Classification: Biological Sciences – Psychological and Cognitive Sciences **and** Neuroscience

Title: Post-retrieval new learning does not reliably induce human memory updating via reconsolidation.

Short title: Investigating memory updating via reconsolidation.

Authors: Tom E. Hardwicke[a,1], Mahdi Taqi[a], & David R. Shanks[a]

Author affiliation: [a]Department of Experimental Psychology, University College London, 26 Bedford Way, London, WC1H 0AP

Corresponding author: [1]Tom E. Hardwicke, Department of Experimental Psychology, University College London, 26 Bedford Way, London, WC1H 0AP, t.hardwicke.12@ucl.ac.uk

Keywords: reconsolidation, sequence learning, memory updating, forgetting, replication

Abstract: Reconsolidation theory proposes that retrieval can destabilize an existing memory trace, opening a time-dependent window during which that trace is amenable to modification. Support for the theory is largely drawn from non-human animal studies that use invasive pharmacological or electroconvulsive interventions to disrupt a putative post-retrieval restabilization ('reconsolidation') process. In human reconsolidation studies however, it is often claimed that post-retrieval *new learning* can be employed as a means of 'updating' or 'rewriting' existing memory traces. This proposal warrants close scrutiny because the ability to modify information stored in the memory system has profound theoretical, clinical, and ethical implications. The present study aimed to replicate and extend a prominent 3-day motor-sequence learning study (Walker, Brakefield, Hobson, & Stickgold, 2003) that is widely cited as a convincing demonstration of human reconsolidation. However, in four direct replication attempts (N = 64) we did not observe the critical impairment effect that has previously been taken to indicate disruption of an existing motor memory trace. In three additional conceptual replications (N = 48), we explored the broader validity of reconsolidation-updating theory by using a declarative recall task and sequences similar to phone numbers or computer passwords. Rather than inducing vulnerability to interference, memory retrieval appeared to aid the preservation of existing sequence knowledge relative to a no-retrieval control group. These findings suggest that memory retrieval followed by new learning does not reliably induce memory updating via reconsolidation.

Significance Statement: Reconsolidation-updating theory suggests that existing memory traces can be modified, or even erased, by post-retrieval new learning. Compelling empirical support for this claim could have profound theoretical, clinical, and ethical implications. However, demonstrating reconsolidation-mediated memory updating in humans has proved particularly challenging. In four direct and three conceptual replication attempts of a prominent human reconsolidation study, we did not observe any reconsolidation effects when testing either procedural or declarative recall of sequence knowledge. These findings suggest that the considerable theoretical weight attributed to the original study is unwarranted and that post-retrieval new learning does not reliably induce human memory updating via reconsolidation.

Reconsolidation theory proposes that retrieval of existing memory traces causes them to destabilize, triggering a transient molecular restabilization ('reconsolidation') process during which they are open to modification (1, 2). If reconsolidation enables memory modification in humans, it could have profound theoretical (3), clinical (4), and ethical (5) implications. For example, the ability to erase 'pathological' memory traces that contribute to post-traumatic stress disorder, addiction, and phobias, offers the potential of permanent relief from these conditions (4).

Proponents of reconsolidation theory suggest that there is broad empirical support across a range of species, tasks, and memory types (2, 4, 6), but several authors have expressed skepticism about the extent to which existing studies rule out alternative explanations (7-9). Extending reconsolidation investigations to human participants has proved particularly challenging. Support for the theory is largely based on non-human animal studies in which invasive interventions, such as electroconvulsive shock or pharmacological treatment, are delivered following retrieval of an established memory trace (2, 6). By contrast, ethical constraints have led to the use of *new learning* as a post-retrieval intervention in many investigations with humans (e.g., 10-12; for review see 13). In both cases, the observation of substantial trace-dependent performance impairments on a subsequent test is taken as evidence that the intervention has disrupted the reconsolidation of the memory trace, resulting in its modification or destruction. Although physiological interventions are intended to directly disrupt the putative molecular substrates of reconsolidation, considerable ambiguity surrounds the envisioned mechanism by which a behavioural intervention might influence these same processes. Nevertheless, there are prevalent claims about the functional role of reconsolidation as a memory 'updating' mechanism (14-16) whereby existing memory traces are selectively 'rewritten' by post-retrieval new learning (12, 17).

It is worth noting that the reconsolidation controversy is only the latest chapter in an enduring historical debate about the locus of interference and forgetting effects (18). On the one hand, amnesia for previously recallable information has been attributed to *storage deficits*: the permanent physical modification of memory traces by post-encoding and post-retrieval interventions (e.g., *consolidation*, 19; *unlearning*, 20; *destructive updating*, 21; *reconsolidation*, 2). On the other hand, amnesia has been attributed to mechanisms operating during trace retrieval that temporarily modulate trace-dependent performance without necessarily influencing the underlying memory trace (e.g., *response competition*, 22; *cue-*

dependent forgetting, 23; *state-dependent retrieval*, 24; *context-dependent forgetting*, 25). These *retrieval-deficit* accounts can explain experimentally induced amnesia without invoking claims about physical trace disruption that cannot be directly observed. They also provide a more convincing account of the widespread finding that impairments of trace-dependent performance are often temporary and show high propensity for recovery under favorable retrieval conditions (26). This debate is particularly pertinent to the evaluation of reconsolidation studies because retrieval-deficit explanations are often overlooked (7-9).

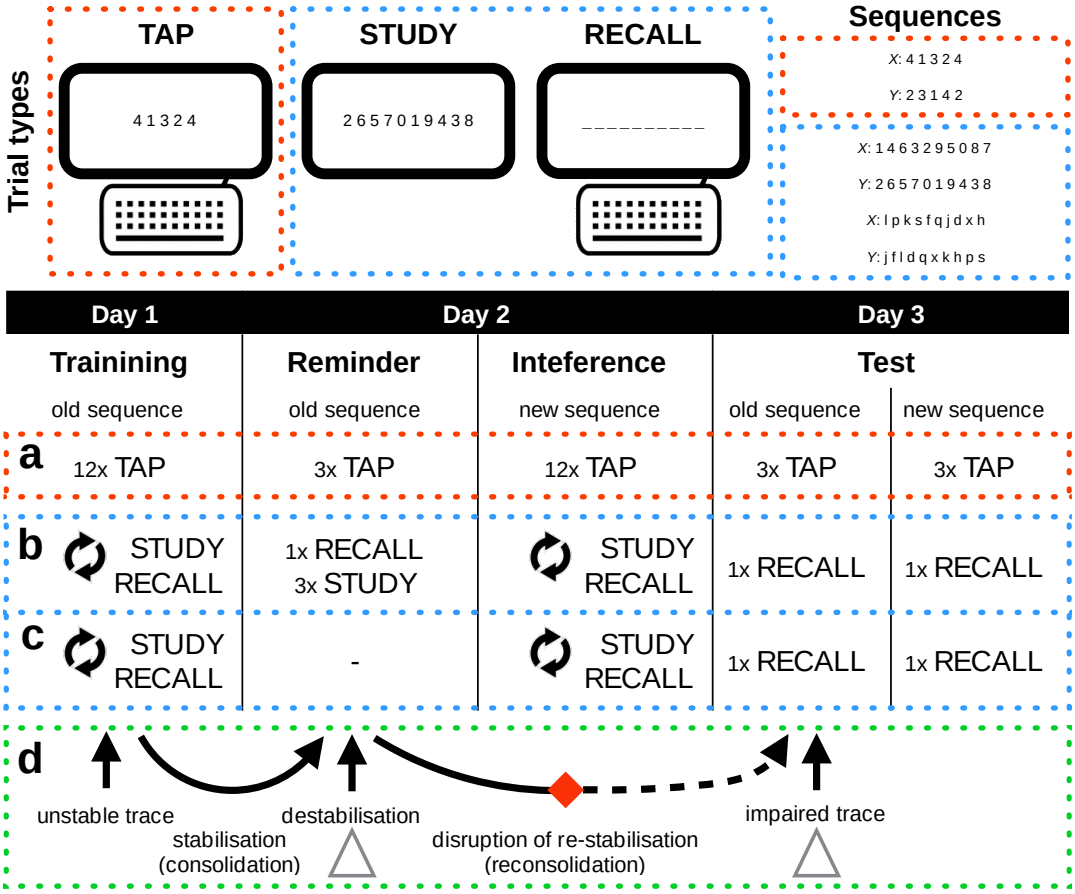

Fig. 1. Study design for Walker et al. (27) and direct replications (**a**, ⋯ red boundary), conceptual replications (⋯ blue boundary) with reminder condition (**b**) and without reminder condition (**c**), and hypothesized underlying mechanisms and events predicted by reconsolidation theory (**d**, ⋯ green boundary). Critical time-points for calculation of the Reconsolidation Score (RS) are indicated by triangle symbols. See main text for details.

