



Repeated failures to obtain selective directed forgetting in lab and online samples and variations in stimuli

Melisa Akan & Lili Sahakyan

To cite this article: Melisa Akan & Lili Sahakyan (2018) Repeated failures to obtain selective directed forgetting in lab and online samples and variations in stimuli, *Memory*, 26:3, 294-305, DOI: [10.1080/09658211.2017.1327600](https://doi.org/10.1080/09658211.2017.1327600)

To link to this article: <https://doi.org/10.1080/09658211.2017.1327600>



Published online: 15 May 2017.



Submit your article to this journal [↗](#)



Article views: 117



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 1 View citing articles [↗](#)



Repeated failures to obtain selective directed forgetting in lab and online samples and variations in stimuli

Melisa Akan^{a,*} and Lili Sahakyan^{a,b,*}

^aDepartment of Psychology, University of Illinois at Urbana-Champaign, Champaign, IL, USA; ^bBeckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Champaign, IL, USA

ABSTRACT

Previous research has produced mixed findings as to whether it is possible to selectively forget a subset of items while maintaining access to the remaining items from the same episode, using a modified version of the list-method directed forgetting (LMDF) paradigm. The present study includes six attempts to obtain the selective directed forgetting (SDF) effect with the aim of exploring its underlying mechanisms. However, despite variations in the stimuli and samples, which included both lab and online participants, we failed to obtain SDF across five experiments. In one of the experiments, we observed what appeared to be an SDF effect; however, the unexpected baseline differences across the conditions make the interpretation of this result equivocal. In contrast, standard directed forgetting effect was obtained when an LMDF condition was included in the design. An evaluation of the previous literature in combination with the present study raises questions about the reliability of the SDF phenomenon.

ARTICLE HISTORY

Received 20 January 2017
Accepted 1 May 2017

KEYWORDS

Selective directed forgetting;
inhibition; context-change

Forgetting is an integral part of successful remembering. Although forgetting is often an unintentional process that we become aware of when we cannot access the information we seek, there are many instances where we intentionally try to forget information that interferes with our current goals. For example, we might want to forget our old password in order to have more efficient access to the current one.

One of the most frequently used paradigms to empirically examine intentional forgetting, list-method directed forgetting (LMDF), typically involves studying two lists of items. In between the presentation of two lists, participants are either told to keep remembering all List 1 (L1) items, or to forget them because they will not be tested on L1. Following the presentation of List 2 (L2), participants are asked to recall both L1 and L2 items. Two typical outcomes that have been observed across numerous studies are (1) reduced recall of L1 in the Forget cue condition – known as the directed forgetting (DF) costs, and (2) enhanced recall of L2 in the Remember cue condition – known as the DF benefits (for reviews, see MacLeod, 1998; Sahakyan, Delaney, Foster, & Abushanab, 2013).

It is now well established that access to L1 can be impaired by instructing participants to forget all the items that comprise the L1 episode. The current investigation was aimed at examining whether it is possible to impair access to only a subset of a list of items – or a

part of an episode – while maintaining access to the rest of the items. In other words, is it possible to engage in *selective* intentional forgetting? This question has important theoretical and practical implications. In everyday life, we commonly need to disregard or selectively forget only some of the information associated with a single event or episode. For example, a juror might have to disregard only some of the information presented by a witness or one might want to forget only the negative aspects of an otherwise pleasurable vacation while maintaining access to its positive aspects.

Selective directed forgetting (SDF) has been investigated in a few studies and produced mixed findings for as yet unknown reasons. Furthermore, there has been no rigorous investigation of its underlying mechanisms. The original goal of our investigation was to examine the underlying mechanisms of SDF by comparing it to the standard DF within the same experiment. Unfortunately, we failed to obtain SDF in six experiments across variations in stimuli and participant samples, and what started out as the investigation of the mechanisms eventually turned into the chase of the phenomenon itself.

Previous SDF research

Previous research has used two types of designs to examine SDF – two-list and three-list designs. In the two-

CONTACT Lili Sahakyan  lsahaky@illinois.edu  Department of Psychology, University of Illinois at Urbana-Champaign, 603 E. Daniel St., Champaign, IL 61820, USA

*Melisa Akan and Lili Sahakyan are joint first authors.

list design, participants are asked to study a list of items, half of which differ from the other half in terms of a particular attribute. The majority of SDF studies presented sentences about two different characters – Tom and Alex – in alternating order, as a part of L1 (Aguirre, Gómez-Ariza, Bajo, Andrés, & Mazzoni, 2014; Delaney, Nghiem, & Waldum, 2009; Gómez-Ariza et al., 2013; Storm, Koppel, & Wilson, 2013). Following L1, the selective forget cue instructs participants to forget a subset of these items (e.g., “forget the sentences about Tom”). Thus, in the two-list design, the to-be-forgotten (TBF) and the to-be-remembered (TBR) items are intermixed within the same temporal episode, and the critical question is whether participants can selectively forget some of the items from that episode while maintaining access to the remaining items.

In contrast to the two-list design, in the three-list design, the TBF and TBR items constitute separate temporal episodes (Kliegl, Pastötter, & Bäuml, 2013; Sahakyan, 2004). Specifically, participants are presented with three lists of items, and after studying the first two lists, they are instructed to forget the middle list, but to keep maintaining access to the first list (and the third list). The critical question is whether participants can forget the middle list without forgetting also the first list, or whether such selective forgetting is impossible, and the first list becomes forgotten along with the middle list. Both types of designs have produced inconsistent findings. Our investigation utilised mainly the two-list design, but for completeness we will discuss the inconsistencies in the literature across both design types.

Two-list designs

Delaney et al. (2009) was the first to investigate SDF using a two-list design. After studying a number of sentences about two characters named Tom and Alex, they had participants study a second list of sentences about a new character. After L1, half of the participants were asked to either forget all the sentences about Tom while keeping Alex in mind, whereas the remaining half was told to keep remembering all L1 sentences. After L2 study, all participants were asked to recall both lists, including the items they were earlier instructed to forget. An additional variable included in their design was the relatedness of the sentences. Half of the participants studied sentences that were unrelated about a given character (e.g., Tom sentences were unrelated to each other, and Alex sentences were unrelated to each other), whereas the remaining participants studied related sentences (e.g., all Tom sentences were related to each other and described Tom’s character, while all Alex sentences described Alex’s character). Delaney et al. (2009) found that instructions to forget Tom (i.e., SDF cue) resulted in a memory impairment for Tom sentences, while leaving recall of Alex sentences intact, demonstrating SDF. However, SDF was observed only in unrelated sentences condition; in contrast, when the sentences were thematically related, no SDF effects

were obtained (there was neither selective forgetting of Tom sentences, nor overall forgetting of sentences regarding both characters).

