



Retrieval practice promotes semantic integration of real-world memories without compromising episodic integrity

Lauren A. Homann¹ · Mursal Jahed¹ · Morgan D. Barense^{1,2}

Received: 25 November 2025 / Accepted: 18 August 2026
© The Psychonomic Society, Inc. 2026

Abstract

Actively recalling information is a well-established memory strengthening technique that reliably outperforms passive review. Termed “the testing effect,” retrieval practice may promote semantic transformation and reshape related, unreviewed episodic information, beyond merely strengthening reviewed content. These broader effects of retrieval practice are difficult to capture with simplified laboratory materials but can be examined using rich, real-world events that evolve across multiple episodic and semantic dimensions. However, retrieval practice’s impact on such real-world memories remains poorly understood. We therefore investigated how actively retrieving event details (retrieval practice) versus passively reviewing the same details (restudy) influenced memory for an episodically rich staged event after two weeks ($n = 26$ per group). Retrieval increased the incorporation of preexisting semantic information into recall narratives, with greater semantic integration after retrieval practice predicting more episodically rich narratives. These findings suggest that retrieval practice enhanced integration of event details with semantic knowledge, perhaps via co-activation of semantically related content during retrieval. We found tentative evidence of a testing effect for real-world events: Retrieval improved retention of tour details (so long as they were successfully retrieved during retrieval practice) but did not otherwise affect broader episodic memory for unreviewed content, narrative accuracy, narrative temporal and semantic organization, or subjective phenomenology. Together, these findings suggest that retrieval reshapes memories by integrating recalled content into semantic knowledge networks, while preserving the overall integrity of episodic representations. Our work reveals that the testing effect may be underpinned by distinct transformations of episodic and semantic memory, with semantic transformation enhancing rather than compromising episodic richness.

Keywords Episodic memory · Retrieval practice · Integration · Naturalistic

Retrieval does not merely index stored knowledge—it reactivates and reshapes memory itself. This is exemplified by a cornerstone finding known as the *testing effect*—the phenomenon that *retrieval practice* (i.e., actively recalling information) improves memory for reviewed content more than *restudy* (i.e., reactivation via passive review; Carpenter, 2009; Kornell et al., 2011; Pyc & Rawson, 2010; Roediger & Karpicke, 2006; Rowland, 2014). Retrieval practice promotes robust, generalizable memory improvements for explicitly reviewed content across diverse test materials and formats

(Karpicke, 2017), demonstrating how retrieval uniquely strengthens memory beyond reexposure to the same content. Moreover, recent accounts suggest retrieval not only strengthens explicitly reviewed content but also influences memory representation and organization within broader memory networks (Antony et al., 2017; Ritvo et al., 2019).

However, uncovering these broader principles of retrieval practice may require moving beyond simplified laboratory materials to stimuli that capture real-world complexity. For example, does retrieval reshape unreviewed but related episodic content, or transform episodic memories semantically by integrating them with existing knowledge? These mechanisms might underlie the testing effect but are difficult to address within traditional paradigms typically reliant on structurally simple stimuli and outcome measures. In contrast, real-world events encompass vast interrelated event-specific details

✉ Lauren A. Homann
lauren.homann@mail.utoronto.ca

¹ Department of Psychology, University of Toronto, Toronto, ON, Canada

² Rotman Research Institute at Baycrest, Toronto, ON, Canada

embedded in spatiotemporal contexts and preexisting semantic knowledge networks. Real world events' multidimensionality permits studying retrieval's influence not only explicitly reviewed episodic content but also related but unreviewed details, relationships across episodes, and semantic transformation in an ecologically valid context. Although retrieval practice has been studied extensively in laboratory settings, its effects on real-world memories remain largely unexplored. Emmerdinger and Kuhbandner (2018) found no retrieval practice benefit over restudy for associations between laboratory-presented words and retrospectively recalled autobiographical memories (weeks-to-months old). However, because this paradigm assessed only cue-based access to memories formed long before the intervention, it remains unclear how retrieval practice affects autobiographical memory stabilization, fidelity, and transformation.

Retrieval practice may transform real-world memories by co-activating semantically related content and integrating reviewed information into broader semantic networks, representational changes that could stabilize new memories and drive the testing effect. Retrieval, more so than restudy, preferentially co-activates contextually related memories (Jonker et al., 2018), potentially facilitating integration of transient, hippocampus-dependent episodic memories with stable neocortical representations (Antony et al., 2017; Ritvo et al., 2019). For example, recalling an episode (e.g., a historic building tour) may activate semantically related knowledge about its content (e.g., details about the era when it was built), embedding the episode within neocortical semantic systems. This renders fleeting episodic content more robust by anchoring it to a stable neocortical trace.

These mnemonic transformations need not be limited to relationships with semantic content (Ritvo et al., 2019); relative to restudy, retrieval practice may preferentially co-activate related but unreviewed episodic content, altering and integrating it with explicitly reviewed memories. Importantly, the strength of co-activation may determine its modification (Ritvo et al., 2019): strong co-activation may strengthen and integrate unreviewed content with explicitly reviewed memories, weak co-activation may leave it weakened and differentiated from explicitly reviewed content, and minimal co-activation may leave it unchanged. This mechanism may explain why unreviewed details are sometimes strengthened, weakened, or unchanged following retrieval practice (Anderson, 2003; Chan et al., 2006; Jonker et al., 2018; Liu & Ranganath, 2021; Pan & Rickard, 2018; Pickering et al., 2021; Rowland & DeLosh, 2014; Wimber et al., 2015). For example, explicitly reviewing some episodic content (e.g., audio-guide building tour details) may strongly co-activate related but unreviewed content (e.g., the building's appearance), strengthening and integrating

unreviewed content with explicitly reviewed content, forming richer episodic memory.

Counterintuitively, retrieval practice may not always enhance memory accuracy, especially for narratively coherent real-world memories. First, retrieval is constructive: Rather than accessing a veridical record, we reassemble episodic fragments under schemas, making memories vulnerable to distortion (Reagh & Ranganath, 2023; Spens & Burgess, 2024). Second, co-activation of thematically related content may paradoxically introduce inaccuracies if these elements become integrated. For example, recalling one tour stop may co-activate a thematically related tour stop from 30 min later, linking the two. This can cause temporal disorganization (e.g., time-separated events are misremembered as consecutive) or content errors (e.g., details from different stops are recombined erroneously). Consistent with this, retrieval increases false recall of semantically related but nonpresented story elements (Denis et al., 2024), and retrieval errors can bias later memory (Bridge & Paller, 2012; Roediger & Marsh, 2005). These nuanced dynamics highlight the importance of studying retrieval effects in rich, real-world contexts, as simpler, decontextualized, and less semantically rich stimuli may limit co-activation and representational change.