A particularly prominent finding reported by Walker et al. (Group 7; 27) is widely cited as a convincing demonstration of reconsolidation-mediated memory updating in humans (e.g., 2, 4, 6, 13, 14). The results are especially compelling because the experiment conformed to the

canonical 3-day reconsolidation protocol (Fig. 1) typically used in non-human animal studies, thus meeting several key criteria necessary for a robust investigation of reconsolidation (2, 4, 6). On Day 1, participants used a computer keyboard to repeatedly tap a simple sequence of on-screen digits (e.g., 41342). Speed and accuracy improvements were observed as participants learned this initial ('Old') sequence. On Day 2, participants in the Reminder Group (n = 16) practised the Old Sequence immediately prior to learning a New Sequence. The No-Reminder Group did not practise the Old Sequence prior to new learning. The No-Intervention Group practised the Old Sequence but did not learn a New Sequence. On Day 3, sequence performance was tested for all groups. The key finding was that the Reminder Group's Old Sequence accuracy suffered a substantial decline (~57%) between the Reminder Stage and the Test Stage, although only minor decrements were observed on the speed measure (~2%). By contrast, improvements in accuracy and speed between Training and Test stages were observed in the No-Reminder and No-Intervention Groups. Therefore, it would appear that the accuracy impairment in the Reminder Group was contingent on the time-dependent interaction of the reminder and intervention as demonstrated in similar non-human animal studies (1) and widely accepted as evidence for reconsolidation (2, 4, 6). Consistent with the view that the Old Sequence memory trace had been 'rewritten' by the new learning (12, 17), the authors suggested that reconsolidation may have 'functional significance', allowing the 'continued refinement and reshaping of previously learned movement skills' (27, p. 618).

However, from the perspective of the aforementioned storage-retrieval debate (18) this interpretation should be viewed with caution, especially as retrieval-deficit explanations were not explored. For example, it was not clear whether the effect endured beyond the 3-day study period, or showed propensity for recovery under favorable retrieval conditions (26), effects that have been observed in several investigations of reconsolidation with non-human animals (e.g., 28-30). In the present study we initially sought to replicate and extend the reported reconsolidation effect (27, Group 7) by examining whether it could be accounted for by retrieval deficits rather than the storage deficit mechanisms outlined under reconsolidation theory (our investigation does not address other findings, unrelated to reconsolidation, reported in the same article). We conducted a replication battery (31) consisting of both *direct replications* (32) that followed the methodology of the original study as closely as possible, and *conceptual replications* (33) that manipulated key task parameters in order to explore the broader validity of the reconsolidation-updating theory.

To foreshadow our findings, the complete absence of a reconsolidation effect in any of our experiments precluded any further investigation of a retrieval-deficit account. Instead, we made several attempts to reproduce the effect in repeated direct replications (N = 64) using our own software (Experiment 1), software provided by the original researchers (Experiment 2), and under conditions intended to increase task difficulty (Experiments 3 and 4). In our conceptual replications (N = 48), we used 'declarative' recall conditions more consistent with the wider human reconsolidation literature (e.g., 10, 11). These experiments also involved sequence learning within a 3-day reconsolidation protocol (see Fig 1), but used sequences similar in length and structure to phone numbers (Experiments 5 and 7) or computer passwords (Experiment 6). A No-Reminder control group (Experiment 7) enabled us to ascertain whether performance impairments were contingent on retrieval-induced vulnerability as predicted by reconsolidation theory.

Results

All data (Dataset S1, Dataset S2) and analysis scripts are publically available on the Open Science Framework (https://osf.io/gpeq4/). All experiments and measures are reported. Unequal variances in between-subject comparisons were addressed by using Welch *t*-tests. Statistical significance was defined at the .05 level.

Direct replications (Experiments 1-4)

Consistent with the original experiment, we observed time-dependent improvements in accuracy and speed across the course of the Training and Interference Stages, and overnight between stages (see Fig. 2 and SI Results). The critical index of a reconsolidation effect (the percentage difference between Old Sequence performance at the Reminder Stage and Test Stage; from herein *Reconsolidation Score or RS*; see Fig. 1 triangles), completely contradicted the finding of the original study (27): we observed only minor fluctuations around zero for both accuracy (Fig. 3a) and speed (Fig. 3b) in all four direct replication attempts (Experiments 1-4). Minor procedural differences between the replications and the original study (variability in participant age and time of testing) were ruled out as potential confounds through additional analyses (see SI Results). Averaged across experiments, mean RS declined by <1% for accuracy, compared to ~57% in the original study, and increased by ~4% for speed. One-sample *t*-tests (one-tailed) indicated that none of the Reconsolidation Scores (Table 1) obtained in the direct replications were significantly less than zero.

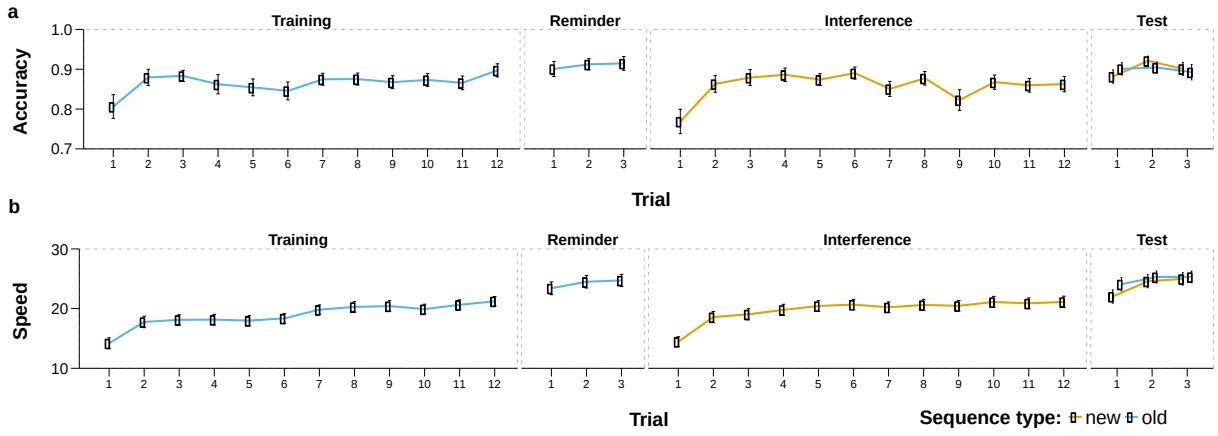

Fig. 2. Full study timeline showing mean accuracy (**a;** number of errors made relative to the number of complete sequences achieved) and mean speed (**b;** number of complete sequences achieved) by stage (Training, Reminder, Interference, and Test), trial, and sequence type, for Experiments 1-4 (pooled). A full definition of these dependent variables is available in the SI Methods. Error bars show ± SEM.