Unlike Delaney et al. (2009), Storm et al. (2013) were unable to find SDF using the original unrelated sentences and the same procedures, despite recruiting substantially larger samples. L1 recall in the SDF condition did not differ at all from the Remember condition, implying that participants either did not attempt to engage in SDF, or attempted but were unsuccessful at it. In a follow-up experiment, Storm and colleagues considered the possibility that Delaney et al.’s participants might have selectively rehearsed Alex sentences upon receiving the SDF cue, enabling them to forget Tom sentences. Thus, following L1 presentation, Storm et al. explicitly instructed participants to rehearse only the L1 TBR items (i.e., Alex sentences). The explicit rehearsal instructions led to enhanced memory for Alex sentences in the SDF condition; however, there was still no evidence for selective forgetting of Tom sentences. In addition, in one of their experiments, they also presented the TBF and TBR items in L1 in a blocked fashion, rather than in alternating order to facilitate the selective rehearsal of only TBR items. However, this blocking factor appeared to have no effect on SDF. Finally, Storm et al. tested whether any DF would be obtained with those stimuli using a standard LMDF task, where all L1 sentences are instructed to-be-forgotten. The results revealed an overall reduction in L1 recall (i.e., both Tom and Alex sentences were successfully forgotten) along with enhancement in L2 recall. To summarise, Storm et al. (2013) failed to replicate the SDF effect using larger samples, and they also failed to obtain SDF upon explicitly instructing participants to selectively rehearse L1 TBR items. Finally, they obtained overall DF using the same set of stimuli, suggesting that it was not a generalised failure to obtain DF, but more specifically, a failure to obtain SDF. Considering the null SDF results of Storm et al. (2013), one might speculate that Delaney et al.’s (2009) positive findings in their unrelated condition were due to a Type 1 error, while their thematic condition reveals the true state of affairs (as opposed to the effect of integration on SDF).

However, this issue is complicated by the findings of Gómez-Ariza et al. (2013) and Aguirre et al. (2014), who tested special populations using the same sentence stimuli and procedure, and obtained SDF in their control samples. Both of these studies were motivated by the assumption that selective forgetting involved the operation of an inhibitory mechanism targeting only the TBF items (akin to inhibition involved in retrieval-induced forgetting (RIF) or think–no-think (TNT) research). According to the authors, the selective nature of this inhibitory operation might require greater inhibitory capacity compared to a standard LMDF task, and therefore, they examined populations known to have reduced executive control. Aguirre et al. (2014) compared older adults and younger adults, whereas Gómez-Ariza et al. (2013) compared

adolescents with social anxiety disorder (SAD) and age-matched nonclinical controls.

Aguirre et al.'s younger adults showed SDF (i.e., they forgot Tom, without forgetting Alex), while older adults did not show SDF – their memory for Tom and Alex sentences remained comparable to the Remember group. In a subsequent experiment, Aguirre et al. (2014) tested only older adults and again failed to obtain SDF with older adults. Finally, their last experiment included both age groups and used the standard LMDF task, which instructed participants to forget both Tom and Alex sentences. Both young and older adults showed the standard LMDF. The authors thus argued the older adult's insensitivity to SDF was specific to SDF, rather than reflecting a generalised deficit in intentional forgetting per se. Although their findings are consistent with their interpretations, it is noteworthy to point out that the SDF effect was demonstrated only once in Aguirre et al. (2014) study, using a modest sample size of younger adults ($N = 48$ younger adults, randomly assigned to SDF and Remember conditions).

Conceptually similar findings were reported also by Gómez-Ariza et al. (2013), who showed that while their SAD sample was unable to selectively forget Tom sentences, their non-anxious control sample demonstrated SDF. However, Gómez-Ariza et al. (2013) used idiosyncratic exclusion criteria that differed from all other SDF studies in that participants with lower memory performance were assumed to have not attempted to learn the material and were eliminated from the analyses (the specific criterion for low memory performance was not reported). One concern over this exclusion criterion is that whereas a low memory in the Remember condition might reflect "not engaging in the task", the same low performance in the SDF condition could mean the exact opposite – namely, successfully engaging in forgetting. Thus, eliminating low performers from both conditions using low-performance criterion could obscure the findings. Furthermore, it appears that more participants were excluded from the Remember than the SDF condition because when SAD and matched control groups were compared across the Remember condition, the reported degrees of freedom were substantially smaller than when those two samples were compared across the SDF condition. Because the results were not reported without exclusions (and it was unspecified how many participants were originally assigned to the Remember and SDF groups within SAD and matched control conditions), it remains possible that the specific exclusion criteria might have produced the SDF effect in their study.

Finally, in one of their experiments, Kliegl et al. (2013) obtained SDF effects using word stimuli and manipulating the gender of the speaker's voice who presented the words to differentiate TBF and TBR items. However, this study differed from prior SDF work in a critical detail – namely, the final memory test required the participants to retrieve TBF and TBR subsets separately, as opposed to retrieving all L1 items together.

Three-list designs

The three-list studies have also revealed inconsistent findings regarding SDF. Sahakyan (2004) pioneered research on SDF, where participants studied three lists, and the critical SDF condition involved instructing participants to forget L2 but to remember L1 and L3, following L2 presentation. Compared to the Remember condition, instruction to selectively forget L2 led to the forgetting of both L1 and L2. Thus, participants did not appear to be able to selectively forget only the middle list; rather, the selective forget instruction led to overall forgetting of both the first and the middle list. Sahakyan (2004) replicated those findings in a subsequent experiment, using categorised lists. In contrast, Kliegl et al. (2013) presented evidence for SDF costs using a similar three-list design, but substantially shorter lists. Specifically, their participants were able to forget the middle list without forgetting the first list. The obvious difference between the two studies was the list length, which Kliegl et al. speculated could have made it easier to discriminate TBR and TBF lists, facilitating selective forgetting. No studies have yet tested the list length hypothesis to discern the reasons for inconsistencies.

Summary of SDF studies

The SDF effect remains elusive across both types of paradigms. Of all the reviewed SDF studies, Storm et al. (2013) used the largest sample sizes and presumably had the greatest power to detect potential SDF effects. However, they failed to obtain SDF in all of their experiments. On the other hand, two studies examining special populations had smaller samples, and they observed SDF in their control samples. Across all two-list SDF studies, the aggregated sample size that showed SDF ($N = 336$) was almost equal to the aggregate sample size that did not show SDF ($N = 338$), see Table 1. Note that the SAD sample in Gómez-Ariza et al. (2013) and the older adult sample in Aguirre et al. (2014), who were not expected to show SDF effects were not included in the aggregated sample sizes. Given that the stimuli and procedures were exactly the same across these studies except the Kliegl

Table 1. The summary of previous SDF studies across two-list and three-list designs.

	Sample sizes	Selective DF	Total
<i>Two-list designs</i>			
Storm et al. (2013)	Exp 1 $n = 64$ Exp 2 $n = 128$ Exp 3 $n = 146$	SDF effect not obtained	$N = 338$
Delaney et al. (2009) Gómez-Ariza et al. (2013) Aguirre et al. (2014) Kliegl et al. (2013)	$n = 64$ $n = 104$ $n = 48$ Exp 3 $n = 120$	SDF effect obtained	$N = 336$
<i>Three-list designs</i>			
Sahakyan (2004)	Exp 1 & Exp 2	SDF effect not obtained	$N = 108$
Kliegl et al. (2013)	Exp 1, Exp 2, Exp 3	SDF effect obtained	$N = 376$

et al. (2013) study, inconsistent outcomes are puzzling and require further investigation.

The current studies implemented the SDF design along with the standard LMDF within the same experiment (Experiments 1–4). The stimuli in the first four studies involved words presented in different font colours. The final two experiments included only the SDF and Remember groups (the standard LMDF was omitted from the design). The stimuli in Experiment 5 were unrelated words that formed ad hoc categories – they consisted of things that are naturally red or green. We reasoned that the ad hoc categories might improve the identification of TBF set upon receiving the selective cue compared to arbitrary word–colour pairings used in Experiments 1–4. Finally, in Experiment 6, we used the unrelated sentences used in published research.