Here, to investigate how retrieval uniquely shapes memory across multiple dimensions, beyond mere reexposure, we examined how retrieval practice versus restudy modifies complex, real-world memories. Participants experienced a staged historical building tour and then completed three tour review sessions across several days. In the *retrieve group*, participants answered cued-recall questions about tour details without feedback. In the *restudy group*, participants passively reviewed the same details by viewing the question and typing the simultaneously provided answers. In the final memory assessment (on Day 14, 8 days after the last review session), we employed multifaceted measures to capture retrieval-driven memory dynamics beyond what is possible with laboratory-based stimuli. First, we explored whether retrieval promotes event integration with preexisting semantic knowledge, which may occur via co-activation of semantically related information during review. We predicted the retrieve group would incorporate more preexisting semantic knowledge into event narratives, reflecting enhanced episode-semantic linkage. Our hypothesis is not that retrieval strengthens this semantic knowledge itself, but that it promotes integration of the episode with broader semantic networks. Second, we examined whether retrieval influences broader episodic memory. Because retrieval may holistically co-activate contextually and thematically related nontarget episodic elements, we predicted the retrieve group would show richer memory for reviewed and unreviewed episodic content.

However, because retrieval-based reconstruction is prone to errors and distortions, it may also reduce memory accuracy. Moreover, because tour events were spatially segregated (likely reducing co-activation of temporally adjacent events), retrieval may promote co-activation of thematically related content, biasing memory toward thematic rather than chronological organization. Together, these changes may alter subjective memory phenomenology. Third, while our primary focus was on how retrieval transforms memories within the broader memory system, we also investigated whether the classic testing effect (i.e., strengthened memory for explicitly reviewed content) extends to real-world memories. We predicted the retrieve group would better retain successfully reviewed content (versus restudy) after the 8-day delay. Importantly, these predictions, considering parallel semantic and episodic transformations, align with theories emphasizing the dynamic nature of memory consolidation and reorganization (Gilboa & Moscovitch, 2021; Winocur & Moscovitch, 2011). Retrieval may simultaneously foster semantic connections with neocortically supported preexisting knowledge and enrich hippocampally supported event-specific episodic detail due to distinct retrieval-induced consolidation pressures acting across memory systems.

Methods

Participants

Participants were randomly assigned to retrieve (retrieve group; $n = 26$; 13 men, 13 women) or restudy tour details (restudy group; $n = 26$; 21 men, five women; Table 1 displays sample statistics). A between-subjects design was chosen to avoid spillover effects across conditions. Groups were matched on episodic memory ability using Survey of Autobiographical Memory scores acquired prior to the tour (Palombo et al., 2013). Participants were recruited from the local community through a mix of social media (e.g., Facebook, Reddit, Instagram) and snowball sampling methods.

Participants were excluded if they had uncorrected visual or auditory difficulties (to ensure they could properly experience the tour) or if they had any diagnosed

psychological disorders (which could influence memory). Participants who had previously visited the tour building were generally excluded to ensure that the tour components were episodic across all participants. However, two exceptions were made for participants who had visited the location more than 10 years prior and had no self-reported memory of the building (one participant in each group). Two participants (one participant in each group) reported having visited or seen the building between the tour and the final assessment. These participants were included in the final analyses, although the pattern of results remained unchanged regardless of their inclusion. Three participants were excluded before data scoring and statistical analyses due to not attempting to recall the tour during the final assessment ($n = 2$ in the retrieve group) and due to a different experimenter conducting the final assessment ($n = 1$ in the restudy group). For all other participants, the same researcher conducted both the tour and the final assessment session. We excluded the participant with a different experimenter because having a different experimenter would have substantially changed the context and narrative style of the assessment, introducing substantial variability unrelated to the experimental manipulation. Specifically, the participant would be recounting their experience to a completely different person they had never met before, which would likely substantially alter the style of their narrative and require them to provide more context. This decision was made prior to data scoring and analysis.

A power analysis was conducted using G*Power (Faul et al., 2009) to detect differences in the number of internal details produced in participant narratives between the restudy and retrieve groups. We aimed for 80% power in a two-tailed Poisson regression (to account for the discrete, count-based nature of the data), with a significance level of .05. We estimated the base rate at 139 internal details per narrative. Prior research indicated that younger adults produced an average of 92.45 internal details per narrative during a naturalistic tour of Baycrest Hospital (Diamond & Levine, 2020). We increased this value by 1.5, as our tour contained more content compared with the Baycrest tour, and we provided more extensive instructions to encourage participants to report more details in their event narratives. Given the lack of studies with similar manipulations and outcome measures, a predicted effect size was difficult

Table 1 Sample demographic statistics

Group	Age			Years of education			Survey of Autobiographical Memory raw episodic subscale score		
	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range
Restudy	25.42	6.09	18–40	15.50	2.58	12–22	29.31	6.62	16–40
Retrieve	26.27	5.47	19–38	16.96	2.34	12–22	28.46	5.26	16–38

to estimate. Nevertheless, based on findings of very large testing effects observed in studies where initial test performance exceeded 75% or feedback was given (mean weighted Hedges's g from a meta-analysis = 0.97; Rowland, 2014), and a study showing substantial retrieval-induced facilitation with a similar manipulation (10% difference between retrieval and restudy groups, $d = .69$; Chan et al., 2006) we powered the analysis for a moderate rate ratio of 1.1 (corresponding to a 10% increase in internal details from restudy to retrieve groups). In summary, based on a two-tailed Poisson regression with 80% power, a significance level of .05, a rate ratio of 1.1, and a base rate of 139, the required sample size was determined to be 24 individuals per group, meaning a minimum of 48 participants was needed.

The tour staged event

Participants completed a 25-min, 13-stop interactive walking tour of a historic building, designed by the first author to procure an event memory that was immersive, episodically rich, and meaningful. Participants experienced the tour one-on-one with the experimenter to ensure comparability across participants and proper attention to all tour components. The experience combined audio-guided and researcher-guided components (Fig. 1B). At each stop, participants listened to descriptive audio guides while the extensively trained experimenter facilitated the experience by asking self-referential questions, presenting props, guiding the participant through activities, stating additional facts, and performing staged actions (full experimenter script, audio guides, and testing materials available on OSF: https://osf.io/t6kc3/overview?view_only=c031116695c44853a4ec7aab9d563636). Participants were unaware of what the Day 14 final assessment session would entail at the time of the tour and were instructed to treat the experience as a typical walking tour.

Tour review sessions

Participants reviewed (retrieved or restudied) verifiable tour details across three review sessions (Fig. 1A, 1C): immediately post tour (in-lab), 24 hours post tour (online), and five days post tour (online). This expanding interval schedule was designed to balance retrieval effort and success, as expanding schedules show superiority following low-level initial training, which may characterize event details from one-shot experiences (Toppino et al., 2018). Details regarding review session timing compliance are in the supplement ('Review Session Timing Compliance'). Review sessions were constructed to probe all main facets of each tour component, while ensuring coverage of distinct categories of information, including participant and researcher actions, audio guide content, spatial location, and perceptual details.

Retrieve group

The retrieve group answered 39 short-answer questions (Table S1) probing verifiable tour details without feedback. We chose a no-feedback retrieval procedure for two reasons. First, it better reflects the ecological conditions under which retrieval typically occurs for real-world events, where retrieval would not be routinely followed by corrective feedback. Second, it allowed us to isolate the effects of retrieval itself, without introducing additional influences associated with feedback, such as prediction error or an additional opportunity to restudy the event. Guessing was encouraged to promote engagement with the questions. Multiple questions per tour stop covered different aspects of that tour stop (e.g., tour audio content, perceptual details, researcher/participant actions, spatial information). Participants responded to the exact same questions at each review session. Most questions were open response, except for those requiring participants to indicate a location, which presented acceptable response options in a multiple-choice format to constrain answers, although participants were still required to type their answer. To promote context reinstatement, questions were grouped by tour stop and presented as blocks in the order of tour administration, with questions for one stop appearing on the same page (this organizational structure was not explicitly disclosed to participants). Questions spanning the tour (e.g., about the researcher) were also presented as a block. Each question was allotted 25 s, with an additional 5 s for pages with headers (pages advanced automatically). Questions were scored manually as correct or incorrect in line with an answer key, which was a document that specified acceptable responses for each item. Only answers that unambiguously referred to the correct content were accepted. For example, for the question "What scene did the stained glass window depict?", acceptable answers could include "wildlife," "Canadian wildlife," or "animals in the sunshine."

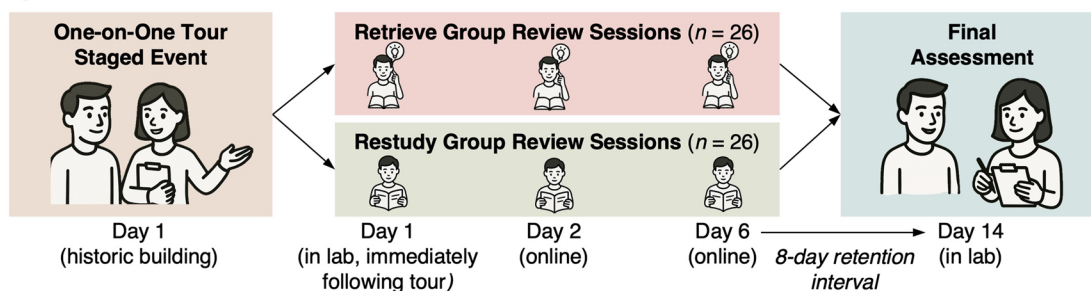
Restudy group

The restudy group followed the same procedure as the retrieve group but viewed questions alongside simultaneously provided correct answers and copied them into a text box (Fig. 1C). Performance was nearly perfect (see supplement 'Restudy Performance').

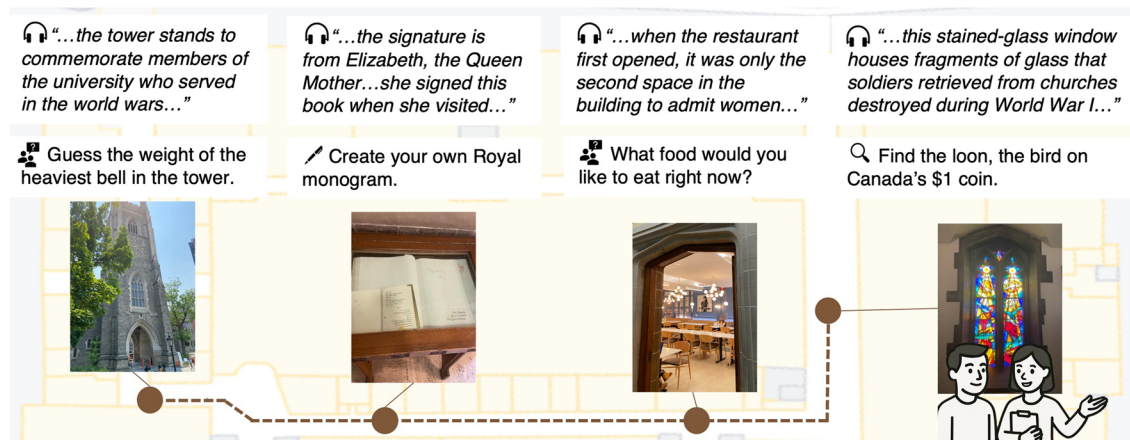
Tour review session phenomenology and metacognition questions

After each review session, participants answered three questions. First, self-reported *episodic reexperiencing* ("To what extent did you feel like you were reliving the tour when completing this review session?"; 1 = *extremely small extent*; 7 = *extremely large extent*); second, perceived review session

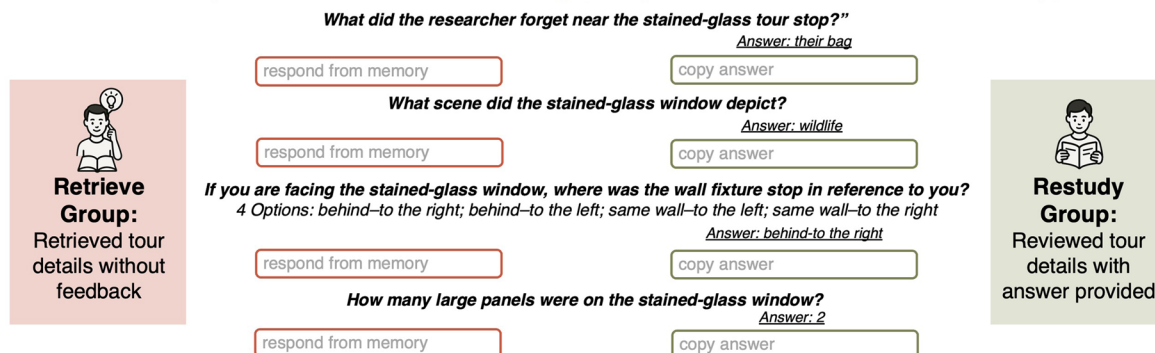
A Experimental Overview



B One-on-One Tour Staged Event: Sample Stops, Audio Guide Fragments, and Activities



C Critical Manipulation: Retrieve vs. Restudy (Sample Questions from One Tour Stop)



D Key Final Assessment Components

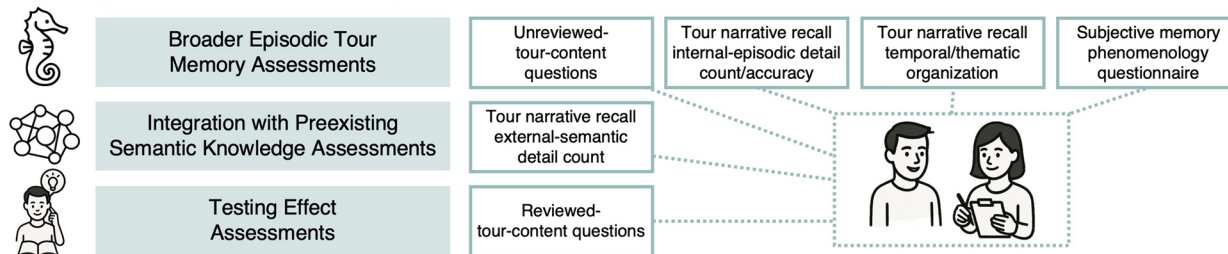


Fig. 1 Study methodology and procedure overview. **A** Our key question was how the type of review session (retrieval vs. restudy) influenced staged-event memory after an 8-day delay. **B** Participants experienced a one-on-one, interactive, factually detailed researcher-guided and audio-guided tour. **C** During tour review sessions, the retrieve group answered questions from memory without feedback, while the restudy group copied simultaneously provided answers into a text box. Questions were designed such that review sessions would

probe distinct types of details, such as events, researcher actions, audio guide details, spatial location, and perceptual details. Questions that asked participants to recall a spatial location included predefined response options to constrain possible answers. **D** The study had three core aims: examining the integration of the event with semantic knowledge structures, broader episodic tour memory, and the testing effect—each assessed with distinct measures. (Color figure online)

efficacy ("How effective do you think this review session was for improving your memory for the tour?"; 1 = *extremely ineffective*; 7 = *extremely effective*), and third, *anticipated future tour memory* ("Compared to a typical memory, how well do you think you will remember the tour?"; 1 = *much worse*; 7 = *much better*).

Final assessment administration

Eight days after the final review session (two weeks post tour), participants completed multiple memory assessments in-lab (Fig. 1D). Tour narratives involved verbally recalling all they remembered from the tour in as much detail as possible (prompts were adapted from the free recall and general probe sections of the Autobiographical Interview; Levine et al., 2002). Narratives were recorded, transcribed, and scored by research assistants using a manual developed by the first author, adapted from prior work (Diamond & Levine, 2020; Levine et al., 2002; supplement provides details for scorer training 'Training on Tour Narrative Scoring Protocol' and assessment administration 'Final Assessment'). To facilitate context reinstatement and reduce variability in recall styles across participants, the same experimenter always conducted both the tour and the final assessment. This ensured that all participants were recalling to a co-witness, rather than some participants experiencing recall without one.

Integration with preexisting semantic knowledge assessments

External-semantic details in tour narratives

Details were classified as *external-semantic* if they reflected factual knowledge not learned during the tour, such as general knowledge (e.g., "Gothic architecture originated in France") or self knowledge (e.g., "I enjoy Gothic architecture"; Fig. 2A–B). In contrast, tour-learned facts (e.g., "the stone carving in the tour building was designed with a Gothic influence") were considered *internal-episodic*, as they were presumed to be sufficiently novel and event-specific.

Broader episodic tour memory assessments

Unreviewed-tour-content questions

Participants answered novel questions probing verifiable tour aspects not covered during review sessions (Table S2), using the same presentation and scoring procedures as for reviewed-tour-content questions.

Internal-episodic details in tour narratives

Internal-episodic details were those that were tour-related, including learned factual information, perceptual/sensory impressions, actions taken/observed, spatiotemporal details, and participants' thoughts/feelings. Details were further categorized as verifiable (objectively confirmable) or unverifiable. Verifiable details were scored as correct or incorrect (Fig. 2A–B). A detailed tour runsheet was completed by the experimenter in real time, such that all recorded details (e.g., participants' responses to questions and idiosyncratic events) were treated as verifiable and scored as correct or incorrect.

Temporal and thematic organization in tour narratives

Recall order of broader tour events in narratives was scored (Fig. 2C), counting each event as a single recall, irrespective of detail provided. There were 18 events scored, the majority being the 13 tour stops (e.g., the stained-glass stop), but other staged events/researcher actions were also included (e.g., when participants were instructed to check the time on the audio device, when the researcher handed participants their binder to hold). Any detail provided that pertained to an event was scored as a recall of that event (e.g., the location, the audio guide, etc.).

Temporal clustering scores (lag-rank analysis) measured the tendency to recall events close in time and space, with 0.5 indicating chance and higher scores reflecting stronger clustering. For temporal clustering scores, for each recall event, we calculated the absolute lag of all remaining tour events and determined their percentile ranks. The recalled item's rank was then scaled from 0 (*most distant event*) to 1 (*nearest event*) and averaged across transitions.

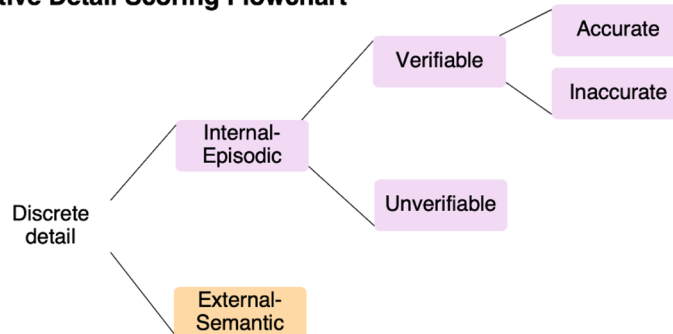
Thematic clustering scores (semantic lag-rank analysis) assessed recall organization by semantic relatedness, with scores above 0.5 indicating preferential recall of semantically related stops. For thematic clustering scores, for each recall event, we calculated the semantic distance of all remaining tour stops and determined their percentile ranks. The recalled item's rank was then scaled from 0 (*most semantically distant event*) to 1 (*most semantically related event*) and averaged across transitions. Semantic similarity of events was quantified using cosine similarities between vector embeddings of transcribed audio guides. Vector embeddings represent text as high-dimensional numerical vectors that capture semantic meaning, such that semantically similar texts have vectors that point in similar directions. Cosine similarity measures the angle between these vectors, ranging from 0 (*semantically unrelated*) to 1 (*semantically identical*). We computed embeddings using the Universal Sentence Encoder (Cer et al., 2018; Fig. 2D).

A Example of Tour Narrative Detail Scoring

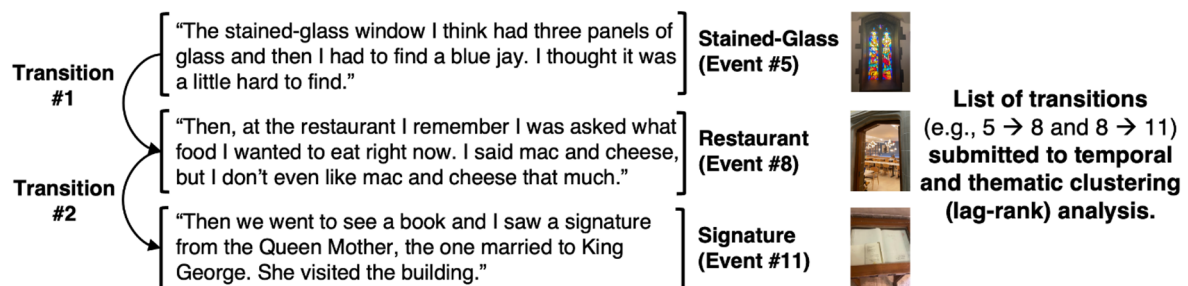


Internal-episodic, verifiable—accurate Internal-episodic, verifiable—inaccurate Internal-episodic, verifiable—inaccurate
 “The stained-glass window I think had three panels of glass and then I had to find a blue jay. I
 Internal-episodic, unverifiable Internal-episodic, verifiable—accurate Internal-episodic, verifiable—accurate
thought it was a little hard to find. Then, at the restaurant I remember I was asked what food I
 Internal-episodic, verifiable—accurate External-semantic
wanted to eat right now. I said mac and cheese, but I don’t even like mac and cheese that much.
 Internal-episodic, verifiable—accurate Internal-episodic, verifiable—accurate External-semantic
 Then we went to see a book and I saw a signature from The Queen Mother, the one married to
 Internal-episodic, verifiable—accurate
King George. She visited the building.

B Tour Narrative Detail Scoring Flowchart



C Example of Tour Narrative Temporal and Thematic Organization Scoring



D Quantifying Thematic Similarity Between Tour Stops

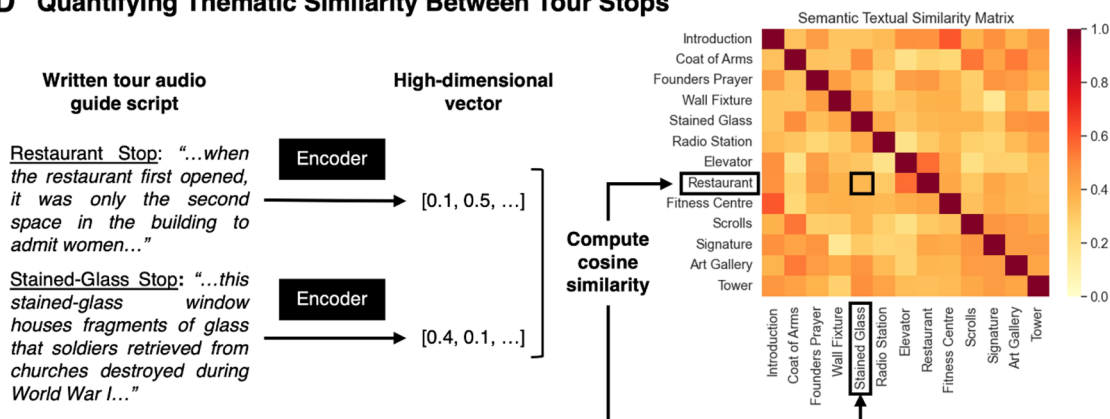


Fig. 2 Tour narrative scoring procedures. **A, B** Tour narrative details were classified as internal-episodic (tour-specific; verifiable/unverifiable and correct/incorrect) or external-semantic (facts not learned during the tour). Internal-episodic details were further categorized as verifiable or unverifiable, and verifiable details were additionally coded as accurate or inaccurate. This tour narrative is fictitious and was created solely for demonstration purposes, though it was inspired

by the types of responses participants provided. **C** Narratives were scored for event order, and transitions were analyzed for temporal and thematic clustering. **D** Thematic (semantic) similarity between tour stops was computed using cosine similarity of high-dimensional textual embeddings from the tour audio guides. (Color figure online)

Subjective memory phenomenology

The Memory Experiences Questionnaire–Short Form (MEQ-SF; excluding irrelevant sharing and distancing subscales; Luchetti & Sutin, 2016) assessed subjective memory phenomenology.

The testing effect assessments

Reviewed-tour-content questions

Participants answered the same questions as in the tour review sessions (Table S1), grouped by tour stop (e.g., stained glass stop) or peripheral content (e.g., questions about the researcher) and presented as blocks (e.g., all questions about the researcher were presented together, sequentially). Unlike in the review sessions, one question appeared per page, question block order was randomized, and response time was unlimited. Responses were scored manually as correct or incorrect in line with an answer key, which specified acceptable responses for each item. This measure examined how retrieval practice of event-specific details affected retention of those details, differing from past work investigating how retrieving an event from a cue word created in-lab influenced later access to that event through the identical cue word (Emmerdinger & Kuhbandner, 2018).

Statistical analyses

Generalized-mixed, linear-mixed, and general models were chosen to best fit the data structure, with random intercepts and slopes included as appropriate. For all models in which group was treatment coded, retrieve was coded as 0 and restudy was coded as 1. Key claims of no effect were supported by Bayes factors comparing models with and without the effect in question (BF_{01} ; more details pertaining to model selection, priors, functions, and packages are in the Supplement ‘Statistical Analyses’. Tour narrative detail counts considered free recall and general probe sections, while temporal and semantic organization analyses used only free recall to avoid influences from the probe.

Data and material availability and artificial intelligence disclosures

Materials, most primary data, and analysis scripts are publicly available on OSF (https://osf.io/t6kc3/overview?view_only=8b8fc90c2f8a4cb6af5671e1ddaabb78). The tour narrative interview data and tour runsheet data cannot be shared because it contains identifying information which

cannot be removed; however, it is available upon request to the corresponding author.

ChatGPT was used to create line drawings of people for figures, provide minor grammatical suggestions for the manuscript text, and troubleshoot analysis code.

Results

Review session phenomenology and metacognition questions

Cumulative link mixed models predicted response profiles (group fixed effect, treatment coded; session fixed effect, effect coded; group $\times$ session interaction; random intercept for participant). Type III Wald chi-square tests indicated the restudy group reported greater episodic reexperiencing, $\chi^2(1) = 4.47$, $p = .034$, and perceived review session effectiveness, $\chi^2(1) = 6.92$, $p = .009$, than the retrieve group (Fig. S1–S2), with no difference in anticipated future tour memory, $\chi^2(1) = 0.60$, $p = .440$, $BF_{01} = 10.58$ (Fig. S3). No other main effects/interactions were significant (p values $> .069$).

Integration with preexisting semantic knowledge

External-semantic details in tour narratives

A negative binomial model predicted external-semantic detail counts in final assessment tour narratives (group fixed effect, treatment coded). The retrieve group produced ~2.3 times more external-semantic details, $b = -0.84$, 95% CI $[-1.66, -0.03]$, $SE = 0.41$, $IRR = 0.43$, $z = -2.04$, $p = .041$, indicating enhanced event integration with preexisting knowledge (Fig. 3A; Fig. S4 displays counts by subtype; i.e., personal vs. general knowledge).

Spearman correlations showed a positive association between external-semantic and internal-episodic recall in the retrieve group, $\rho = 0.40$, 95% CI $[0.02, 0.68]$, $p = .041$, but not in the restudy group, $\rho < 0.01$, 95% CI $[-0.39, 0.39]$, $p = .997$, $BF_{01} = 3.22$ (Fig. 3B).

Broader episodic tour memory

Unreviewed-tour-content questions

A binomial mixed model predicting final assessment unreviewed-tour-content question accuracy (group fixed effect, treatment coded; random intercepts for question and participant; random slope for group within question; adjusted ICC = 0.398) showed no group effect, $b = 0.05$, 95% CI $[-0.21, 0.28]$, $SE = 0.13$, $OR = 1.05$, $z = 0.36$, $p = .718$,

Integration with Preexisting Semantic Knowledge Results



Tour narrative recall external-semantic details (e.g., “I saw a signature from *The Queen Mother*, the one married to King George”)

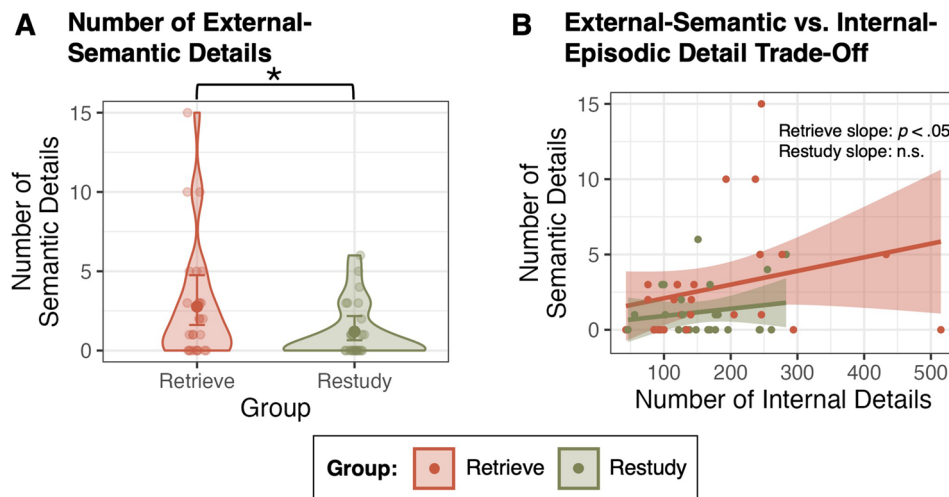


Fig. 3 Integration with preexisting semantic knowledge results. The retrieve group included more external-semantic details (i.e., non-tour-acquired personal/general facts) in tour narratives at the final assessment. In the retrieve group, more external-semantic details corresponded with more internal-episodic details. **A** Number of external-semantic details in tour narratives (large dots = model estimates with

95% confidence intervals; small dots = participant scores). **B** The relationship between internal-episodic and external-semantic detail counts (lines = slope estimates; small dots = participant scores; ribbons = 95% confidence intervals). * $p < .05$. (Color figure online)

$BF_{01} = 4.93$, indicating no strengthening of nontarget episodic details (Fig. 4A).

Internal-episodic details in tour narratives

A negative binomial regression predicting internal-episodic detail count in final assessment narratives (group fixed effect, treatment coded) found no group effect, $b = -0.11$, 95% CI $[-0.38, 0.17]$, $SE = 0.14$, $IRR = 0.90$, $z = -0.75$, $p = .452$, $BF_{01} = 3.66$, indicating retrieval did not enrich episodic tour recall at the group level (Fig. 4B; Figs. S6–S8 display internal-episodic detail counts by subtype—e.g., event, perceptual, thought/emotion, place, time—verifiability, and whether they were reviewed in review sessions).

A binomial mixed model predicting internal-episodic detail accuracy (group fixed effect, treatment coded; random intercept for participant; adjusted ICC = 0.071) found no group effect, $b = 0.12$, 95% CI $[-0.21, 0.45]$, $SE = 0.17$, $OR = 1.13$, $z = 0.71$, $p = .479$, $BF_{01} = 3.26$, indicating retrieval did not modulate episode accuracy (Fig. 4C; Fig. S9 displays errors by subtype; i.e., confabulations, intrusions across events or source memory errors, sequence errors, and tour content errors).

Temporal and thematic organization in tour narratives

A beta regression predicting temporal clustering values in tour narratives (group fixed effect; retrieve = -1 , restudy = 1) revealed a significant intercept, $b = 1.42$, 95% CI $[1.18, 1.67]$, $SE = 0.12$, $z = 11.47$, $p < .001$, but no group effect, $b = -0.13$, 95% CI $[-0.35, 0.10]$, $SE = 0.11$, $z = -1.12$, $p = .264$, $BF_{01} = 3.39$, indicating participants clustered narratives temporally (clustering was above chance at ~ 0.81 across groups) and retrieval did not induce temporal disorganization (Fig. 4D; findings corroborated by lag-conditional response probability curves and explicit temporal sequencing task, Figs. S10–S11).

A beta regression model predicting thematic clustering values in tour narratives (group fixed effect, effect coded: retrieve = -1 , restudy = 1) revealed a non-significant intercept, $b = 0.10$, 95% CI $[-0.04, 0.25]$, $SE = 0.07$, $z = 1.39$, $p = .165$, and no group effect, $b = -0.07$, 95% CI $[-0.22, 0.07]$, $SE = 0.07$, $z = -1.01$, $p = .312$, $BF_{01} = 5.28$, indicating participants did not cluster narratives thematically (clustering was not above chance), and retrieval did not promote thematic clustering (Fig. 4E).

Broader Episodic Tour Memory Results



Unreviewed-tour-content questions (i.e., novel questions probing verifiable tour details; e.g., “What picture was on the researcher’s tote bag”)

Tour narrative recall internal-episodic detail count and accuracy (e.g., “*I saw a signature from The Queen Mother, the one married to King George*”)

Tour narrative recall temporal and thematic organization

Subjective memory phenomenology

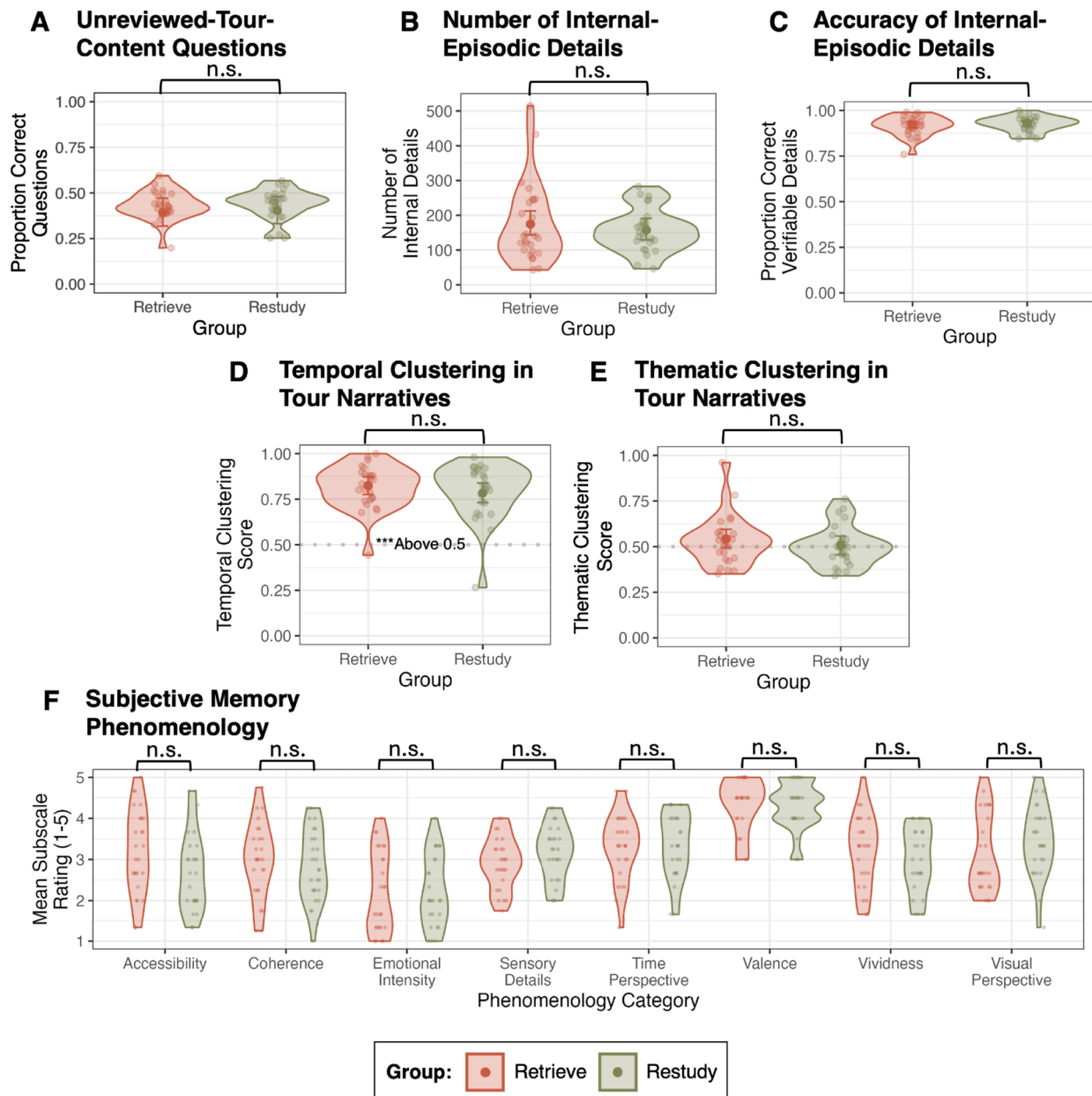


Fig. 4 Broader episodic tour memory results. Overall, there was no significant difference in broader episodic tour memory between the retrieve and restudy groups at final assessment (large dots = model estimates with 95% confidence intervals; small dots = participant scores). **A** Proportion of unreviewed-tour-content questions answered correctly. Table S2 displays performance by question. **B** Number of internal-episodic details in tour narratives. **C** Proportion of internal-

episodic, verifiable details that were accurate (vs. inaccurate). **D–E** Temporal and thematic lag rank values, with chance clustering corresponding to 0.5 (dotted line). Temporal (but not thematic) lag rank values were significantly above 0.5 for both groups. **F** Average Memory Experiences Questionnaire scores for each subscale. *** $p < .001$. (Color figure online)

Subjective memory phenomenology

Linear regressions (group fixed effect, treatment coded) predicting mean MEQ-SF subscale scores for accessibility, coherence, emotional intensity, sensory details, time perspective, valence, vividness, and visual perspective found no significant group effects (p values $\geq .083$; $BF_{01} \geq 1.76$, except accessibility, $BF_{01} = 0.89$; see Table S3 for model results). Thus, retrieval did not influence final assessment subjective memory phenomenology (Fig. 4F).

The testing effect

Reviewed-tour-content questions

During review sessions, the retrieve group received no feedback and reviewed only recallable details (Fig. 5A shows retrieve-group performance during review sessions; see the supplement, “Retrieve-Group Performance During Tour Review Sessions,” for more details on retrieve-group performance across sessions; Fig. S12 shows performance by question), whereas the restudy group reviewed all details three times, creating a mismatch in reviewed information. Because retrieval effects depend on initial success, we conditionalized final performance on each participant’s review success (Rowland & DeLosh, 2015), analyzing only items recalled in all three review sessions. This ensured all analyzed items had been reviewed equally and addressed our key question: whether successfully retrieved information was better remembered than restudied information after an eight-day delay. A binomial mixed model (group fixed effect, random intercepts for question and participant, random slopes for group within question; adjusted ICC = 0.443) predicted correct responses to reviewed-tour-content questions. The retrieve group was ~3.6 times more likely to answer correctly, $b = -1.29$, $OR = 0.28$, 95% CI $[-2.17, -0.41]$, $SE = 0.45$, $z = -2.88$, $p = .004$ (Fig. 5B), with the largest advantage for questions with lower baseline success rates at the final test (i.e., harder questions; question intercepts x slopes $r = -0.94$). To test whether the retrieve group’s advantage merely reflected inclusion of generally easier items, we compared conditionalized retrieve-participant scores to restudy-group performance on the same items (Fig. S13). The proportion of retrieve participants scoring above the restudy-matched median (65%) did not significantly exceed 50% ($p = .084$). An unconditionalized model with the same structure but including all items for all participants (ICC = 0.437) revealed a nonsignificant numerical trend favouring restudy, $b = 0.49$, 95% CI $[-0.14, 1.13]$, $SE = 0.32$, $z = 1.53$, $p = .126$, $BF_{01} = 0.87$ (Fig. 5C).

Discussion

The present study leveraged real-world memories’ multidimensionality to investigate how retrieval practice shapes and reorganizes memory, revealing nuanced changes across semantic and episodic dimensions. Participants experienced a staged event, completed three retrieval or restudy sessions across 5 days, and underwent a final memory test 8 days after the final review session. Three key findings emerged. First, retrieval practice enhanced preexisting semantic knowledge incorporation into event narratives, with the degree of incorporation predicting episodic narrative richness after retrieval. Second, retrieval practice did not affect cued recall for unreviewed content, narrative accuracy, episodic richness, temporal or thematic organization, or subjective phenomenology. Third, to our knowledge, we provide the first demonstration of the classic testing effect for real-world event memory: Event details successfully retrieved during review were remembered better after an 8-day delay than those merely restudied (though as discussed below, this pattern warrants cautious interpretation). Overall, by assessing retrieval’s impact on real-world memories using nuanced measures capturing mnemonic changes across multiple dimensions, our study reveals that retrieval shapes long-term memory by integrating episodes with semantic networks, enhancing episodic richness.

First, we show retrieval practice promotes semantic integration of episodes: Participants incorporated more preexisting (i.e., not acquired during the tour) semantic information into recall narratives following retrieval than restudy. Interestingly, higher external-semantic counts in the retrieve group were associated with higher internal-episodic counts, suggesting semantic content scaffolds rather than compensates for episodic memory. This supports theories suggesting retrieval, like sleep, embeds initially hippocampus-dependent episodic memories into neocortical semantic networks (Antony et al., 2017) and aligns with elaborative retrieval accounts, where activation spreads to related concepts during retrieval, creating multiple semantic pathways to the target (Carpenter, 2009). Our findings also align with evidence that retrieval improves retention of conceptual object features (Lifanov et al., 2021) and increases neural similarity among semantically related objects (Ferreira et al., 2019). These findings underscore the need for theories of memory consolidation to account for its multidimensional nature—where episodic and semantic representations dynamically coexist, are expressed concurrently, and evolve independently, rather than one supplanting the other (Gilboa & Moscovitch, 2021; Winocur & Moscovitch, 2011). The

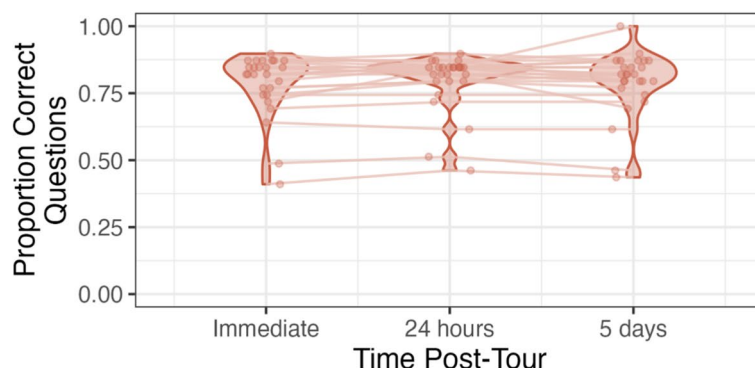
Testing Effect Results



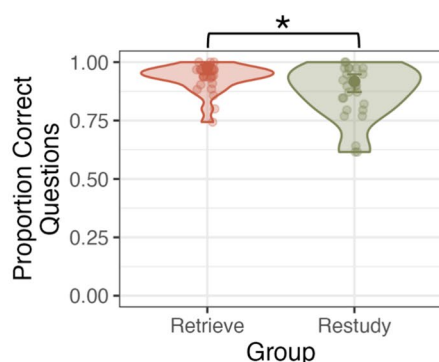
Reviewed-tour-content questions

(i.e., questions repeated from the review sessions probing verifiable tour details)
(e.g., “What did the researcher forget near the stained-glass tour stop?”)

A Raw Review Session Performance (Retrieve Group Only; Reviewed-Tour-Content Questions)



B Conditionalized Analysis: Successfully-Reviewed Tour Content Retained from Third Review Session to Final Assessment



C Unconditionalized Analysis: Total Tour Content Retained from Third Review Session to Final Assessment

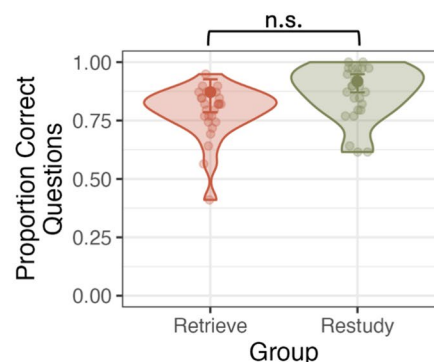


Fig. 5 The testing effect results. **A** Retrieve-group performance during the review sessions was generally consistent across all three sessions (small dots = participant scores). **B** Conditionalized analysis: The retrieve group outperformed the restudy group in retaining successfully reviewed tour content from the final review session to the final assessment (8-day delay). To equate groups on item review status, only reviewed-tour-content questions that were answered correctly in all three review sessions were included (large dots = model

estimates with 95% confidence intervals; small dots = participant scores). **C** