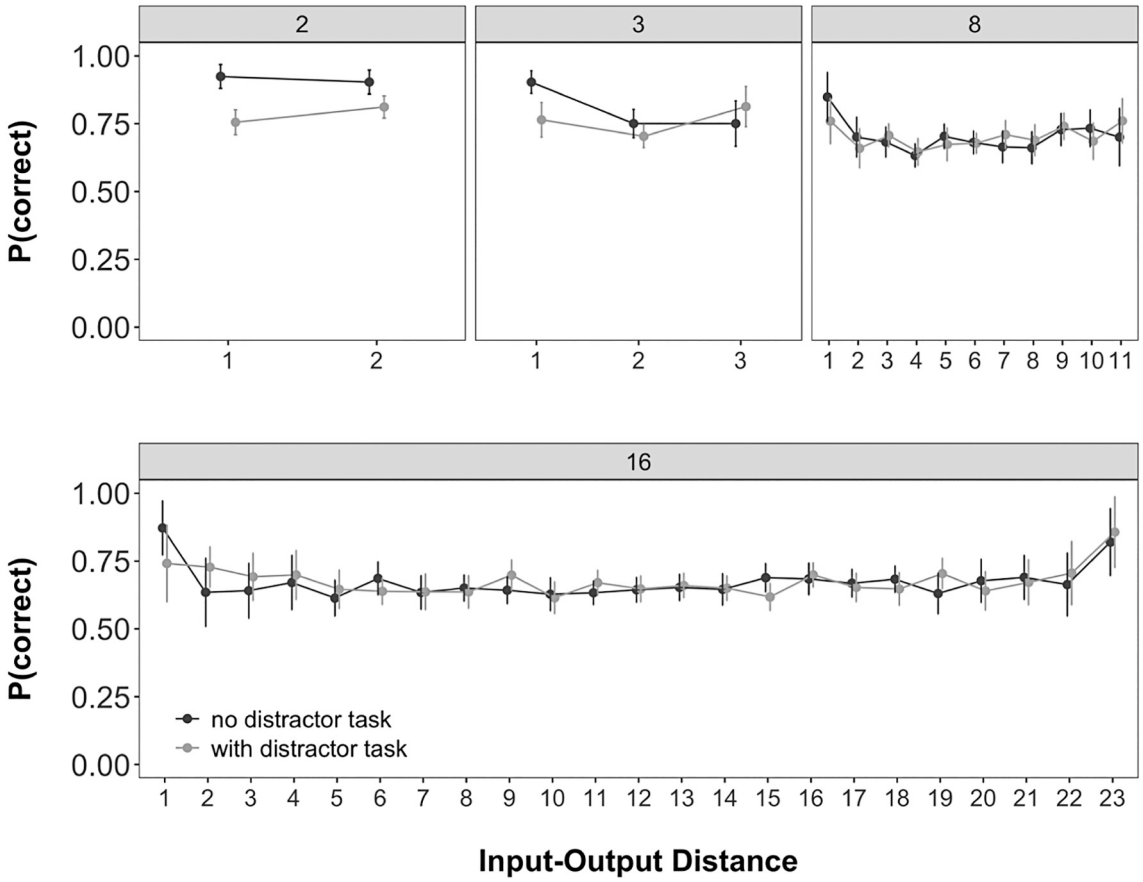

Fig. 13. Probability of choosing the correct item in Experiment 4 under conditions with and without a distractor task across input-output distances and set sizes. Error bars represent the between subject confidence interval.

working memory: First, we provided a new test of the binding hypothesis of WM capacity by testing people's ability to maintain bindings between items. Across five experiments we showed that WM capacity is indeed limited predominantly – though perhaps not exclusively – by the number of bindings between items. Second, we showed a double dissociation of contributions of WM and episodic LTM to performance in this immediate memory task: Performance at set sizes larger than 3 were specifically affected by proactive interference – but were immune to influences from a distractor filled delay. In contrast, performance at set size 2 was unaffected by proactive interference but harmed by a distractor-filled delay.

5.1. *WM capacity limits on the bindings between items*

Our first goal was to test the binding hypothesis with a new WM test which deconfounds the number of bindings from the distinctiveness of the retrieval cues to which stimuli are bound. In this way, the effect of set size on performance can be attributed more unambiguously to the capacity limit of WM.