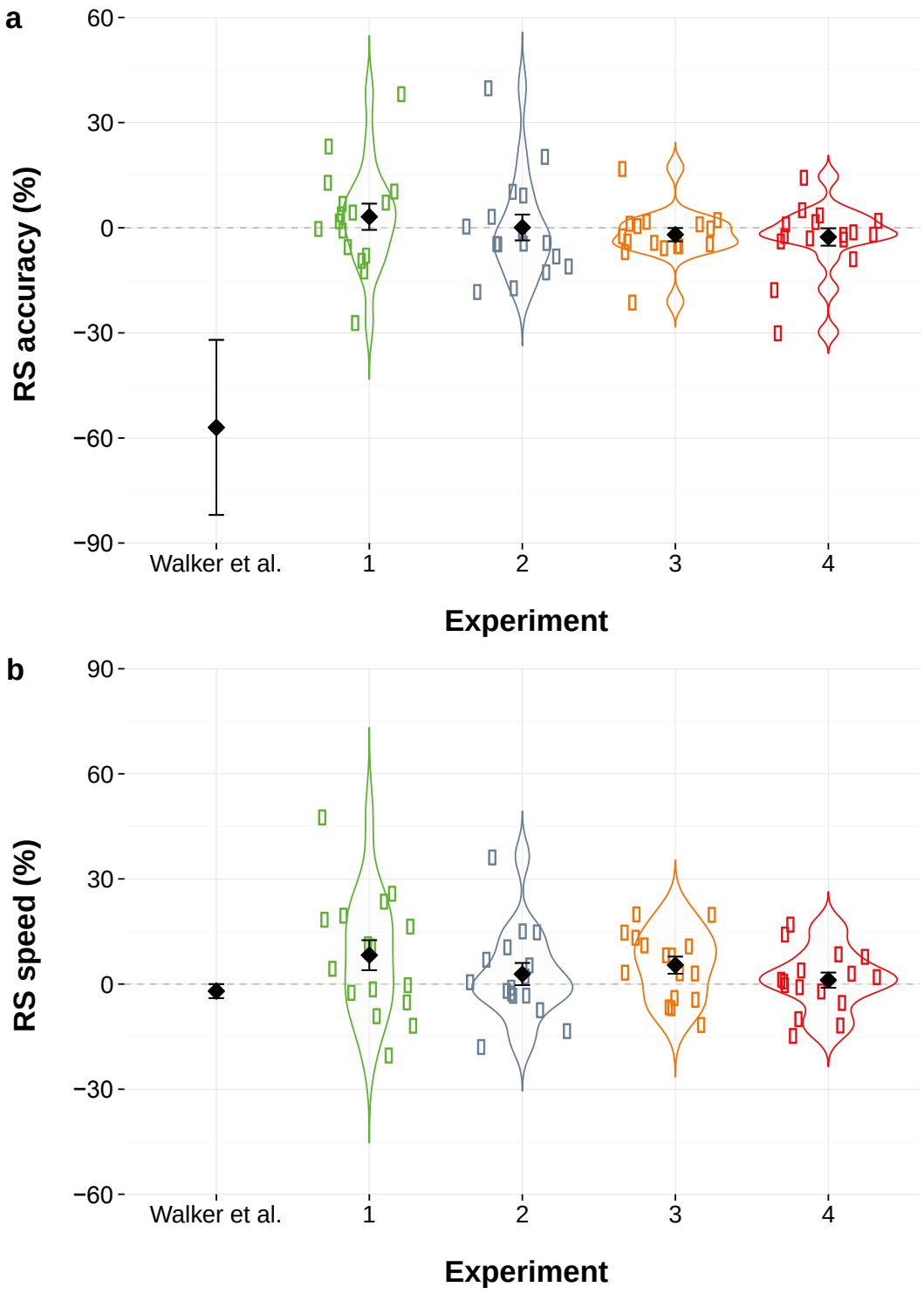

Fig. 3. Accuracy (**a**) and speed (**b**) reconsolidation scores (RS) for Walker et al. (27); n = 16) and Experiments 1-4 (N = 64). Black diamonds represent means and error bars show SEM. Where raw data are available (Experiments 1-4), individual participant scores (circles) and kernel density distributions are also depicted.

As the inherent limitations of Null Hypothesis Significance Testing constrain the degree to which one can determine the strength of evidence in favor of the null hypothesis (34), we also conducted a Bayesian analysis which enabled us to quantify the evidence in favor of the null hypothesis H_0 (RS = 0) relative to the reconsolidation hypothesis H_1 (RS < 0). Specifically, we calculated directional Bayes Factors (35) using an 'objective' JZS prior (Cauchy distribution with scale $r = 1$). H_1 was based on the general prediction of reconsolidation theory that trace-dependent performance should be reduced following disrupted reconsolidation of the reactivated trace (2, 4, 6). In all experiments, Bayes Factors (BF_{01}; see Table 1) were larger than 1, indicating greater evidentiary support for H_0 relative to H_1.

Table 1. Direct replication Reconsolidation Score statistics for accuracy and speed.

Exp	*DV*	*RS*	*SD*	*t(15)*	*p*	BF_{01}
1	Accuracy	3.13	14.95	.84	.79	8.97
	Speed	8.24	17.14	1.92	.96	13.59
2	Accuracy	.08	14.75	.02	.51	5.38
	Speed	2.85	12.74	.89	. 81	9.23
3	Accuracy	-1.98	7.58	-1.04	.16	1.90
	Speed	5.41	9.75	2.22	.98	14.68
4	Accuracy	-2.64	9.88	-1.07	.15	1.85
	Speed	1.12	8.70	.52	.69	7.52

Exp, experiment. *DV*, dependent variable. *RS*, mean Reconsolidation Score. *SD*, standard deviation. BF_{01}, Bayes Factor quantifying evidence in favor of the null hypothesis (RS = 0) relative to the reconsolidation hypothesis (RS< 0).

Meta-analysis. A primary goal of replication attempts is to facilitate more precise estimates of effect-size magnitude (36). However, in light of the stark discrepancy between the finding

observed in the original experiment (27, $N = 16$) and the four direct replications ($N = 64$), we focused on assessing the extent to which the collated evidence indicated that the phenomenon exists *at all*. That is to say, we aimed to establish whether the effect is qualitatively reproducible, as non-replication will preclude attempts to derive greater quantitative precision in the estimation of the effect's magnitude.

Directional meta-analytic Bayes Factors employing *t*-values for Experiments 1-4 (see Table 1) indicated greater evidentiary support for the null hypothesis (RS= 0) relative to the alternative hypothesis (RS< 0) for both accuracy ($BF_{01} = 5.743$) and speed ($BF_{01} = 36.027$). This pattern remained after incorporating an estimated *t*-value for the original study (accuracy: $BF_{01} = 2.080$; speed: $BF_{01} = 31.317$). The complete absence of predicted outcomes across these four experiments suggests that the reconsolidation effect reported in Group 7 of the original study (27) is not robust.

Conceptual replications (Experiments 5-7)

In the second component of the replication battery we aimed to evaluate the broader validity of reconsolidation-updating theory. These experiments also involved sequence learning within a 3-day reconsolidation protocol (see Fig 1), but used sequences similar in length and structure to phone numbers (Experiments 5 and 7) or computer passwords (Experiment 6), and required declarative (rather than procedural) recall at the Test Stage. Performance was assessed using a string-matching algorithm that provided an index of similarity between the target sequence and the sequence entered by the user (see SI Methods). This afforded a sensitive measure of partial (or 'chunked') sequence knowledge. As the pattern of performance did not vary significantly between Experiment 5 and Experiment 6 ('Reminder Experiments'), these data were pooled for display (Fig. 4) and subsequent analyses. Participants learned either a number (Experiment 5) or letter (Experiment 6) sequence to a criterion on Day 1 (Training Stage). On Day 2 these sequences were recalled (Reminder Stage) prior to new learning (Interference Stage), and on Day 3 recall of the sequences was evaluated (Test Stage). In the No-Reminder control group (Experiment 7) there was no Reminder Stage, permitting a comparison of Day 3 recall in the presence or absence of the Day 2 reminder.