Experiment 1

Method

Participants

One hundred and seventeen undergraduates from the University of Illinois at Urbana-Champaign (UIUC) were individually tested in the lab in exchange for course credit. Upon completion of the experiment, participants who indicated not believing or conforming to the forget instruction in their post-experimental questionnaire were dropped (6 from SDF and 15 from DF conditions), and additional participants were tested to replace them. In addition, two participants were excluded because they did not follow the encoding instructions, and one participant was excluded due to experimenter error. The final sample consisted of 114 participants (38 in the SDF, 36 in the DF, and 40 in the R group).

Materials

Thirty-two unrelated medium-frequency English nouns were drawn from the Toronto Word Pool. The words were divided into two lists of 16 each. The presentation order of words within each list, as well as their assignment to each list and colour was random.

Procedure

Participants were informed that they would be presented with words in different colours to learn for a later memory test. The encoding instructions did not specify to intentionally learn the colour of the items. Items were presented at a rate of 4 s followed by a 1-s inter-stimulus interval (ISI). Half of the L1 items was presented in red and the remaining half was presented in green colour, and all L2 items were presented in green. The word–colour matching was arbitrary as the colours were randomly assigned to each word. The presentation order of green and red words in L1 was randomised with the constraint that no more than two words of the same colour could follow each other. Following L1 presentation,

participants received one of the three types of instruction. They were either told to forget only the items presented in red font colour (SDF condition), or to forget all L1 items (DF condition), or to continue remembering all L1 items (R condition). Participants receiving either type of forget instruction were told that they would not be tested on the items they were asked to forget, and it was important to try to forget them as this would help them better remember the remaining items. The cue was followed by L2 presentation, which all participants were instructed to remember. Following a 30-s distractor task involving simple arithmetic problems, participants were given 90 s to type as many words as they could recall without worrying about the colours, including the items they were earlier told to forget. The recall order of L1 and L2 was counterbalanced, with half of the participants recalling L1 first, followed by L2, and the remaining participants recalling the lists in reverse order. After the memory test, participants completed a brief questionnaire that inquired about their encoding and forgetting strategies.

Results and discussion

Responses were counted as correct if they were recalled with the correct list.¹ All SDF results were analysed with a 2×2 mixed analysis of variance (ANOVA) on correct L1 recall with Cue (SDF vs. R) as the between-subjects, and Item-type (TBF = red items vs. TBR = green items) as the within-subjects variable. Typically, an SDF effect is captured by an Item-type $\times$ Cue interaction. This interaction could result either from the selective impairment of TBF items, accompanied by an enhancement of TBR items in the SDF group relative to the R group, or from the greater impairment of TBF items compared to the impairment of L1 TBR items – that is, when both types of L1 items are forgotten, but TBF items are forgotten to a greater extent. The SDF results are presented in Table 2.

There was no significant interaction between Item-type and Cue, revealing no overall SDF effects, $F(1, 76) = 1.10$, $MSE = .03$, $p = .30$, $\eta^2 = .01$. The recall of TBF (red) items was not reduced in the SDF compared to the R group, $t(76) = 0.13$, $p = .90$, $d = 0.03$. There was also no recall difference in L1 TBR (green) recall in the SDF relative to the R condition, $t(76) = 1.02$, $p = .31$, $d = 0.23$. There was neither a main effect of Cue [$F(1, 76) = 0.28$, $MSE = .018$, $p = .60$, $\eta^2 = .004$] nor a main effect of Item-type, [$F(1, 76) = 0.20$, $p = .65$, $\eta^2 = .003$]. In short, receiving a selective forget cue did not result in any changes in memory performance compared to receiving a remember cue. There were neither SDF effects nor overall forgetting of L1. The two groups also did not differ in L2 recall, $t(76) = 0.13$, $p = .90$.

The standard DF results were analysed with a mixed ANOVA using Cue (DF vs. R) as a between-subjects factor and List (L1 vs. L2) as a within-subjects factor. A standard DF effect is captured by a List $\times$ Cue interaction, which manifests itself as impaired L1 recall and enhanced L2

Table 2. Mean recall proportions of L1 TBF and L1 TBR items in SDF and R conditions across experiments.

Experiment	List 1 TBF		List 1 TBR		List 2	
	SDF	R	SDF	R	SDF	R
1	.31 (.20)	.31 (.23)	.27 (.20)	.32 (.22)	.31 (.21)	.21 (.23)
2	.20 (.25)	.29 (.21)	.25 (.27)	.29 (.22)	.28 (.25)	.25 (.24)
3	.16 (.17)	.20 (.17)	.24 (.21)	.23 (.19)	.19 (.19)	.24 (.20)
4	.22 (.20)	.35 (.29)	.24 (.22)	.28 (.29)	.26 (.26)	.27 (.28)
5	.42 (.18)	.43 (.10)	.45 (.16)	.43 (.10)	.47 (.17)	.48 (.16)
6	.28 (.21)	.31 (.26)	.28 (.21)	.30 (.17)	.34 (.21)	.33 (.22)

Note: Standard deviations are in parentheses.

recall in the DF group compared to the R group. The standard DF results are presented in Table 3.

In contrast to the null SDF findings, reliable DF effects were obtained. There was a significant List $\times$ Cue interaction, $F(1, 74) = 5.02$, $MSE = .03$, $p = .03$, $\eta^2 = .06$. L1 recall was impaired in the DF compared to the R condition, $t(74) = 2.45$, $p = .02$, $d = 0.57$. However, DF costs were not accompanied by an enhancement of L2 recall in the DF relative to the R condition, $t(74) = 0.50$, $p = .62$, revealing no DF benefits. Overall, L2 recall ($M = .32$, $SD = .22$) was numerically higher than L1 recall ($M = .27$, $SD = .18$), $F(1, 74) = 3.82$, $p = .054$, $\eta^2 = .05$. Finally, Cue did not have a significant main effect, $F(1, 74) = 0.96$, $MSE = .05$, $p = .33$, $\eta^2 = .01$.

To summarise, whereas the stimuli presented in different colours were susceptible to overall DF, they were not susceptible to selective DF. It is possible that participants in the SDF group were not able to successfully differentiate red items from the green items, making it impossible to selectively target TBF and/or TBR set of items. Hence, in Experiments 2–4, items were presented as colour–word pairs, and participants were explicitly instructed to encode colour of the words.

Experiments 2–4

Experiments 2–4 had the same design and procedure as Experiment 1 with the exception of a few changes. First, these studies were run online, using Amazon's Mechanical Turk (MTurk). Second, participants were forewarned to expect a forget cue, in order to minimise suspicion and non-compliance triggered by the unexpected forget instruction in Experiment 1. Before the study phase, all participants were informed that the study was about intentional forgetting of unwanted information, and that after studying a list of words, they would either be asked to forget all or some of the items they have studied or alternatively, they would be asked to remember all of them.

Table 3. Mean recall proportions of L1 and L2 items in DF and R conditions.

Experiment	List 1		List 2	
	DF	R	DF	R
1	.22 (.15)	.31 (.19)	.33 (.21)	.31 (.23)
2	.23 (.17)	.29 (.19)	.44 (.32)	.24 (.24)
3	.14 (.19)	.22 (.15)	.30 (.24)	.24 (.20)
4	.19 (.19)	.31 (.27)	.41 (.29)	.27 (.28)

Note: Standard deviations are in parentheses.