The binding hypothesis predicts that with increasing set size, memory for bindings should decline, whereas memory for items should be largely unimpaired (Oberauer, 2019). In Experiments 1 and 2 we provided support for this prediction with a new WM task in which the distinctiveness of the retrieval cues – words or pictures – is independent of set size. We showed that the set size effect was driven by reduced memory for bindings, which became manifest in an increase in probability to choose the same-trial lure but not of the remaining options. By contrast, item memory – reflected in the ability to reject new lures – was unaffected by set size. Experiments 3, 4 and 5 replicated the finding that the large majority of errors committed at larger set sizes were binding errors, that is, errors of selecting an item from the current trial, but belonging to another pair than the tested one. However, in these experiments there was also evidence for a set-size effect on the other two error types – choosing old items, and new items. The shape of these set-size effects differed from that on within-trial lures: Whereas the prevalence of the latter increased steeply from set size 2 to 4, and decelerated afterwards, the proportion of old and new lures increased in a shallow, approximately linear fashion with set size. This gradual increase is reminiscent of the fairly shallow list-length effect observed in item recognition tests of episodic LTM (Annis, Lenes, Westfall, Criss, & Malmberg, 2015; Osth & Dennis, 2015, for reviews and potential explanations of that effect). Moreover, the increase of old-trial lures with set size was modulated by PI, suggesting that it arises from the fact that, at larger set sizes, most pairs are retrieved from episodic LTM. Nevertheless, we cannot rule out that WM capacity also plays a role in limiting item memory at higher set sizes in Experiments 3 and 4 (see Cowan, 2022). We conclude that the WM capacity limit is primarily, but perhaps not exclusively, a limit on the ability to maintain bindings.

5.2. *The contribution of LTM to tests of WM*

Based on the unexpected finding that the detrimental effect of set size on WM decelerated after 4 pairs – with participants performing comparably in the immediate memory task across set sizes 4, 6, and 8 – we investigated the differential contribution of LTM to our relational WM task. The results across Experiments 2–5 showed that our manipulations that either targeted WM (distractor-filled retention interval, Exp. 4 and 5) or LTM (PI, Exp. 2 and 3) also affected the types of errors differentially: Specifically, in Experiments 4 and 5, including a distractor-filled retention interval specifically increased the probability