During the Training and Interference Stages all participants successfully reached the criterion of 5 consecutive errorless sequence recalls (i.e., a maximum similarity score of 1.0) indicating successful learning of both the Old and New Sequences (see SI Results). A one-way

repeated-measures ANOVA indicated significant changes across stages (Training, Reminder, Test) for the Reminder Experiments ($F(2,90) = 8.68$, $p < .001$). Follow-up paired *t*-tests (one-tailed) showed that there was significant decline from the Training Stage (1.0) to the Reminder Stage ($M = .750$, $SD = .246$; $t(31) = 5.742$, $p < .001$), and from Reminder Stage to the Test Stage ($M = .638$, $SD = .315$; $t(31) = 2.645$, $p < .001$). The No-Reminder control group (Experiment 7) enabled us to ascertain whether the observed recall impairments could be causally attributed to the time-dependent interaction of memory reactivation and interference as predicted by reconsolidation theory (2, 4, 6). Despite the absence of a Reminder Stage, these participants also showed a substantial performance decrement from Training (1.0) to Test ($M = .488$, $SD = .363$). A paired-samples *t*-test (two-tailed) confirmed that this decline was significant ($t(15) = 5.646$, $p < .001$).

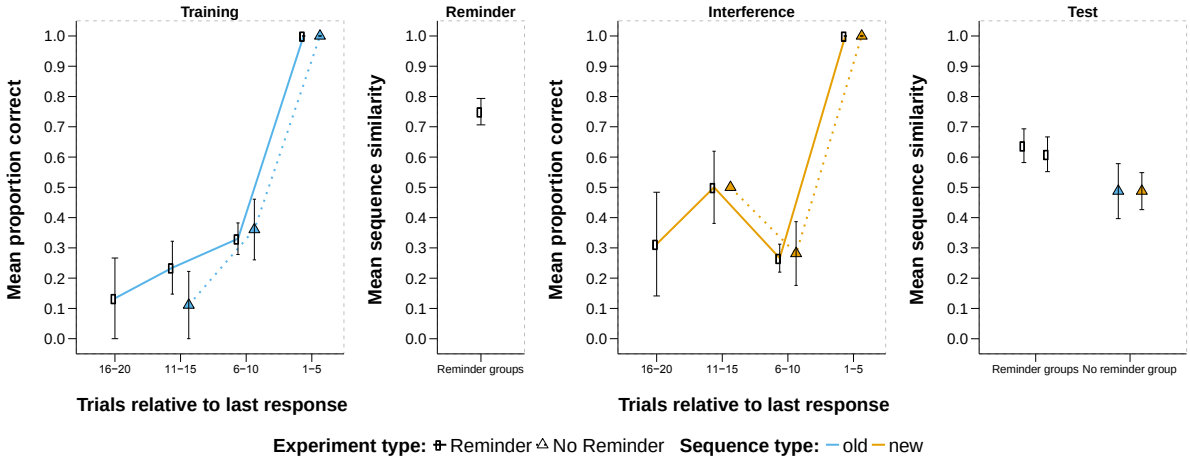

Fig. 4. Full study timeline showing performance in Experiments 5 and 6 pooled (Reminder groups; $n = 32$), and Experiment 7 (No Reminder group; $n = 16$). The Training and Interference panels show mean proportion correct on $\text{Recall}_{\text{feedback}}$ trials across 5 trial bins plotted relative to participants' final response of the stage. All participants reached the performance criterion (5 correct trials in a row) but required a different number of trials to do so (see SI Results). The small number of participants who took more than 20 trials to reach criterion (Training: $n = 2$, maximum trials = 29; Interference: $n = 1$, maximum trials = 22) contribute to all relevant analyses. The Reminder and Test panels show mean sequence similarity between the target sequence and the user-entered sequence assessed on a single $\text{Recall}_{\text{NoFeedback}}$ trial for each previously learned sequence (Old and New). Error bars show SEM.

These findings imply that at least some of the recall impairment observed in the Reminder Experiments was not contingent on the provision of a reminder-triggered reconsolidation process. Furthermore, a between-group comparison of Test Stage performance indicated poorer recall in the No-Reminder Experiment (M = .488, SD = .363) than in the Reminder Experiments (M = .638, SD = .315), and a two-sample t-test (one-tailed) indicated no significant difference ($t(26.59) = -1.409$, $p = .915$). Rather than inducing a state of increased susceptibility to interference, memory reactivation resulted in numerically *less* recall impairment of the Old Sequence at the Test Stage relative to no memory reactivation, an effect in the opposite direction to that predicted by reconsolidation theory.

Discussion

Reconsolidation-updating theory suggests that retrieval of an existing trace in the human memory system can render that trace vulnerable to modification from post-retrieval new learning. In the present investigation we attempted to replicate and extend a critical finding (27) widely considered to provide a compelling demonstration of reconsolidation-mediated memory updating in humans. In four direct replication attempts involving procedural recall and three conceptual replication attempts involving declarative recall, we did not observe the critical impairment effects observed in the original study and predicted by reconsolidation theory (2, 4, 6).

The findings of our direct replications are consistent with several recent investigations that employed variations of the original paradigm (27), with either transcranial magnetic stimulation (37, 38) or new learning (39) interventions. Although the findings of these studies were interpreted as favorable evidence for reconsolidation theory, the expected Reminder-Test performance decrement was actually absent in most conditions, including close replications of the original study (27, 39). When trace-dependent performance decrements were observed at short reminder lengths, they were modest, and rapidly recovered within the test session (39). It is difficult to reconcile the absence of performance impairments and the presence of recovery effects with the prediction of permanent trace modification (2). Thus, the outcomes of all three studies (37-39), are inconsistent with both the original study (27) and reconsolidation theory in general (2, 4, 6).

The findings of our conceptual replications cast further doubt on the veracity of claims that memory updating can be mediated by reconsolidation processes (12, 14-17). These

experiments adhered to the canonical 3-day reconsolidation protocol and aimed to increase external validity through the use of sequences similar in structure to phone numbers or computer passwords. In addition, consistent with several studies in the human reconsolidation literature (e.g., 10-12), participants completed a declarative recall task. Under these conditions, performance impairments occurred in the both the presence and absence of memory retrieval. Rather than triggering a state of heightened trace-vulnerability, retrieval actually led to numerically higher performance than in the no-retrieval control group. This finding is consistent with previous investigations of reconsolidation that found retrieval practice can afford some protection against interference (e.g., 40), and a considerable body of evidence suggesting that retrieval aids rather than impairs subsequent recall (41).

Two notable aspects of human reconsolidation research are not directly addressed by the present investigation. Firstly, there is evidence that post-retrieval pharmacological interventions can attenuate emotional responding in a fear-conditioning paradigm (42). However, the reliability of these effects has also recently come under scrutiny (43). Similarly, initially promising findings based on using post-retrieval extinction to disrupt reconsolidation in a fear-conditioning paradigm (17) have proved elusive in subsequent replication attempts (44, 45). It is striking that declarative recall of the CS-US contingency remains intact across these fear-conditioning studies, either in the presence or absence of effects on emotional responding.

Additionally, it has been suggested that 'prediction error' is a necessary reconsolidation trigger (11, 46). If this were the case, it could explain the absence of reconsolidation effects in the present replications. However, it would also be surprising that a reconsolidation effect was observed in the original study because the reminder protocol required participants to practice the Old Sequence in its entirety, and thus presumably did not invoke prediction error. To justify an auxiliary theoretical assumption about prediction error one would need to reconcile a considerable amount of contradictory evidence. For example, relative to controls, *no impairment of declarative recall* is observed in the aforementioned prediction error studies (11, 46), only attenuation of emotional responding (46), or ambiguous null effects on an indirect measure of trace integrity (retrieval-induced forgetting; 11). Furthermore, reconsolidation-like effects have been reported when the reminder involves reinforced trials (and thus no prediction error) in both non-human animals (47) and humans (12), and impairment effects are absent even in studies where prediction error would be expected (44, 45). At present therefore, it is

unclear whether prediction error is either necessary or sufficient for reconsolidation effects to emerge.