They were further told that the programme randomly determined which instruction they would receive and thus were asked to put equal effort into learning all the items.² Third, single item presentation was changed to colour–word pair presentation format, with the first word of the pair referring to the colour and the second word denoting the word (e.g., GREEN – DIAMOND, RED – HAMMER).³ The assignment of font colours to words was random as in Experiment 1, and the pairs were always printed in the assigned font colours (e.g., GREEN – DIAMOND word pair was presented in green font whenever DIAMOND was assigned to the green colour condition). Finally, participants were explicitly instructed to encode the colour of the words along with the words.

The only difference across Experiments 2–4 was the colour of L2 items. The reason for having different L2 colours across experiments was to examine whether similarity between L1 and L2 might moderate the SDF effects by influencing the ease of differentiating TBF items from the remaining items in memory. Previous DF literature showed that the similarity of TBF and TBR items is an important factor in whether DF is obtained, such that presenting categorically dissimilar items, or items that differ in presentation format (i.e., words vs. line drawings) across the two lists eliminates DF (Hupbach & Sahakyan, 2014; Wilson, Kipp, & Chapman, 2003). Thus, we used the colour manipulation with the purpose of influencing the likelihood that TBF items could be differentiated not just from the L1 TBR items, but also L2 items. In Experiment 2, all L2 items were presented in green font, which maximally differentiates the red TBF items from all remaining items in memory, because they are all green. In Experiment 3, half of the L2 items were presented in red, and the other half in green, which reduces the differentiation of TBF items since they share similarity with some L2 items. Finally, in Experiment 4, all L2 items were in blue font, which makes TBF items differentiated both from L1 TBR items and L2 items.

Experiment 2: green L2

Method

Participants. Participants were recruited from MTurk. Participants had to be located in the US, and to have a hit approval rate of 95% to participate. A total of 137

participants took part in this experiment, and they were paid 15¢ for their participation. Data from 3 participants who reported having cheated and 6 participants who were above the age of 60, or who did not report their age were eliminated. Older participants were excluded from analyses because DF is reduced in ageing (for a meta-analysis, see Titz & Verhaegen, 2010). The final sample consisted of 128 participants (40 in the SDF, 48 in the DF, and 40 in the R group).

Procedure. All L2 items were presented in green font as in E1. Everything was identical to E1 except for the above-mentioned changes.

Results

SDF effects were analysed by an Item-type (TBF = red items vs. TBR = green items) by Cue (SDF vs. R) mixed ANOVA on L1 recall. There was no significant interaction, $F(1, 78) = 1.04$, $MSE = .03$, $p = .31$, $\eta^2 = .01$. The recall of TBF (red) items was not significantly impaired in the SDF compared to the R group, $t(78) = 1.81$, $p = .07$, $d = 0.54$. There was also no recall difference in L1 TBR items across the SDF and R conditions, $t(78) = 0.68$, $p = .50$, $d = 0.18$. There was neither a main effect of Item-type, $F(1, 78) = 1.04$, $MSE = .30$, $p = .31$, $\eta^2 = .01$, nor a main effect of Cue, $F(1, 78) = 2.05$, $MSE = .08$, $p = .16$, $\eta^2 = .03$. There was also no difference in L2 recall across the two groups, $t(78) = 0.60$, $p = .55$, $d = 0.13$. Taken together, these results suggest that there was no significant SDF in the current experiment.

For the overall DF analyses, a mixed ANOVA with List (L1 vs. L2) and Cue (DF vs. R) revealed a significant interaction, indicating a reliable overall DF effect, $F(1, 86) = 19.62$, $MSE = .04$, $p < .0001$, $\eta^2 = .19$. L1 recall was lower in the DF compared to the R group, $t(86) = 1.65$, $p = .10$, $d = 0.35$, even though this difference did not reach significance. L2 recall was higher in the DF compared to the R group, $t(86) = 3.21$, $p = .002$, $d = 0.70$. List had a significant main effect, L2 recall ($M = .35$, $SD = .30$) being higher than L1 recall ($M = .26$, $SD = .18$), $F(1, 86) = 8.42$, $p = .005$, $\eta^2 = .09$. There was no main effect of Cue, $F(1, 86) = 2.53$, $MSE = .08$, $p = .11$, $\eta^2 = .03$.

Experiment 3: green and red L2

Method

Participants. One hundred and thirty-six MTurk participants were recruited. Participation criteria and payment were the same as in Experiment 2. Five participants who reported having cheated or copied down the items, together with 5 participants who were above the age of 60 were eliminated from the analyses. The final sample consisted of 126 participants (42 in the SDF, 41 in the DF, and 44 in the R group).

Procedure. The stimuli and procedure were exactly the same as in Experiment 2 except that L2 words were presented in red and green colours. The presentation order of red and green words within each list was randomised using the same constraining criteria as before.

Results

SDF effects were analysed by an Item-type (TBF = red items vs. TBR = green items) by Cue (SDF vs. R) mixed ANOVA on L1 recall. There was no significant interaction, $F(1, 84) = 0.96$, $MSE = .02$, $p = .33$, $\eta^2 = .01$, indicating no SDF effects. The recall of TBF (red) items was not reduced in the SDF compared to the R group, $t(84) = 0.94$, $p = .35$, $d = 0.20$, and there was no recall difference between L1 TBR (green) items in the SDF compared to the R group, $t(84) = 0.19$, $p = .85$, $d = 0.04$. Item-type had a significant main effect; with higher recall of L1 green items ($M = .24$, $SD = .20$) relative to the L1 red items ($M = .18$, $SD = .17$), $F(1, 84) = 6.39$, $MSE = .02$, $p = .013$, $\eta^2 = .07$. There was no main effect of Cue, $F(84) = .16$, $MSE = .05$, $p = .69$, $\eta^2 = .002$. There was also no difference in L2 recall across the two groups, $t(84) = 1.18$, $p = .24$, $d = 0.25$.

Contrary to the null SDF effects, reliable overall DF effects were obtained revealed by a significant List $\times$ Cue interaction, $F(1, 83) = 10.82$, $MSE = .02$, $p = .001$, $\eta^2 = .11$. L1 recall was impaired in the DF compared to the R group, $t(83) = 2.00$, $p = .049$, $d = 0.47$, indicating DF costs. L2 recall was numerically in the expected direction, being higher in the DF compared to the R group; however, this difference was not statistically significant, $t(83) = 1.41$, $p = .16$, $d = 0.30$. There was a significant main effect of List, $F(1, 83) = 18.31$, $p < .001$, $\eta^2 = .18$, with L2 recall being higher ($M = .27$, $SD = .22$) than L1 recall ($M = .18$, $SD = .17$). The main effect of Cue was not significant, $F(1, 83) = .008$, $MSE = .06$, $p = .93$. When L1 and L2 items were presented in the same colours as in previous DF studies, a reliable DF effect was observed, but there was no evidence of SDF.

Experiment 4: blue L2

Method

Participants. One hundred and forty-nine MTurk participants took part in this experiment. Seven participants who reported having cheated or copying down the words were eliminated together with 18 participants who were either above the age of 60 or who did not report their age. The final sample included 124 participants (39 in the SDF, 44 in the DF, and 41 in the R group).

Procedure. All L2 words were presented in blue font.

Results and discussion

The interaction between Cue and Item-type was not statistically significant, $F(1, 78) = 2.75$, $MSE = .03$, $p = .10$, $\eta^2 = .03$. TBF recall was lower in the SDF compared to the R group, $t(78) = 2.30$, $p = .02$, $d = 0.52$, whereas no difference was observed in L1 TBR recall, $t(78) = 0.69$, $p = .49$. There was neither a main effect of Item-type, $F(1, 78) = 1.13$, $p = .29$, $\eta^2 = .01$, nor Cue, $F(1, 78) = 2.80$, $MSE = .10$, $p = .10$, $\eta^2 = .03$. Finally, L2 recall did not differ across the two groups, $t(78) = 0.06$, $p = .95$.