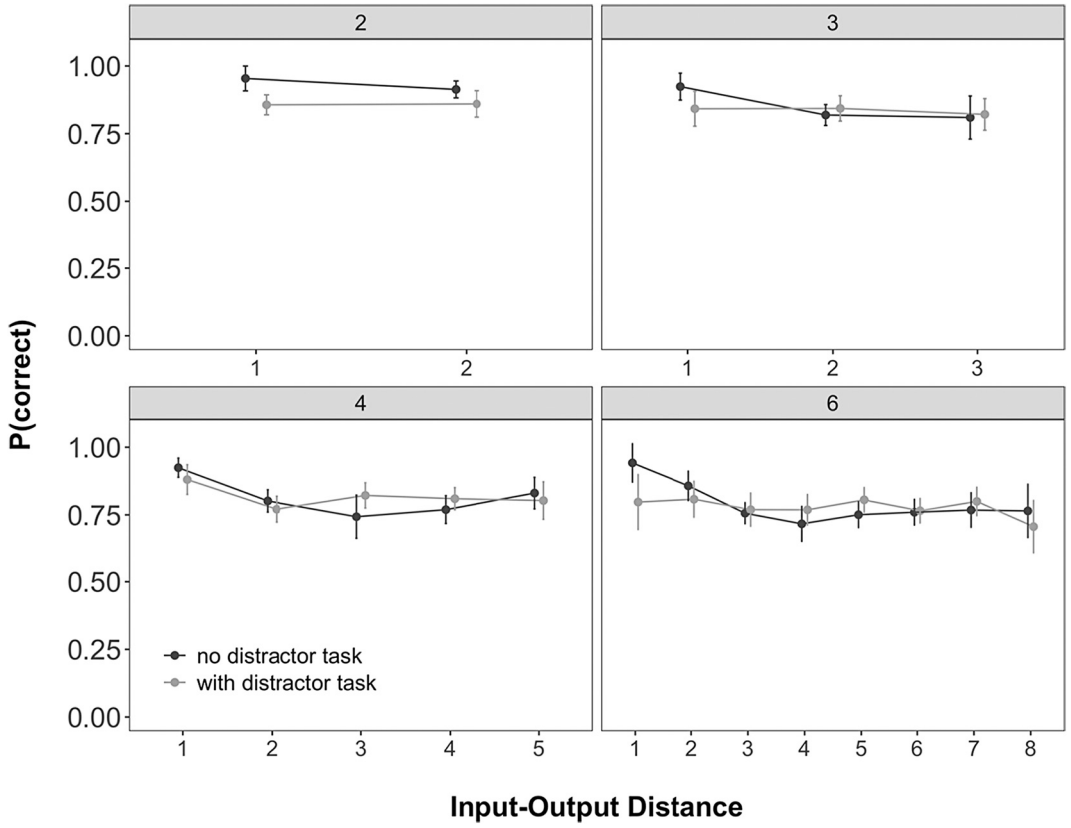

Fig. 14. Probability of choosing the correct item in Experiment 5 under conditions with and without a distractor task across input-output distances and set sizes. Error bars represent the between subject confidence interval.

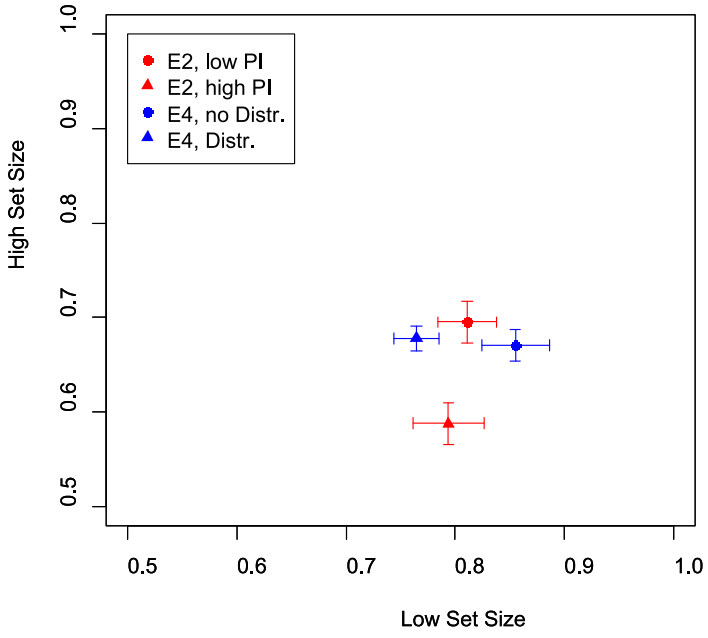

Fig. 15. State-trace plot of the data from Experiments 2 and 4. The proportion of correct responses of each participant was averaged over low set sizes (2, 3, or 4 pairs), and high set sizes (> 4 pairs, excluding 30 pairs in Experiment 2). Each data point represents one experimental condition for which we have a measure of performance at both set-size levels. Error bars are 95% confidence intervals for within-subjects' comparisons.

of choosing the same-trial lure at the small set size 2 – but not at larger set sizes. This is what we should expect if WM is specifically responsible for maintaining bindings at small set sizes, and the distractor task selectively impaired WM. In Experiments 2 and 3, which manipulated the amount of PI and thereby the usefulness of episodic LTM, PI resulted in a differential increase of errors of choosing the old trial lure at larger set sizes. This is to be expected as PI reduces the distinctiveness between trials in episodic LTM, making it harder to determine whether a stimulus

Table 4
Bayes Factors of the main effects of set size and distractor task across Experiments 4 and 5.

	main effect of set size		main effect of distractor	
	E4	E5	E4	E5
within-trial lures	2506	3.45×10^{15}	0.016	0.187
old lures	7.97×10^{9}	1065	2.83	2.63
new lures	8.27×10^{6}	5.07	0.017	0.10

Table 5
Bayes Factors of the effect of distractor task across set sizes, depending on input-output distance of Experiment 4. Values larger than 3 represent evidence for an effect of the distractor task and are printed in bold, values below 0.33 evidence against an effect.