Taken together, our findings cast doubt on the efficacy of 'new learning' interventions as a means for disrupting the reconsolidation of procedural or declarative memory in humans. The absence of reconsolidation effects in all four direct replications suggests that the considerable theoretical weight attributed to the original study (2, 4, 6, 13, 14) is unwarranted. Furthermore, the absence of retrieval-contingent impairment in the conceptual replications is inconsistent with the purported functional role of reconsolidation as an adaptive mechanism that underlies memory updating (12, 14-17). Replication will be an essential tool in future reconsolidation investigations as researchers seek to verify the reliability of existing findings, identify genuine boundary conditions, and foster theoretical progress.

Method

All experimental programs and verbatim materials are publically available on The Open Science Framework (https://osf.io/gpeq4/). All participants provided informed consent and the local UCL ethics committee approved the study.

Direct replications (Experiments 1-4)

Participants. 16 participants were randomly allocated to each of the 4 direct replication experiments affording a total sample size of 64 individuals (49 female; median age = 22 years, age range = 18-54 years). 2 additional participants were excluded for typing an incorrect sequence at the Reminder Stage and 4 additional participants did not complete all 3 stages of the study. Participants were recruited from the University College London (UCL) mixed-occupation subject pool and received either monetary compensation or course credits. All participants reported that they were right-handed and had no history of neurological, psychiatric, or sleep disorder.

Design. Participants performed a 'finger-tapping' sequence learning task in three discrete sessions taking place on consecutive days (see Fig. 1). Two five-digit sequences (*X*: 4–1–3–2–4; *Y*: 2–3–1–4–2) were assigned to be the *Old Sequence* and the *New Sequence* in counter-balanced order. On Day 1, participants completed 12 Old Sequence trials (*Training*). On Day 2, participants performed 3 Old Sequence trials (*Reminder*)

immediately prior to 12 New Sequence trials (*Interference*). On Day 3, participants completed 3 trials of both the Old Sequence and the New Sequence in counterbalanced order (*Test*). The dependent variables (see SI Methods for details) were the number of sequences completed during each 30-second trial (‘*speed’*) and the ratio of errors to speed [‘*accuracy’*; 1-(errors/speed)].

Procedure. Unless otherwise stated (see SI Methods), the following procedures were used in all direct replications and precisely matched those reported in the original study (27). Ambiguous or missing information was clarified through contact with the senior author of the original research team. Participants were seated in front of a computer screen in a quiet room and used the four fingers of their left (non-dominant) hand to respond using the four top-row numeric keys 1, 2, 3 and 4 of a standard keyboard. The task involved repeatedly tapping a five-element sequence that was displayed on the screen for 30 seconds (including on ‘test’ trials), followed by 30 seconds of rest during which the sequence was absent. Key presses were acknowledged with white dots that accumulated on screen, but there was no feedback regarding response accuracy. A 30 second countdown timer was displayed during the rest phase to signal the approaching test phase. During the tapping phase the screen background was green and during the rest phase it was red. Participants were instructed to “tap out the sequence as quickly and accurately as possible.” There was no within- or between-subjects timing variability in the original study because all sessions were conducted at 1 p.m. In the present experiments there was also no within-subject variability: participants completed sessions at precise 24-hour intervals (± 15 minutes), however session times varied between participants (9am-6pm).

Conceptual replications

Participants. 16 participants were randomly allocated to each of the 3 conceptual replication experiments affording a total sample size of 48 individuals (38 female; median age = 22 years, age range = 18-52 years). 3 additional participants were excluded as they did not complete all 3 stages of the study. Participants were recruited from the UCL mixed-occupation subject pool and received either monetary compensation or course credits. All participants reported that they were right-handed and had no history of neurological, psychiatric, or sleep disorder.

Design. Participants performed a sequence-learning task in three discrete sessions taking place on consecutive days (see Fig. 1). Two ten-item sequences with independent grammars (see SI Methods) were assigned to be the *Old Sequence* and the *New Sequence* in counter-balanced order. For Experiments 5 and 7, the sequences were numbers (*X*: 1-4-6-3-2-9-5-0-8-7; *Y*: 2-6-5-7-0-1-9-4-3-8). For Experiment 6, the sequences were letters (*X*: l-p-k-s-f-q-j-d-x-h; *Y*: j-f-l-d-q-x-k-h-p-s). On Day 1, an adaptive test-feedback protocol was employed to ensure that all participants could recall the Old Sequence unassisted 5 times in a row (*Training*). On Day 2, Participants in Experiments 5 and 6 recalled and re-studied the Old Sequence immediately prior to new learning (*Reminder*). All participants learned the New Sequence in the same manner as Old Sequence Training (*Interference*). On Day 3, participants were asked to recall both sequences in counter-balanced order (*Test*). The dependent variable was a metric of the similarity between the target (Old/New) sequence at a given stage and the sequence entered by the user (*'sequence similarity'*; see SI Methods for details).

Procedure. Participants were seated in front of a computer screen in a quiet room and responded using a standard keyboard. On STUDY trials, participants were instructed to memorize the sequence whilst it was displayed on screen for 5-seconds. No response was required. On $\text{RECALL}_{\text{Feedback}}$ trials, participants were asked to enter the sequence from memory into 10 blank placeholders (_). Correctly entered items appeared in green. Entering an item in an incorrect order caused that item to flash in red and black (4x0.5s flashes over 2s) followed by replacement with the correct item, which flashed in green and black (4x0.5s flashes over 2s), and early termination of the trial. On $\text{RECALL}_{\text{NoFeedback}}$ trials, participants also had to enter the sequence from memory, however the trial was not interrupted if they entered items in an incorrect order and they could make corrections if they wished. All items appeared in black so there was no feedback on these trials.

The Training and Interference Stages involved iterative cycles of STUDY and $\text{RECALL}_{\text{Feedback}}$ trials starting with the former. Accurately entering the whole sequence on a $\text{RECALL}_{\text{Feedback}}$ trial led to additional $\text{RECALL}_{\text{Feedback}}$ trials. Failure to complete a $\text{RECALL}_{\text{Feedback}}$ trial resulted in a STUDY trial and the cumulative $\text{RECALL}_{\text{Feedback}}$ counter was reset. When the participant had achieved 5 accurate $\text{RECALL}_{\text{Feedback}}$ trials in a row the stage was terminated.

The Reminder Stage involved a single $RECALL_{NoFeedback}$ trial followed by two STUDY trials. The Test Stage involved two $RECALL_{NoFeedback}$ trials where participants were asked to "Recall the OLD sequence from day one and enter it on the next screen" and, separately, "Recall the NEW sequence from day two and enter it on the next screen". Full participant instructions for each stage are available (see SI Methods). Participants completed sessions at precise 24-hour intervals (± 15 minutes), however session times varied between participants (9am-6pm).

Acknowledgments. We thank the UCL Institute of Making for assistance with apparatus for Experiment 4. This work was supported by the Economic and Social Research Council [grant number ES/J500185/1].

References

1. Nader K, Schafe GE, Le Doux JE (2000) Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature* 406(6797):722–726.

2. Nader K, Hardt O (2009) A single standard for memory: the case for reconsolidation. *Nat Rev Neurosci* 10(3):224–234.

3. Nadel L, Land C (2000) Memory traces revisited. *Nat Rev Neurosci* 1(3):209–212.

4. Schwabe L, Nader K, Pruessner JC (2014) Reconsolidation of human memory: Brain mechanisms and clinical relevance. *Biol Psychiat* 76(4):274–280.

5. Hui K, Fisher CE (2014) The ethics of molecular memory modification. *J Med Ethics* 0:1–6.

6. Tronson NC, Taylor JR (2007) Molecular mechanisms of memory reconsolidation. *Nat Rev Neurosci* 8(4):262–275.

7. Cahill L, McGaugh JL, Weinberger NM (2001) The neurobiology of learning and memory: some reminders to remember. *Trends Neurosci* 24(10):578–581.