Reliable DF effects were detected as revealed by the significant List $\times$ Cue interaction, $F(1, 83) = 26.54$, $MSE = .03$, p

$< .001$, $\eta^2 = .24$. L1 recall was lower in the DF compared to the R group, $t(83) = 2.36$, $p = .02$, $d = 0.51$, whereas L2 recall was higher in the DF group compared to the R group, $t(83) = 2.30$, $p = .02$, $d = 0.49$. L2 was overall better recalled ($M = .34$, $SD = .29$) than L1 ($M = .25$, $SD = .24$), as revealed by a main effect of List, $F(1, 83) = 11.23$, $p = .001$, $\eta^2 = .12$. There was no main effect of Cue, $F(1, 83) = 0.05$, $MSE = .11$, $p = .82$.

This was the first experiment where significant SDF costs were detected. Furthermore, despite having dissimilar lists, DF effects were obtained. This might be indicative of colour not acting as a particularly useful cue in differentiating the lists, which typically eliminates the standard DF effect. Therefore, it is unclear whether the low similarity across lists facilitated obtaining SDF or whether the apparent SDF costs in the current experiment may have been a Type 1 error.

In this experiment, the recall rates of red and green items from L1 showed an unexpected pattern in the R condition. Specifically, red items had higher recall ($M = .35$, $SD = .29$) relative to green items ($M = .28$, $SD = .29$) by 7%, $t(40) = 2.27$, $p = .02$. Given that the assignment of words to red and green colours was random, there was no a priori reason to expect such a difference. Moreover, no such difference was observed in any of the preceding experiments. Thus, the inflated baseline for red items (the designated TBF category in the SDF condition) in the R condition may have been responsible for the detection of SDF costs.⁴

When the data were pooled across Experiments 1–4, and L1 red and green item recall was compared in the R group, there was no significant difference between the recall of two item types, $t < 1$. Comparing L1 red (TBF) recall across the SDF ($M = .22$, $SD = .21$) and the R group ($M = .28$, $SD = .23$) revealed significant SDF costs, $t(322) = 2.54$, $p = .01$, $d = 0.28$. [Comparing L1 green (TBR) items across the SDF ($M = .25$, $SD = .23$) and R groups ($M = .28$, $SD = .23$) did not reveal significant differences, $t(322) = 1.12$, $p = .26$, $d = 0.12$].

Given that the SDF impairment reached significance only when the data were pooled across multiple experiments, we also examined the data by estimating a Bayes factor, comparing the fit of the data under the null hypothesis and the alternative hypothesis. With large sample sizes, p -values might indicate significance for effects that actually provide just as much evidence for the null hypothesis as for the alternative, and they tend to overestimate evidence against the null (Masson, 2011; Wagenmakers, 2007). In such cases, Bayes factor becomes more informative (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). An estimated Bayes factor on the combined data revealed $BF_{01} = 0.35$, which suggests only anecdotal evidence in favour of the alternative over the null hypothesis (when the Bayes factor is estimated only for Experiment 4, it is $BF_{01} = 0.45$, which again provides anecdotal evidence). In sum, while the null hypothesis testing revealed significant SDF impairment, Bayesian analyses did not provide support for this finding.

Experiment 5: ad hoc categories

Given that the previous four experiments did not provide much evidence for SDF, in the current study we used items from ad hoc categories, which consisted of things that are naturally red or green. These items served as L1 items, whereas L2 consisted of items that make noise. We reasoned that ad hoc categories might make it easier for participants to classify the items into their designated sets compared to the colour manipulation used in the previous experiments. It could be that the random assignment of words to colours made the identification of the TBF set difficult, and when participants received the selective forget cue, in order to identify which items they should forget, they had to retrieve those items first, thereby strengthening their memory, and reducing the likelihood of obtaining SDF. Using the TNT paradigm (Anderson & Green, 2001), Detre, Natarajan, Gershman, and Norman (2013) showed that moderate levels of activation of no-think targets were associated with subsequent forgetting, whereas strong activation of no-think targets produced the opposite effect (enhanced performance on the final test). Paralleling Detre et al.'s results, if the cue to selectively forget moderately activated the TBF items, then we might expect to obtain SDF, whereas if the cue led to the retrieval of these items, we would expect the opposite.

In addition to implementing ad hoc categories, this experiment aimed to more closely follow the procedures of the previous SDF studies that typically employ sentences, including the uneven length of L1 and L2, the alternating order of TBF and TBR items, the inclusion of a brief distractor activity after each list. Upon seeking the input of researchers from other labs, who obtained the SDF effect using sentence stimuli, we slightly modified the selective forget instruction by making it more emphatic and including a justification for the forgetting requirement. See Appendix for the exact wording of the instructions.

In addition, we examined the impact of forewarning on SDF, given that Experiments 2–4 involved warning participants to expect a forget cue. Although warning manipulations are gaining popularity in standard DF research, and have been shown to have no impact on the DF outcomes (e.g., Foster & Sahakyan, 2011), there has been no systematic investigation of this variable in SDF literature. Finally, because the central goal of the investigation was the SDF phenomenon, to maximise the number of available participants in the relevant conditions, we dropped the standard DF group from the design.

Method

Participants

The participants were 120 UIUC undergraduates, who participated for course credit. They were tested individually in the lab. Participants were randomly assigned to one of the four groups, which were formed by factorial combination of Surprise/Warning, and SDF/R conditions. There were

30 participants in each of the conditions. Four participants in the Surprise-SDF group were replaced as they had indicated not believing the forget cue and expecting a memory test in the post-experimental questionnaire. Within each of the forget conditions, 15 participants were told to forget the red items, and 15 participants were told to forget the green items.

Stimuli

The items on L1 consisted of 16 words, with half referring to “things that are green” (i.e., dollar and grasshopper), and half referring to “things that are red” (i.e., blood and ketchup). All L2 items consisted of 14 words referring to “things that make noise” (e.g., vacuum and engine). Stimuli were selected from the updated Battig and Montague norms (Van Overschelde, Rawson, & Dunlosky, 2004). For L1 items, we selected low frequency responses (i.e., production probability of each selected word was under .15 in the norms). For L2 items, the production probabilities were under .30 in the norms.

Procedure

All participants (both warning and surprise conditions) were told that they should memorise a list of words, representing different things in the world, some of which are naturally green, and others are red. They were encouraged to think about the meaning of the words and their colour because it would be important later on. In the warning condition, *prior to receiving* the L1 encoding instruction, participants were forewarned about the possibility of receiving the forget cue, using the instructions described in Appendix. The instructions made it clear to participants that it was impossible to know in advance whether they would get the forget or the remember instruction, and therefore they were encouraged to put equal effort into learning all the words. In contrast, in the surprise condition, participants received the forget cue as a surprise, after L1 encoding, and they were not warned about the possibility of receiving the forget cue prior to encoding L1 words. Items were presented in black font at a rate of 4 s followed by a 1-s ISI. The red and green ad hoc exemplars were alternated during presentation, with a randomly selected exemplar from each category. After L1 presentation, half of the participants received the forget instruction. In the warning condition, the forget instruction stated that they only need to keep in mind the red objects (or green objects), and that they should forget the green objects (or red objects) because they would not be tested on them. In the surprise condition, the forget instruction was more emphatic to better motivate it and emphasise the need to forget (see Appendix). Participants receiving the remember instruction after L1 were told to keep all items in mind. Afterwards all participants engaged in a brief distractor activity for 60 s, which involved solving arithmetic problems presented one at a time on the computer screen. This was followed by L2 presentation,

and participants were told that all items on L2 refer to things that make noise. Following L2 presentation, participants solved arithmetic problems for 60 s, and then took the memory test. Recall order was fixed, with L1 being tested before L2, with 90 s allocated for the recall of each list. The test instructions specified to recall any word from L1 they could remember, including the words they were previously told to forget. Afterwards, L2 memory was tested. Following the test, participants completed a post-experimental questionnaire on their encoding and forgetting strategies.

Results and discussion

First, we compared the recall of two types of L1 items in the R group, using Item colour (red vs. green) $\times$ Condition (surprise vs. warning) mixed ANOVA to determine if there were differences in the baseline recall of the two ad hoc categories. Neither the main of item colour [$F(1, 118) = 1.10$, $MSE = .01$, $p = .30$, $\eta^2 = .01$] nor the main effect of condition [$F(1, 188) = 0.71$, $MSE = .03$, $p = .40$, $\eta^2 = .01$] nor the interaction [$F(1, 118) = 0.15$, $p = .69$, $\eta^2 = .01$] was significant. The mean recall in the surprise group was $M_{red} = .41$ ($SD = .16$), $M_{green} = .43$ ($SD = .17$), and in the warning group was $M_{red} = .44$ ($SD = .13$), $M_{green} = .45$ ($SD = .13$). Therefore, for the subsequent analyses, we collapsed across the red and green item recall in the R condition, and used the average item recall as the basis of comparison to the SDFgroup.

To examine SDF, we analysed L1 TBF and L1 TBR recall separately using Cue (SDF vs. R) $\times$ Condition (Surprise vs. Warning) factorial ANOVA. The Cue $\times$ Warning Condition ANOVA on L1 TBF recall showed neither a main effect of cue [$F(1, 116) = 0.29$, $MSE = .02$, $p = .59$, $\eta^2 = .003$] nor a main effect of warning condition [$F(1, 116) = 0.10$, $p = .76$, $\eta^2 = .001$]. The interaction was also not significant, $F(1, 116) = 0.38$, $p = .54$, $\eta^2 = .003$. This means that there was neither evidence of SDF, nor any effect of surprise/warning manipulation. Combined across the warning condition variable, SDF group ($M = .42$, $SD = .18$) did not show impaired recall of the designated L1 TBF items compared to the R group ($M = .43$, $SD = .11$), $t(118) = 0.55$, $p = .59$.

The same analyses on L1 TBR recall yielded similar outcomes. The main effect of condition was not significant, $F(1, 116) = 1.37$, $MSE = .02$, $p = .24$, $\eta^2 = .01$, and neither was the main effect of cue [$F(1, 116) = 0.85$, $p = .36$, $\eta^2 = .007$], nor the interaction, $F(1, 116) = 0.03$, $p = .87$, $\eta^2 = .000$. Collapsing across the warning condition variable, SDF group ($M = .45$, $SD = .16$) did not recall significantly more L1 TBR items than the R group, $t(118) = 0.92$, $p = .36$.

For completeness, we also assessed L2 recall using Cue and Condition as variables. Neither the main effect of Cue [$F(1, 116) = 0.06$, $p = .81$], nor the main effect of condition [$F(1, 116) = 1.43$, $p = .23$], nor the interaction were significant [$F(1, 116) = 1.84$, $p = .18$]. Collectively, these results suggest that surprise/warning manipulation had no effect

on the findings, and we did not find any evidence of SDF using ad hoc categories.

Experiment 6: sentences

In this experiment, we chose sentences as stimuli and closely followed the procedures of prior work with the hope of obtaining the SDF effect. It could be that our manipulations of font colour or ad hoc categories were not strong enough, and in our final attempt, we used the published sentences from previous research, and employed a procedure that was more typical of prior research. Namely, each list was followed by a distractor task, the TBF and TBR items were presented in alternating order, and the forget cue was presented as a surprise.

Method

Participants

There were 72 UIUC undergraduates, who participated for course credit. They were tested individually in the lab. Participants were randomly assigned to SDF or R groups, with 36 participants in each condition. Three participants in the SDF condition were replaced as they had indicated in the post-experimental questionnaire not believing the forget cue and expecting to be tested on TBF items.

Stimuli

The stimuli were taken from the published Delaney et al. (2009) study, from their random/unrelated condition. The first list contained 8 sentences about a character named Tom and 8 sentences about a character named Alex, and the second list contained 14 sentences about a character named Joe.

Procedure

Prior to L1 presentation, all participants were told they would be studying sentences referring to two characters named Tom and Alex, and that they should attempt to memorise the sentences for a later memory test. Then Tom and Alex sentences were presented in alternating order at a rate of 8 s, with 1 s ISI. Tom and Alex characters were assigned equally often to each sentence. After the last sentence, participants receiving the selective forget instruction were told to forget Tom sentences and remember only Alex sentences because they would not be tested on the Tom sentences (see Appendix for instructions). After completing a 60 s delay filled with arithmetic problems, participants were told that they were going to study a second list composed of sentences about a character named Joe. Upon L2 presentation, and another 60-s delay filled with arithmetic, participants were asked to recall L1 (including the sentences they were previously told to forget), followed by L2, with 120 s allocated for the recall of each list. Finally, participants completed a post-experimental questionnaire inquiring about their encoding and forgetting strategies.

Results and discussion

Recall was scored following the same criteria as in Delaney et al. (2009). Sentences were counted correct if the gist of the sentence was preserved, and substitutions did not change its meaning (e.g., brushed vs. combed his hair). Substitutions that altered the meaning of the sentence were counted as incorrect. In addition, the character–action combination had to be preserved for the item to be counted as correct (e.g., Tom sentences assigned to the Alex character, or vice versa, were counted as incorrect) consistent with previous research. Overall, the character confusions were quite rare (the average number of sentences that were recalled with character confusions across all participants were $M = .58$, $SD = 1.16$). We examined the number of character confusions as a function of Character (Tom vs. Alex) and Cue (SDF vs. R) using mixed ANOVA. The results revealed neither main effects (both $F_s < 1$), nor an interaction, $F(1, 70) = 1.12$, $p = .29$.

L1 recall was analysed with Character (Tom vs. Alex) $\times$ Cue (SDF vs. R) mixed ANOVA. Neither the main effects of character [$F(1, 70) = .000$, $MSE = .02$, $p = 1.00$], nor cue [$F(1, 70) = .23$, $MSE = .07$, $p = .64$, $\eta^2 = .003$], nor interaction were significant [$F(1, 70) = .02$, $p = .89$, $\eta^2 = .000$]. Both characters were recalled at the same rate in the R group, and in the SDF group. These findings do not change if character confusions are treated as part of correct recall (all $F_s < 1$). Analyses of L2 recall also revealed equivalent recall of Joe sentences between the R and the SDF groups, $t(70) = 0.04$, $p = .97$. In sum, there was no evidence of SDF in this experiment.

Given that Experiments 5 and 6 were procedurally similar to each other, we combined the data from these two experiments and performed Bayesian analyses. Comparing the recall of L1 TBF items across the SDF and R groups revealed an estimated Bayes factor of $BF_{01} = 5.23$, which means that the results are 5.23 times more likely to occur under the null than the alternative. The same analyses on the recall of L1 TBR items revealed an estimated $BF_{01} = 6.10$, again suggesting that the results are 6.10 more likely to be observed under the null hypothesis than the alternative. These results provide additional evidence for the lack of an SDF effect.