Set size	Input-output distance	Effect of distractor task
2	1	$\mathbf{BF_{10} = 67.55}$
	2	$\mathbf{BF_{10} = 25.92}$
3	1	$\mathbf{BF_{10} = 5.54}$
	2	$BF_{10} = 0.15$
8	1	$BF_{10} = 0.26$
	2	$BF_{10} = 0.09$
16	1	$BF_{01} = 0.26$
	2	$BF_{01} = 0.18$

Table 6
Bayes Factors of the effect of distractor task across set sizes, depending on input-output distance of Experiment 5. Values larger than 3 represent evidence for an effect of the distractor task and are printed in bold, values below 0.33 evidence against.

Set size	Input-output distance	Effect of distractor task
2	1	$\mathbf{BF_{10} = 2.67}$
	2	$BF_{10} = 0.71$
3	1	$BF_{10} = 0.68$
	2	$BF_{10} = 0.10$
4	1	$BF_{10} = 0.27$
	2	$BF_{10} = 0.06$
6	1	$BF_{01} = 0.98$
	2	$BF_{01} = 0.08$

has been presented in the current or in a previous trial.

A further exploratory analysis revealed that performance was immune to PI even at larger set sizes when very few encoding or retrieval events intervened between the presentation of a pair and its test. This suggests that the last-presented pairs are maintained in WM, and are accessed from WM as long as they are not corrupted by output interference. However, we did not find that these pairs are vulnerable to a distractor task in Experiment 4 and 5, so that our conclusion about which pairs are preferentially maintained in WM at larger set sizes remains tentative.

How do WM and episodic LTM work together? Based on the present results we envision their cooperation as follows: Each successive pair results in the activation of representations of its elements (i.e., the concepts corresponding to the word and the object) in semantic memory. The relation between them is encoded in WM through a temporary binding. In addition, an episodic memory trace is created that represents the elements and their relation as an integrated event (see Cowan, 2019; Cowan & Chen, 2008). This means, that items even in short lists are represented simultaneously in both eLTM and WM. As successive pairs are presented, interference between bindings in WM builds up. If we assume that interference is predominantly retroactive, this leads to a gradual loss of accessibility of earlier pairs in WM. Alternatively, interference could be symmetric, but earlier-presented pairs are removed from WM to prevent them from interfering with later-presented pairs. Either way, the bindings of the last-presented pairs are most accessible in WM at the onset of test.

When a pair is tested, the given element is first used as retrieval cue to access its counterpart through their binding in WM. If that fails, an attempt is made to retrieve the tested pair from episodic memory. Both retrieval attempts draw on additional information from the activation of representations in semantic memory: The re-activation of a word or object through cue-based retrieval – from WM or episodic LTM – is combined with its persistent activation in semantic memory. Whereas the former contributes information about the relations between items, the latter contributes item information. It helps excluding new words or objects from the response alternatives, and in the low PI condition, it also helps to exclude items from previous trials. In the high PI condition, however, all words and objects in the small pool are chronically activated in semantic memory, and therefore, item activation carries little information. Discriminating between the current and previous trials has to rely on cue-based retrieval, using either bindings in WM, or episodic memory representations. To the extent that it relies on WM, we should expect the prevalence of intrusions from old lists to increase with set size, as we found to be the case in the high-PI condition.

As mentioned in the introduction, a contribution of episodic LTM to performance in WM tasks has been discussed for some time. Nevertheless, the assumption of such a contribution leaves us with a conundrum: Episodic LTM is not constrained by a capacity limit – so why is performance in WM tests so severely limited? In particular, why is people's ability to remember information in WM tests virtually perfect at set sizes up to about 2 or 3, and then breaks down rapidly as set size is increased further, as illustrated in Fig. 1? Part of the answer is probably that typical WM tasks involve a high level of proactive interference between trials: In serial-recall trials, memory for order is maintained by binding each list item to a representation of its list position, so that the position can be used as retrieval cue for the item at recall (Lewandowsky & Farrell, 2008). Across trials, the same position representations act as retrieval cues (Fischer-baum & McCloskey, 2015), so that proactive interference builds up in episodic memory: Each position is associated to more and more items – one from every preceding trial – that compete for retrieval with the target item from the current trial. Likewise, in most tasks for investigating visual WM, stimuli are presented as spatially distributed arrays, and at test, the target stimulus is identified by its location. The same, or very similar, locations are re-used across trials, so that locations as retrieval cues are associated with more and more different stimuli in episodic LTM. By contrast, WM escapes proactive interference between trials because WM is cleared after every trial, and the contents of WM are protected against interference from LTM (Oberauer, 2009b; Oberauer et al., 2017). Therefore, in settings with high PI across trials, the contribution of episodic LTM is likely to be modest, and performance is strongly limited by the capacity of WM, leading to a precipitous decline with increasing set size.

By contrast, when PI is much reduced, people can rely on episodic LTM to circumvent the capacity limit of WM. Previous research has already shown this for immediate tests of item memory (Endress & Potter, 2014). The novel contribution of the present work is to show this phenomenon also for relational memory: In conditions without proactive interference, people can memorize 20 – and perhaps more – rapidly presented arbitrary pairs of stimuli, far exceeding any estimate of WM capacity. Proactive interference impairs that ability, showing that it relies on episodic LTM. A further indication of a stronger contribution of LTM to the WM task was that at larger set sizes the type of errors shifted from same-trial intrusions to old list intrusions.

Nevertheless, PI alone cannot fully explain why the present task revealed a less severe limitation on immediate memory than more typical WM tasks. Even in the conditions with high PI, the set-size effects beyond four pairs were rather shallow, and the capacity estimates from the highest set sizes – in particular in Experiment 2 – were unusually high. We suspect that this might be because we used retrieval cues that are relatively well discriminable because they have representations in a high-dimensional semantic (and, in the case of pictures, also visual)

feature space. In contrast, the retrieval cues of typical WM tasks – in most cases, ordinal list positions or spatial locations in an array – are less discriminable, as they inhabit one- or two-dimensional spaces, which become increasingly crowded as set size increases. Highly distinctive retrieval cues could enable episodic LTM to work reasonably well even in the face of proactive interference.

6. Conclusion

The present study makes three contributions to our understanding of working memory: First, the WM capacity limit is primarily, but perhaps not exclusively, a limit on the ability to maintain bindings – in this case to maintain bindings between items. Second, we showed a double dissociation of contributions of WM and episodic LTM to performance in this immediate memory task: At set sizes larger than 3, LTM strongly contributed to performance, whereas performance at set size 2 was driven by WM. Third, we propose, and empirically support, methods for gauging the relative contributions of WM and episodic LTM to performance in the immediate memory tasks that researchers use to test and investigate WM: Proactive interference can be used as an index of the contribution of episodic LTM. Conversely, vulnerability to a distractor task can be used as an index of the contribution of WM.

Author note

We thank Atalia Adank, Joscha Dutli, Gary Hoppeler, Samuel Pawel, and Dawid Strzelczyk for helping with data collection. The data and the analysis scripts can be accessed in the Open Science Framework (https://osf.io/ymgkq). This research was supported by a grant from the Swiss National Science Foundation to K. Oberauer (project 100014_179002).

CRediT authorship contribution statement

Lea M. Bartsch: Conceptualization, Methodology, Data curation, Formal analysis, Software, Visualization, Writing – original draft. **Klaus Oberauer:** Conceptualization, Methodology, Funding acquisition, Supervision, Software, Writing – review & editing.

Data availability

The data and the analysis scripts can be accessed in the Open Science Framework (https://osf.io/ymgkq).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2022.105330.

References

Adam, K. C. S., Mance, I., Fukuda, K., & Vogel, E. K. (2015). The contribution of attentional lapses to individual differences in visual working memory capacity. *Journal of Cognitive Neuroscience, 27*(8), 1601–1616.

Adam, K. C. S., Vogel, E. K., & Awh, E. (2017). Clear evidence for item limits in visual working memory. *Cognitive Psychology, 97*, 79–97. https://doi.org/10.1016/j.cogpsych.2017.07.001

Annis, J., Lenes, J. G., Westfall, H. A., Criss, A. H., & Malmberg, K. J. (2015). The list-length effect does not discriminate between models of recognition memory. *Journal of Memory and Language, 85*, 27–41.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In *, Vol. 2, Issue C. Psychology of learning and motivation* (pp. 89–195). Elsevier. https://doi.org/10.1016/S0079-7421(08)60422-3.

Baddeley, A. D. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology, 63*, 1–29.

Barrouillet, P., & Camos, V. (2012). As time goes by: Temporal constraints in working memory. *Current Directions in Psychological Science, 21*(6), 413–419. https://doi.org/10.1177/0963721412459513

Bartsch, L. M., Loaiza, V. M. V. M., & Oberauer, K. (2019). Does limited working memory capacity underlie age differences in associative long-term memory? *Psychology and Aging, 34*(2), 268–281. https://doi.org/10.1037/pag0000317

Beukers, A. O., Buschman, T. J., Cohen, J. D., & Norman, K. A. (2021). Is activity silent working memory simply episodic memory? *Trends in Cognitive Sciences, 25*(4), 284–293.

Bjork, R. A., & Whitten, W. B. (1974). Recency-sensitive retrieval processes in long-term free recall. *Cognitive Psychology, 6*(2), 173–189.

Botvinick, M., & Watanabe, T. (2007). From numerosity to ordinal rank: A gain-field model of serial order representation in cortical working memory. *Journal of Neuroscience, 27*(32), 8636–8642.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences, 105*(38), 14325–14329. https://doi.org/10.1073/pnas.0803390105

Brown, G., Neath, I., & Chater, N. (2002). *A ratio model of scale-invariant memory and identification*. Manuscript Submitted for Publication.

Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology, 10*(1), 12–21. https://doi.org/10.1080/17470215808416249

Bunting, M., Cowan, N., & Scott Saults, J. (2006). How does running memory span work? *Quarterly Journal of Experimental Psychology, 59*(10), 1691–1700.

Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28.

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *R Journal, 10*(1), 395–411. https://doi.org/10.32614/rj-2018-017

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). *Stan*: A probabilistic programming language. *Journal of Statistical Software, 76*(1). https://doi.org/10.18637/jss.v076.i01

Chuderski, A. (2019). Even a single trivial binding of information is critical for fluid intelligence. *Intelligence, 77*, Article 101396.

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

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*(1), 87–114. https://doi.org/10.1017/S0140525X01003922

Cowan, N. (2005). Working memory capacity. In *Working memory capacity*. Psychology Press. https://doi.org/10.4324/9780203342398.

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? *Progress in Brain Research, 169*(07), 323–338. https://doi.org/10.1016/S0079-6123(07)00020-9

Cowan, N. (2019). *Short-term memory based on activated long-term memory: A review in response to Norris (2017)*.

Cowan, N. (2022). Item-position binding capacity limits and word limits in working memory: A reanalysis of Oberauer (2019). *Journal of Cognition, 5*(1), 1–11. https://doi.org/10.5334/JOC.193

Cowan, N., & Chen, Z. (2008). How chunks form in long-term memory and affect short-term memory limits. In *Interactions between short-term and long-term memory in the verbal domain* (pp. 98–119). Psychology Press.

Cowan, N., Johnson, T. D., & Scott Saults, J. (2005). Capacity limits in list item recognition: Evidence from proactive interference. *Memory, 13*(3–4), 293–299.

Crowder, R. G. (1982). The demise of short-term memory. *Acta Psychologica, 50*(3), 291–323.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Memory and Language, 19*(4), 450.

Davelaar, E. J., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H. J., & Usher, M. (2005). The demise of short-term memory revisited: Empirical and computational investigations of recency effects. *Psychological Review, 112*(1), 3.

Endress, A. D., & Potter, M. C. (2014). Large capacity temporary visual memory. *Journal of Experimental Psychology: General, 143*(2), 548.

Eriksson, J., Vogel, E. K., Lansner, A., Bergström, F., & Nyberg, L. (2015). Neurocognitive architecture of working memory. *Neuron, 88*(1), 33–46. https://doi.org/10.1016/j.neuron.2015.09.020

Fischer-baum, S., & McCloskey, M. (2015). Representation of item position in immediate serial recall: Evidence from intrusion errors. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 41*(5), 1426–1446. https://doi.org/10.1037/xlm0000102

Gardiner, J. M., Craik, F. I. M., & Birtwistle, J. (1972). Retrieval cues and release from proactive inhibition. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 778–783.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman & Hall/CRC.

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

Grenfell-Essam, R., & Ward, G. (2012). Examining the relationship between free recall and immediate serial recall: The role of list length, strategy use, and test expectancy. *Journal of Memory and Language, 67*(1), 106–148.

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2017). *Bridgesampling: An R package for estimating normalizing constants*. ArXiv Preprint ArXiv:1710.08162.

Halford, G. S., Maybery, M. T., & Bain, J. D. (1988). Set-size effects in primary memory: An age-related capacity limitation? *Memory & Cognition, 16*(5), 480–487.

Henson, R. N. A. (1996). Unchained memory: Error patterns rule out chaining models of immediate serial recall. *The Quarterly Journal of Experimental Psychology Section A, 49*(1), 80–115.

Henson, R. N. A. (1998). Short-term memory for serial order: The start-end model. *Cognitive Psychology, 36*(2), 73–137.

Huebner, G. M., & Gegenfurtner, K. R. (2011). The efficiency of encoding: Limits of information transfer into memory. *Attention, Perception, & Psychophysics, 73*(5), 1503–1521.
Jonides, J., Schumacher, E. H., Smith, E. E., Lauber, E. J., Awh, E., Minoshima, S., & Koeppe, R. A. (1997). Verbal working memory load affects regional brain activation as measured by PET. *Journal of Cognitive Neuroscience, 9*(4), 462–475.
Lewandowsky, S., & Farrell, S. (2008). Short-term memory: New data and a model. *Psychology of Learning and Motivation, 49*, 1–48.
Lewandowsky, S., Geiger, S. M., & Oberauer, K. (2008). Interference-based forgetting in verbal short-term memory. *Journal of Memory and Language, 59*(2), 200–222. https://doi.org/10.1016/j.jml.2008.04.004
Lewis-Peacock, J. A., & Postle, B. R. (2008). Temporary activation of long-term memory supports working memory. *28*(35), 8765–8771. https://doi.org/10.1523/JNEUROSCI.1953-08.2008.Temporary
Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience, 17*(3), 347.
McCabe, D. P. (2008). The role of covert retrieval in working memory span tasks: Evidence from delayed recall tests. *Journal of Memory and Language, 58*(2), 480–494. https://doi.org/10.1016/j.jml.2007.04.004
McElree, B., & Dosher, B. A. (1989). Serial position and set size in short-term memory: The time course of recognition. *Journal of Experimental Psychology: General, 118*(4), 346.
Melton, A. W. (1963). Implications of short-term memory for a general theory of memory. *Journal of Verbal Learning and Verbal Behavior, 2*(1), 1–21.
Murdock, B. B., Jr. (1963). Short-term retention of single paired associates. *Journal of Experimental Psychology, 65*(5), 433–443. https://doi.org/10.1037/h0040191
Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition, 18*(3), 251–269.
Nairne, J. S. (2002). Remembering over the short-term: The case against the standard model. *Annual Review of Psychology, 53*(1), 53–81.
Newell, B. R., & Dunn, J. C. (2008). Dimensions in data: Testing psychological models using state-trace analysis. *Trends in cognitive sciences, 12*(8), 285–290. ISO 690.
Oberauer, K. (2002). Access to information in working memory: Exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(3), 411.
Oberauer, K. (2009a). Chapter 2 Design for a Working Memory. *Psychology of Learning and Motivation - Advances in Research and Theory, 51*, 45–100. https://doi.org/10.1016/S0079-7421(09)51002-X
Oberauer, K. (2009b). Design for a working memory. *Psychology of Learning and Motivation, 51*, 45–100.
Oberauer, K. (2019). Working memory capacity limits memory for bindings. *Journal of Cognition, 2*(1).
Oberauer, K., Awh, E., & Sutterer, D. W. (2017). The role of long-term memory in a test of visual working memory: Proactive facilitation but no proactive interference. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 43*(1), 1–22. https://doi.org/10.1037/xlm0000302
Oberauer, K., & Kliegl, R. (2001). Beyond resources: Formal models of complexity effects and age differences in working memory. *European Journal of Cognitive Psychology, 13*(1–2), 187–215. https://doi.org/10.1080/09541440042000278
Oberauer, K., Lewandowsky, S., Awh, E., Brown, G., Conway, A., Cowan, N., … Hurlstone, M. J. (2018). Benchmarks for models of short-term and working memory. *Psychological Bulletin, 144*(9), 885.
Osth, A. F., & Dennis, S. (2015). Sources of interference in item and associative recognition memory. *Psychological Review, 122*(2), 260.
Pertzov, Y., Dong, M. Y., Peich, M.-C., & Husain, M. (2012). Forgetting what was where: The fragility of object-location binding. *PLoS One, 7*(10), Article e48214.
Peterson, L., & Peterson, M. J. (1959). Short-term retention of individual verbal items. In *, vol. 58, Issue 3. Journal of experimental psychology* (pp. 193–198). American Psychological Association. https://doi.org/10.1037/h0049234.
Ranganath, C. (2006). Working memory for visual objects: Complementary roles of inferior temporal, medial temporal, and prefrontal cortex. *Neuroscience, 139*(1), 277–289. https://doi.org/10.1016/j.neuroscience.2005.06.092
Ranganath, C., & Blumenfeld, R. S. (2005). Doubts about double dissociations between short- and long-term memory. *Trends in Cognitive Sciences, 9*(8), 374–380. https://doi.org/10.1016/j.tics.2005.06.009
Ranganath, C., Cohen, M. M. X., & Brozinsky, C. C. J. (2005). Working memory maintenance contributes to long-term memory formation: Neural and behavioral evidence. *Journal of Cognitive Neuroscience, 17*(7), 994–1010. https://doi.org/10.1162/0898929054475118
Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review, 21*(2), 301–308.
Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods, 22*(2), 322.
Sederberg, P. B., Howard, M. W., & Kahana, M. J. (2008). A context-based theory of recency and contiguity in free recall. *Psychological Review, 115*(4), 893.
Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—Retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*(2), 145–166.
Tamez, E., Myerson, J., & Hale, S. (2008). Learning, working memory, and intelligence revisited. *Behavioural Processes, 78*(2), 240–245.
Tamez, E., Myerson, J., & Hale, S. (2012). Contributions of associative learning to age and individual differences in fluid intelligence. *Intelligence, 40*(5), 518–529.
Unsworth, N. (2010). Interference control, working memory capacity, and cognitive abilities: A latent variable analysis. *Intelligence, 38*(2), 255–267.
Unsworth, N., & Engle, R. W. (2006). Simple and complex memory spans and their relation to fluid abilities: Evidence from list-length effects. *Journal of Memory and Language, 54*(1), 68–80.
Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review, 114*(1), 104.
Vergauwe, E., Barrouillet, P., & Camos, V. (2010). Do mental processes share a domain-general resource? *Psychological science, 21*(3), 384–390.
Verhaeghen, P., & Basak, C. (2005). Ageing and switching of the focus of attention in working memory: Results from a modified N-Back task. *The Quarterly Journal of Experimental Psychology Section A, 58*(1), 134–154.
Verhaeghen, P., Vandenbroucke, V., & Dierckx, A. (1998). Growing slower and less accurate: Adult age differences in time-accuracy functions for recall and recognition from episodic memory. *Experimental Aging Research, 24*(1), 3–19. https://doi.org/10.1080/036107398244337
Vo, M. L. H., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J., & Jacobs, A. M. (2009). The Berlin affective word list reloaded (BAWL-R). *Behavior Research Methods, 41*(2), 534–538.
Ward, G., Tan, L., & Grenfell-Essam, R. (2010). Examining the relationship between free recall and immediate serial recall: The effects of list length and output order. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(5), 1207.
Wickens, D. D., Born, D. G., & Allen, C. K. (1963). Proactive inhibition and item similarity in short-term memory. *Journal of Verbal Learning and Verbal Behavior, 2*(5–6), 440–445.
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), 1.
Wilhelm, O., Hildebrandt, A., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Frontiers in Psychology, 4*(JUL), 1–22. https://doi.org/10.3389/fpsyg.2013.00433
Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453*(7192), 233.