8. Riccio DC, Millin PM, Bogart AR (2006) Reconsolidation: A brief history, a retrieval view, and some recent issues. *Learn Mem* 13(5):536–544.

9. Millin PM, Moody EW, Riccio DC (2001) Interpretations of retrograde amnesia: old problems redux. *Nat Rev Neurosci* 2(1):68–70.

10. Hupbach A, Gomez R, Hardt O, Nadel L (2007) Reconsolidation of episodic memories: a subtle reminder triggers integration of new information. *Learn Mem* 14(1-2):47–53.

11. Forcato C, Argibay PF, Pedreira ME, Maldonado H (2009) Human reconsolidation does not always occur when a memory is retrieved: The relevance of the reminder structure. *Neurobiol Learn Mem* 91(1):50–57.

12. Chan JCK, LaPaglia JA (2013) Impairing existing declarative memory in humans by disrupting reconsolidation. *P Natl Acad Sci USA* 110(23):9309–9313.

13. Schiller D, Phelps EA (2011) Does reconsolidation occur in humans? *Front Behav Neurosci* 5:1-12.

14. Lee JLC (2009) Reconsolidation: maintaining memory relevance. *Trends Neurosci* 32(8):413–420.

15. Dudai Y (2009) Predicting not to predict too much: how the cellular machinery of memory anticipates the uncertain future. *Philos Trans R Soc Lond B Biol Sci* 364(1521):1255–1262.

16. Hardt O, Einarsson EÖ, Nader K (2010) A bridge over troubled water: Reconsolidation as a link between cognitive and neuroscientific memory research traditions. *Annu Rev Psychol* 61(1):141–167.

17. Schiller D, et al. (2009) Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature* 463(7277):49–53.

18. Miller RR, Matzel LD (2006) Retrieval failure versus memory loss in experimental amnesia: Definitions and processes. *Learn Mem* 13(5):491–497.

19. McGaugh JL (2000) Memory--a century of consolidation. *Science* 287(5451):248–251.

20. Melton AW, Irwin JM (1940) The influence of degree of interpolated learning on retroactive inhibition and the overt transfer of specific responses. *Am J Psychol* 53(2):173–203.

21. Loftus EF (1979) The malleability of human memory: Information introduced after we view an incident can transform memory. *Am Sci* 67(3):312–320.

22. McGeoch JA (1942) *The psychology of human learning: An introduction* (New York: Longmans).

23. Tulving E (1974) Cue-dependent forgetting: When we forget something we once knew, it does not necessarily mean that the memory trace has been lost; it may only be inaccessible. *Am Sci* 62(1):74–82.

24. Eich JE (1980) The cue-dependent nature of state-dependent retrieval. *Mem Cogn* 8(2):157–173.

25. Capaldi EJ, Neath I (1995) Remembering and forgetting as context discrimination. *Learn Mem* 2(3-4):107–132.

26. Bouton ME (2002) Context, ambiguity, and unlearning: sources of relapse after behavioral extinction. *Biol Psychiat* 52(10):976–986.

27. Walker MP, Brakefield T, Allan Hobson J, Stickgold R (2003) Dissociable stages of human memory consolidation and reconsolidation. *Nature* 425(6958):616–620.

28. Power AE, Berlau DJ, McGaugh, JL, Steward, O (2006) Anisomycin infused into the hippocampus fails to block "reconsolidation" but impairs extinction: The role of re-exposure duration. *Learn Mem* 13(1):27–34.

29. Lattal KM, Abel T (2004) Behavioral impairments caused by injections of the protein synthesis inhibitor anisomycin after contextual retrieval reverse with time. *P Natl Acad Sci USA* 101(13):4667–4672.

30. Eisenberg M, Dudai Y (2004) Reconsolidation of fresh, remote, and extinguished fear memory in medaka: old fears don't die. *Eur J Neurosci* 20(12):3397–3403.

31. Rosenthal R (1990) Replication in behavioral research. *J Soc Behav Pers*. 5(4):1-30.

32. Simons DJ (2014) The value of direct replication. *Perspect Psychol Sci* 9(1):76–80.

33. Schmidt S (2009) Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Rev Gen Psychol* 13(2):90–100.

34. Dienes Z (2011) Bayesian versus orthodox statistics: Which side are you on? *Perspect on Psychol Sci* 6(3):274–290.

35. Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G (2009) Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon Bull Rev* 16(2):225–237.

36. Cumming G (2012) *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis* (Routledge, New York).

37. Censor N, Horovitz SG, Cohen LG (2014) Interference with existing memories alters offline intrinsic functional brain connectivity. *Neuron* 81(1):69–76.

38. Censor N, Dimyan MA, Cohen LG (2010) Modification of existing human motor memories is enabled by primary cortical processing during memory reactivation. *Curr Biol* 20(17):1545–1549.

39. de Beukelaar TT, Woolley DG, Wenderoth N (2014) Gone for 60 seconds: reactivation length determines motor memory degradation during reconsolidation. *Cortex* 59:138–145.

40. Potts R, Shanks DR (2012) Can testing immunize memories against interference? *J Exp Psychol Learn Mem Cogn* 38(6):1780–1785.

41. Roediger HL III, Butler AC (2011) The critical role of retrieval practice in long-term retention. *Trends Cog Sci* 15(1):20–27.

42. Kindt M, Soeter M, Vervliet B (2009) Beyond extinction: erasing human fear responses and preventing the return of fear. *Nat Neurosci* 12(3):256–258.

43. Bos MGN, Beckers T, Kindt M (2014) Noradrenergic blockade of memory reconsolidation: A failure to reduce conditioned fear responding. *Front Behav Neurosci*

8:412.

44. Golkar A, Bellander M, Olsson A, Öhman A (2012) Are fear memories erasable?–reconsolidation of learned fear with fear-relevant and fear-irrelevant stimuli. *Front Behav Neurosci* 6:1-10.

45. Kindt M, Soeter M (2013) Reconsolidation in a human fear conditioning study: A test of extinction as updating mechanism. *Biol Psychol* 92(1):43–50.

46. Sevenster D, Beckers T, Kindt M (2014) Prediction error demarcates the transition from retrieval, to reconsolidation, to new learning. *Learn Mem* 21(11):580–584.

47. Duvarci S, Nader K (2004) Characterization of fear memory reconsolidation. *J Neurosci* 24(42):9269–9275.

48. Lakens D (2013) Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Front Psychol* 4:1–12.

49. Suzuki A (2004) Memory reconsolidation and extinction have distinct temporal and biochemical signatures. *J Neurosci* 24(20):4787–4795.

50. Van der Loo M (2014) The stringdist package for approximate string matching. *R J* 6(1):111–122.

Supplementary Information

SI Results

Direct replications (Experiments 1-4)

Old Sequence Training. To establish whether there were performance gains during the Old Sequence Training Stage we employed two separate 4x12 mixed-factorial ANOVAs for accuracy and speed with one between-subjects variable (experiment: *1-4*) and one within-subjects variable (trial: *1-12*).

Accuracy (see Fig. 2a *Training*) increased numerically between trial 1 (M = .806, SD = .238) and trial 12 (M = .879, SD = .132), although neither the linear ($F(1, 60) = 2.96$, p = .091) nor quadratic ($F(1,60)$ = .105, p = .747) trend reached significance. There was no significant interaction between experiment and either linear ($F(3,60)$ = .580, p = .630) or quadratic ($F(3,60)$= 1.203, p = .316) trend of trial.

Speed (see Fig. 2b *Training*) increased numerically between trial 1 (M = 14.200, SD = 6.40) and trial 12 (M = 21.238, SD = 6.356). Linear ($F(1, 60) = 95.398$, $p < .001$) and quadratic

($F(1,60) = 12.908$, $p = .001$) trends both reached statistical significance. There was no significant interaction between experiment and linear trend of trial ($F(3,60) = 1.371$, $p = .260$), but the interaction between experiment and quadratic trend of trial reached significance ($F(3,60) = 3.209$, $p = .029$).

New Sequence Interference. The same ANOVA design was used to assess changes in New Sequence performance across the Interference Stage.

Accuracy (see Fig. 2a *Interference*) increased numerically between trial 1 (M = .769, SD = .246) and trial 12 (M = .862, SD = .155). Across trials there was a significant quadratic trend ($F(1,60) = 7.651$, $p = .008$). The linear trend was not significant ($F(1,60) = 1.087$, $p = .301$). There was no significant interaction between experiment and quadratic ($F(3,60) = 1.254$, $p = .274$) or linear ($F(3,60) = 1.327$, $p = .274$) trend of trial.

Speed (see Fig. 2b *Interference*) increased numerically between trial 1 (M =14.466, SD = 6.593) and trial 12 (M = 21.128, SD = 7.588). Linear ($F(1,60) = 83.075$, $p < .001$), and quadratic ($F(1,60) = 58.072$, $p < .001$) trends both reaching significance. There was no interaction between experiment and linear ($F(3,60) = 1.137$, $p = .342$) or quadratic ($F(3,60) = 1.362$, $p = .263$) trend.

Overnight performance changes. In the original study (27) the following comparisons were made to examine overnight changes in sequence performance:

- *Overnight Score Old* (OSO) was the percentage change between Old Sequence Training (trials 10-12 only) and Old Sequence Reminder (all 3 trials).
- *Overnight Score New* (OSN) was the percentage change between New Sequence Interference (trials 10-12 only) and New Sequence Test (all 3 trials).

Only the final three trials of the Training and Interference Stages have been used because calculating an average across all 12 trials could attenuate the true time-dependent performance changes achieved by the end of these stages (after 27). To establish whether the overnight scores varied between experiments, we employed a series of one-way ANOVAs with experiment (*1-4*) as a between-subjects factor and overnight score (separately for *old/new* and separately for *accuracy/speed*) as a dependent variable.

For accuracy, there was no significant main effect of experiment for OSO ($F(3,60) = 1.287$, $p = .287$) or OSN ($F(3,60) = .986$, $p = .406$). However, for speed there was a significant main effect of experiment for both OSO ($F(3,60) = 3.426$, $p = .023$) and OSN ($F(3,60) = 5.126$, $p = .003$). Consequently, we report follow-up tests for the data pooled across experiments (accuracy) or for each experiment individually (speed). One-sample *t*-tests (one-tailed) were employed to assess whether any performance changes between time-points were significantly greater than zero.

Consistent with the original study, we observed significant overnight accuracy improvements for OSO (OSO = 4.649, SD = 15.089, $t(63) = 2.465$, $p = .008$) and OSN (OSN = 5.638, SD = 15.081, $t(63) = 2.991$, $p = .002$) when data were pooled across experiments. In most cases improvements in speed were larger and more in keeping with the original study for both OSN and OSO (see Table S1).

Table S1. Overnight Scores for direct replications. Speed dependent variable only.

Experiment	Sequence	*OS*	*SD*	*t*(15)	*p*
1	Old	5.93	20.81	1.14	.14
	New	2.89	12.86	.90	.19
2	Old	16.72	14.16	4.72	< .001
	New	15.82	15.66	4.04	< .001
3	Old	19.72	16.87	4.68	< .001
	New	13.63	9.25	5.89	< .001
4	Old	26.00	19.90	5.23	< .001
	New	21.71	16.60	5.23	< .001

OS, Overnight Score. *SD*, standard deviation.

Impact of counterbalancing. Counterbalanced conditions were sequence order (*X* or *Y*; i.e., whether the Old Sequence was designated as 4-1-3-2-4 or 2-3-1-4-2, with the remaining sequence being assigned as the New Sequence) and test order (*A* or *B*; i.e., whether the Old or New Sequence was tested first on the Day 3 Test). Conditions were balanced in all experiments (N = 8 per condition) except in Experiment 2 where researcher error led to unbalanced conditions (test order: A = 12, B = 4; sequence order: X = 7, Y = 9). To establish whether the counter-balancing procedures influenced the Reconsolidation Score, we employed a series of one-way ANOVAs separately for test order (*A*, *B*) or sequence order (*X*, *Y*) as a between-subjects factor and Reconsolidation Score (separately for accuracy and speed), as a dependent variable.

There was no significant main effect of sequence order on RS accuracy ($F(1, 62) = .004$, $p = .948$) or RS speed ($F(1, 62) = .224$, $p = .638$). There was also no significant main effect of test order on RS accuracy ($F(1, 62) = .655$, $p = .421$), however test order did influence RS speed significantly ($F(1, 62) = 5.320$, $p = .024$). Follow-up one-sample *t*-tests (two-tailed) indicated that RS speed was significantly *higher* than zero for test order A (RS = 7.482, SD = 13.479, $t(35) = 3.331$, $p = .002$) and did not differ significantly from zero for test order B (RS = .448, SD = 10.044, $t(27) = .236$, $p = .815$), confirming that there was no reconsolidation effect in either condition.

Impact of training accuracy. To address concerns about ceiling effects in the accuracy data, we conducted a median split on the pooled data across all experiments and repeated the Reconsolidation Score analysis for accuracy. The pooled data were split based on the median accuracy score achieved on the final three trials of Old Sequence Training (.894). In the above-median group there were minor improvements in accuracy (RS = 1.98, $t(31) = .825$, $p = .792$) and the Bayes Factor (BF_{01} = 12.47) indicated greater evidentiary support for the null hypothesis (RS= 0) against the alternative hypothesis (RS< 0). In the below-median group there were minor non-significant decrements in accuracy (RS = -2.689, $t(31) = -1.496$, p = .072) and the Bayes Factor (BF_{01} = 1.374) indicated that the data are inconclusive when comparing the null hypothesis (RS= 0) against the alternative hypothesis (RS< 0). Therefore, even when below-average performers were examined in isolation, reconsolidation effects were not observed.

Extracting and estimating statistics from the original study. As neither raw or summary level data for the original study was available, we used plot-digitising software to extract

Reconsolidation Score values for accuracy (M = -57, SEM = 25) and speed (M = -2, SEM = 2) from the relevant graphs published in the original article (Fig. 4c in 27). Values were rounded to the nearest whole number. As *t*-values were not reported in the original article, we used these means and standard errors to recalculate them for use in meta-analysis.

Statistical power. Cohen's *d* effect sizes for Reconsolidation Scores in the original study (27) were calculated for accuracy (*d* = -.57) and speed (*d* = -.25). This was achieved using the estimated *t*-values (see *extracting and estimating statistics*) in the following formula (48):

$$\boldsymbol{d = \frac{t}{\sqrt{n}}}$$

Given these effect sizes, the use of directional one-sample *t*-tests, and an alpha level of .05, the power for any one of our experiments taken individually (*n* = 16) was 0.70. The combined power of all the direct replications (*N* = 64) was 0.99. As the direct replications overall had high statistical power to detect the effect size reported in the original study, it seems unlikely that our findings reflect a false-negative or "Type II error". In addition, the meta-analytic Bayesian analysis (see main text), which accounts for sample size, indicated greater evidentiary support for the null hypothesis relative to the reconsolidation hypothesis.

Influence of variability in session times and participant age range. Two minor procedural differences between the original study and the direct replications (participant age range and time of testing) were evaluated to see if they influenced the findings. Time of testing in the direct replications (rounded to the nearest hour: median = 15.00h, *SD* = 1.833) differed only slightly from Walker et al. (13.00h) and was not significantly correlated with reconsolidation scores (*r* = .12, *p* = .355). Therefore, time of testing cannot account for the absence of a reconsolidation effect.

Participant age in the direct replications (median = 22 years, range = 18-54 years) covered a larger range than Walker et al. (median unknown, range = 18-27 years). A re-analysis of reconsolidation scores for only those participants who fell within the 18-27 age bracket (*n* = 48), showed that there was still no substantial impairment (mean = -2.05, *SD* = 10.51, *t*(47) = -1.35, *p* = 0.092). Therefore, participant age cannot account for the absence of a reconsolidation effect.

Conceptual replications (Experiments 5-7)

Old Sequence Training. All participants were trained until they reached a performance criterion of 5 consecutive errorless recalls of the Old Sequence and recall failure resulted in additional study trials. This ensured that, regardless of idiosyncratic learning strategies, all participants robustly encoded both the Old and New sequences. There is some evidence from non-human animal studies to suggest that stronger memories are less amenable to reconsolidation than weaker ones (49). However, the specific parameters under which this potential moderator might operate are not well defined (2, 4). Furthermore, this seems unlikely to be an influential factor in this case as participants demonstrated below-ceiling performance at the Reminder Stage (M = .750, SD = .246; see *Old Sequence performance between stages* below). More trials were required to reach criterion in Experiment 6 (*Letters*; M = 13.00, SD = 6.623) than Experiment 5 (*Numbers*; M = 8.688, SD = 2.676) and Experiment 7 (*Numbers No Reminder*; M = 7.813, SD = 3.124). A one-way ANOVA indicated that the number of trials required to reach criterion varied significantly between experiments ($F(2,45) = 6.089$, $p <$ .001). Follow-up Welch two-sample t-tests indicated that Experiments 5 and 7 did not differ significantly ($t(29.31) = .851$, $p = .402$). However, participants required significantly more trials to reach criterion in Experiment 6 compared to Experiment 5 ($t(19.77) = -2.42$, $p = .026$). No participants failed to reach the performance criterion.

New Sequence Interference. The same learn-to-criterion procedure was employed in the Interference Stage as in the Training Stage. Although more trials were required to reach criterion in Experiment 6 (*Letters*; M = 10.375, SD = 4.938) than in Experiment 5 (*Numbers*; M = 8.563, SD = 4.397) and Experiment 7 (*Numbers - No Reminder*; M = 7.125, SD = 2.363), a one-way ANOVA indicated that these differences were not statistically significant ($F(2,45) =$ 2.583, $p = .087$). No participants failed to reach the performance criterion.

Old Sequence performance between stages. Following the Training Stage baseline (1.0), there were performance decrements in Old Sequence performance at the subsequent Reminder ($M =$ .756, SD = .253) and Test Stages (M = .606, SD = .315) in Experiment 5 (*Numbers*). A similar pattern was observed in Experiment 6 (*Letters*) with performance declining at the Reminder (M = .744, SD = .248) and Test Stages (M = .669, SD = .322).

A 2x3 mixed-factorial ANOVA (experiment: *5, 6*; stage: *Training, Reminder, Test*) showed that this performance decline across stages was statistically significant ($F(2,87) =$ 8.629, $p <$ .001). There was no main effect of experiment ($F(1,87)= 1.122$, $p = .292$), or interaction between experiment and stage ($F(1,30) = .672$, $p = .513$). As the overall pattern did

not vary between Experiments 5 and 6 ('Reminder Experiments'), we pooled the data for subsequent analysis (see main text).

New Sequence performance between stages. New Sequence performance declined from the Interference Stage baseline (1.0) to the subsequent Test Stage in Experiment 5 (M = .625, SD = .317), Experiment 6 (M = .594, SD = .342), and Experiment 7 (M = .488, SD = .245). A 3x2 mixed-factorial ANOVA (experiment: *5, 6, 7*; stage: *Interference, Test*) indicated that there was a main effect of stage ($F(1,88) = 15.425$, $p < .001$) and no main effect of experiment ($F(2,88) = .548$, $p = .580$) or interaction between experiment and stage ($F(2,88) = .548$, $p = .580$).

Impact of counterbalancing. We counterbalanced sequence order i.e., whether the Old Sequence (*X*) was designated as 1-4-6-3-2-9-5-0-8-7 / 2-6-5-7-0-1-9-4-3-8 (Experiments 5 and 7) or l-p-k-s-f-q-j-d-x-h / j-f-l-d-q-x-k-h-p-s (Experiment 6), with the remaining sequence being assigned as the New Sequence (*Y*). We also counter-balanced test order (*A* or *B*; i.e., whether the Old or New Sequence was tested first on the Day 3 Test). Conditions were balanced in all experiments (N = 8 per condition). To establish whether the counterbalancing procedures influenced sequence similarity scores at the Test Stage, two separate two-way ANOVAs with experiment (*5, 6, 7*) and either test order (*A, B*) or sequence order (*X, Y*) as between-subjects factors. There was no significant main effect of test order ($F(1,42) = .466$, $p = .498$), or interaction between test order and experiment ($F(2,42) = .723$, $p = .491$), and no main effect of sequence order ($F(1,42) = .714$, $p = .403$) or interaction between sequence order and experiment ($F(2,42) = .162$, $p = .851$).

SI Methods

Direct replications

Procedural variations. Each experiment had minor variations from the general procedure outlined in the main text. Unlike the other experiments, in Experiment 1 the sequence remained on screen during rest trials, there was no countdown timer, and the background color was invariant throughout. Key presses were acknowledged with the transient display of white dots arranged in a row that corresponded to the horizontal order of the physical keys. Experiments 1, 3, and 4 were executed in Python code developed by T.E.H. whereas Experiment 2 was run from an executable file provided by the original research team.

In the original study (27) and Experiments 1 and 2, participants were instructed to tap the sequence, "as quickly and accurately as possible". In Experiments 3 and 4, this instruction was modified to read "as quickly as you can. Try not to make errors, but overall you should emphasise speed over accuracy." The phrase "tap as quickly as you can!" was also displayed continuously on screen during test phases in Experiments 3 and 4. In Experiment 4, the keyboard was positioned in an adapted box file such that the participant was unable to view their hand during task performance. Tactile markers were placed on the response keys to prevent the participants' hand shifting to the incorrect keys. Participants were allowed to lift the lid of the box file during rest phases so they could stretch their fingers and ensure the hand was correctly positioned before closing the lid and starting the next trial.

Operationalizing accuracy and speed. The precise operationalization of the dependent measures reported in the original study (27) was ambiguous: "Performance measures were the number of complete sequences achieved ('speed'), and the number of errors made relative to the number of complete sequences achieved ('accuracy'). The original research team confirmed the following definitions:

- *speed* was the number of complete sequences achieved during a 30-second trial plus any partial sequence the participant was completing when the trial was terminated. For example, a participant who performed 15 complete sequences, and had just entered two correct items when the trial terminated, would receive a speed score of 15.4 (15 + 2/5);
- *accuracy* was 1 - (errors/speed) where a single error was defined as any string of up to 5 contiguous incorrect items that did not match the target sequence. For example, 3 contiguous incorrect items would constitute a single error, but 6 contiguous incorrect items would constitute 2 errors.

Note that under this scheme, it is technically possible for a participant to incur a negative accuracy score *on an individual trial* if error exceeds speed. This could substantially bias between-stage comparisons, as accuracy scores should only range between 0 and 1. Across the four experiments reported here, 5 trials with negative accuracy scores were identified (< .003% of total trials) and converted to zero. This did not impact the qualitative pattern of the results.

Conceptual replications

Operationalizing sequence similarity. The similarity between the target (Old/New) and user-entered sequences was measured using a normalized ratio of the Damerau–Levenshtein edit distance: a metric that indicates the number of 'fundamental' operations (substitution, deletion, insertion, or transposition) required to convert one character string into another and thus reflecting the 'similarity' of the two sequences (50).

Sequence construction. Sequences were generated with relatively unique grammars but used the same items in order to ensure a degree of old-new competition. To do this we first defined a 'base set' of 10 items, which were either randomly selected consonants (*l p k s f q j d x h*; Experiment 6) or single digits (*0-9*; Experiments 5 and 7). The first sequence was generated by randomly shuffling the order of these items. The second sequence was generated by repeatedly shuffling the first sequence until (a) all relative item positions (i.e., pairwise forward and backward transitions) were unique; and (b) all absolute item positions were unique. The same two sequences were used for all participants.