General discussion

We presented six experiments in which we failed to obtain SDF in both university and online participants using a variety of stimuli. In the first four experiments, the TBF and TBR sets in L1 involved unrelated words that were differentiated based on their font colour. In Experiment 5, the TBF and TBR sets included words from the ad hoc categories, involving red and green things in the real world, and manipulating whether the forget cue was delivered as a surprise or with prior warning. Finally, Experiment 6 aimed to replicate previous SDF research by using the original sentences and employing similar procedures and sample sizes as the

Delaney et al. (2009) study.⁵ Experiments 2–4 were run online, whereas the remaining studies were conducted in the lab. Across six studies, significant SDF impairment was obtained only in Experiment 4. However, the findings from Experiment 4 appear equivocal because unexpected baseline differences were detected in the Remember group and may have artificially created significant SDF impairment. In contrast to SDF, standard DF was obtained consistently across Experiments 1–4, in which a DF group was included in the design of the study.⁶ In addition to null hypothesis testing, we also performed Bayesian analyses by pooling the data from similar sets of experiments, such as Experiments 1–4, and Experiments 5 and 6. The Bayesian analyses showed that the obtained results were more likely to occur under the null hypothesis than the alternative hypothesis (Experiments 5 and 6), or at best, they showed anecdotal evidence in favour of SDF (Experiments 1–4 combined, or Experiment 4 by itself).

Our findings are not entirely surprising considering that the published literature on SDF also does not provide strong evidence for the phenomenon. As shown in Table 1, the studies that obtained the SDF effect (using a two-list design) had a total sample size of $N=336$, which was almost equal to the sample size of studies, which failed to obtain the SDF effect. Including the current set of studies, the updated number of participants that fail to produce SDF yields an $N=854$, which is more than twice larger than the total sample of participants that produced SDF. Note that the reported sample sizes do not include the three-list designs. The three-list SDF design differs from the two-list design in that the TBF items are temporally segregated from the TBR items. This aspect of the three-list design might be more conducive for engagement of forgetting mechanisms, as it requires one to reduce access to an entire set of items associated with a particular episode rather than having to select individual items within an episode. Given that the three-list design has been used by only two studies (Kliegl et al., 2013; Sahakyan, 2004), which produced mixed results, more research is needed to examine the role of temporal segregation in SDF.

As reviewed previously, prior SDF research had some features that posed difficulties for the interpretation of their findings. For example, in Delaney et al. (2009), it seems plausible that the SDF impairment emerged due to the inflated baseline in the Remember condition, in which the recall of Tom sentences exceeded the recall of Alex sentences in the Remember condition for no apparent reason (they did not report any formal analyses on this difference). A similar issue arose in our current Experiment 4, which was the only experiment that produced positive SDF results. In Gómez-Ariza et al. (2013) study, the exclusion criterion differed from the other SDF studies and appeared to result in the elimination of more participants from the Remember condition. This could have inflated the critical baseline comparison condition, producing an apparent SDF impairment, which becomes equivocal. Aguirre et al. (2014) found SDF effects using a modest sample size as it was a special

population study. Finally, Kliegl et al. (2013) was the only study to obtain significant SDF effects using word stimuli. However, their study differed from all prior SDF studies in that during the test phase, rather than having participants recall all L1 items together (as in current or prior SDF research), TBF and TBR subsets of L1 were recalled separately. The impact of such testing procedure has not been examined directly, and it is possible that it involves different search strategies, thereby creating different patterns of TBF recall. Another potential reason for the inconsistent findings might be due to the presentation format and the particular attribute Kliegl et al. used to differentiate TBF and TBR subsets – namely the words were presented auditorily by different speakers. The speaker's voice might be a more effective attribute for set differentiation relative to the font colour or the ad hoc categories used in the current study. However, even if that were the case, it still does not explain the null SDF with sentences.

The current set of findings suggest that it is not possible to selectively forget particular aspects of an episode. Even though inconsistent findings in the SDF literature is disconcerting, the current set of findings appear to be consistent with theoretical accounts of intentional forgetting. According to the context-change account of forgetting (Sahakyan & Kelley, 2002), the forget cue leads to a shift in mental context resulting in a mismatch between the pre-cue context and the context at retrieval. In response to the selective forget cue, participants may choose not to change their mental context because they have to retain access to a portion of L1 episode. Thus, changing mental context appears counterproductive for achieving the goal of retaining some items, while having to forget the remaining items. Indeed, neither in the current study nor in Storm et al. (2013) study, there was any forgetting in the SDF condition relative to the Remember condition.

Furthermore, the retrieval inhibition account of LMDF (Geiselman, Bjork, & Fishman, 1983) has been introduced because selective rehearsal by itself could not accommodate for the null SDF effects. More specifically, the seminal study by Geiselman and colleagues demonstrated that a cue to forget to-be-remembered items led to overall forgetting of both to-be-remembered and incidentally encoded items intermixed within the same L1 episode, demonstrating lack of selectivity. It is unclear why selectivity could not be obtained in the standard LMDF, but it would be obtained in a SDF design. Finally, re-exposing a subset of L1 TBF items was shown to release the remaining L1 items from LMDF during the final test (e.g., Bäuml & Sameniev, 2012; Bjork & Bjork, 1996). Therefore, having to retain Alex sentences should act as reminder of Tom sentences, precluding any impairment. If successful SDF is obtained and an inhibitory explanation is proposed, it would have to assume that inhibition acts at the level of individual items (i.e., TBF items), akin to what is observed, for example, in RIF and TNT paradigm. Unfortunately, previous studies that argued for inhibition as the underlying mechanism of SDF have not made it clear when and how

this inhibitory process operates. We note that many researchers have pointed to important differences between LMDF and RIF/TNT both in theory and empirical findings (e.g., Anderson, 2005; Bäuml, Pastötter, & Hanslmayr, 2010; Bjork, Bjork, & Anderson, 1998; Raaijmakers & Jakab, 2013; Sahakyan et al., 2013), and therefore, more conceptual work is needed to identify the parameters under which inhibition is predicted to be observed versus not observed.

Across six experiments, using diverse populations and varied stimuli, we failed to obtain the SDF effect. Although in Experiment 4, the pattern was consistent with SDF impairment, we caution against over-interpreting this finding because of the inflated baseline issues. In general, the baseline condition should receive more scrutiny in future research, because unexpected differences within that condition could give rise to a seeming SDF impairment, when in fact the results are driven by the baseline condition rather than the SDF condition.

We recognise that the absence of evidence is not the evidence of absence – it may be possible to obtain SDF using different manipulations for differentiating the TBF and TBR subsets. One could also speculate that perhaps multiple selective forget instructions may be more successful in obtaining SDF rather than a single instruction akin to multiple trials of no-think instructions that seem necessary to find effects in the TNT paradigm. However, we note that the SDF phenomenon was originally established with sentences, and even with published stimuli we did not replicate SDF, making this investigation the fourth failure to obtain SDF with sentences along with three failed attempts by Storm et al. (2013). Thus, although the current set of experiments do not enhance our theoretical understanding of SDF as we had hoped, they call into question the reliability of the phenomenon itself.

Notes

1. The pattern of results remained unaffected when items were counted as correct regardless of the list with which they were recalled.
2. See Appendix for the wording of the warning instruction.
3. The colour–word pair presentation format of the items is more similar to the stimuli used in most of the previous SDF studies (e.g., Delaney et al., 2009) which consisted of sentences about different characters engaging in various actions (e.g., Tom ate a sandwich, Alex watched TV).
4. A similar concern was also present in Delaney et al.'s (2009) study, where the recall of Tom sentences (TBF items in the SDF condition) was higher than the recall of Alex sentences (TBR items in SDF condition) in the R condition for no apparent reason, and could have given rise to the apparent SDF effect.
5. The only difference between Experiment 6 and Delaney et al. (2009) study was the length of the distractor interval between the two lists (60 vs. 90 s respectively). However, this is unlikely to have affected the results, especially given Storm et al.'s (2013) findings showing no SDF effects with or without the distractor interval.
6. Even though the standard DF was not the focus of the investigation, it is noteworthy to mention that Experiments 2–4 were conducted on Mturk and to the best of our knowledge, they were the first online experiments to demonstrate the standard LMDF.

Acknowledgements

The authors thank Teresa Bajo, Carmen Aguirre, and Carlos Gomez-Ariza for helpful discussions and input on the instructions. They are also grateful to Todd Berkenpas, Stephanie Coghill, Kristin Johnson, Divya Polson, Priyanka Rao, Mariel Reyes, Nathan Davis, and Marisa Allen for their assistance with data collection and coding.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

- Aguirre, C., Gómez-Ariza, C. J., Bajo, M. T., Andrés, P., & Mazzoni, G. (2014). Selective voluntary forgetting in young and older adults. *Psychology and Aging*, 29(1), 128–139.
- Anderson, M. C., & Green, C. (2001). Suppressing unwanted memories by executive control. *Nature*, 410(6826), 366–369.
- Anderson, M. C. (2005). The role of inhibitory control in forgetting unwanted memories: A consideration of three methods. In C. MacLeod & B. Utzl (Eds.), *Dynamic cognitive processes* (pp. 159–190). Tokyo: Springer-Verlag.
- Bjork, E. L., Bjork, R. A., & Anderson, M. C. (1998). Varieties of goal-directed forgetting. In J. M. Golding & C. M. MacLeod (Eds.), *Intentional forgetting: Interdisciplinary approaches* (pp. 103–137). Mahwah, NJ: Lawrence Erlbaum Associates.
- Bjork, E. L., & Bjork, R. A. (1996). Continuing influences of to-be-forgotten information. *Consciousness and Cognition*, 5, 176–196.
- Bäuml, K. H., Pastötter, B., & Hanslmayr, S. (2010). Binding and inhibition in episodic memory – cognitive, emotional, and neural processes. *Neuroscience & Biobehavioral Reviews*, 34(7), 1047–1054.
- Bäuml, K. H. T., & Samenieh, A. (2012). Selective memory retrieval can impair and improve retrieval of other memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(2), 488–494.
- Delaney, P. F., Nghiem, K., & Waldum, E. R. (2009). The selective directed forgetting effect: Can people forget only part of a text? *Quarterly Journal of Experimental Psychology*, 62, 1542–1550.
- Detre, G. J., Natarajan, A., Gershman, S. J., & Norman, K. A. (2013). Moderate levels of activation lead to forgetting in the think/no-think paradigm. *Neuropsychologia*, 51, 2371–2388.
- Foster, N. L., & Sahakyan, L. (2011). The role of forget-cue salience in list-method directed forgetting. *Memory*, 19(1), 110–117.
- Geiselman, R.E., Bjork, R.A., & Fishman, D. (1983). Disrupted retrieval in directed forgetting: A link with posthypnotic amnesia. *Journal of Experimental Psychology: General*, 112, 58–72.
- Gómez-Ariza, C. J., Iglesias-Parro, S., Garcia-Lopez, L. J., Díaz-Castela, M. M., Espinosa-Fernández, L., & Muela, J. A. (2013). Selective intentional forgetting in adolescents with social anxiety disorder. *Psychiatry Research*, 208(2), 151–155.
- Hupbach, A., & Sahakyan, L. (2014). Additional boundary condition for list-method directed forgetting: The effect of presentation format. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(2), 596–601.
- Kliegl, O., Pastötter, B., & Bäuml, K. H. T. (2013). List-method directed forgetting can be selective: Evidence from the 3-list and the 2-list tasks. *Memory & Cognition*, 41(3), 452–464.

- MacLeod, C. M. (1998). Directed forgetting. In J. M. Golding & C. M. MacLeod (Eds.), *Intentional forgetting: Interdisciplinary approaches* (pp. 1–57). Mahwah, NJ: Erlbaum.
- Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null hypothesis significance testing. *Behavioral Research Methods*, 43, 679–690.
- Raaijmakers, J. G. W., & Jakab, E. (2013). Rethinking inhibition theory: On the problematic status of the inhibition theory for forgetting. *Journal of Memory and Language*, 68, 98–122.
- Sahakyan, L., Delaney, P. F., Foster, N. L., & Abushanab, B. (2013). List-method directed forgetting in cognitive and clinical research: A theoretical and methodological review. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 59, pp. 131–190). New York, NY: Elsevier.
- Sahakyan, L., & Kelley, C. M. (2002). A contextual change account of the directed forgetting effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(6), 1064–1072.
- Sahakyan, L. (2004). Destructive effects of “forget” instructions. *Psychonomic Bulletin & Review*, 11(3), 555–559.
- Storm, B. C., Koppel, R. H., & Wilson, B. M. (2013). Selective cues to forget can fail to cause forgetting. *The Quarterly Journal of Experimental Psychology*, 66(1), 29–36.
- Titz, C., & Verhaegen, P. (2010). Aging and directed forgetting in episodic memory: A meta-analysis. *Psychology and Aging*, 25(2), 405–411.
- Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, 50, 289–335.
- Wagenmakers, E. M. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, 14, 779–804.
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J. & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7, 627–633.
- Wilson, S. P., Kipp, K., & Chapman, K. (2003). Limits of the retrieval-inhibition construct: List segregation in directed forgetting. *The Journal of General Psychology*, 130(4), 359–379.

Appendix

Warning instructions

In this experiment, we examine intentional forgetting of unwanted information because we hope to understand how people get rid of things they no longer wish to retain in mind. After studying a list of words, you will receive one of the two types of instructions. You may be asked either to forget some of the words that you studied, or you may be asked to remember all of them. If you receive an instruction to forget some words, it means your memory will not be tested for those words. Please make an effort to forget them. If you receive a remember instruction, it means your memory will be tested for all those words, and you should keep them in mind. There is no way to anticipate which instruction you will get, since it is randomly determined by the computer program. Therefore, please put equal effort into learning all the words.

Selective forget instruction

Experiments 1–4

You just finished studying LIST 1 words. Please try and forget the RED words because you will not be tested on them. It is important that you put them out of your mind. This will help you better remember the GREEN words.

Experiment 5

Before we move on with the experiment, we are going to make your task a bit easier because from now on, you only need to keep in mind the red objects from the first list and do not need to remember the green objects. Those were just fillers to make the task harder at the beginning, but now it gradually gets easier. You will not be tested on the green words. Please do your best to forget the green items because it will help you better remember the red items.

Experiment 6

Before we move on, you need to know that later you will be asked to recall only the sentences concerning Alex. The sentences about Tom are no longer relevant. They were fillers to make the task harder at the beginning, but now it gets easier. Thus, to improve your recall about Alex, please do your best to forget the sentences about Tom because it will help you better remember Alex during the test.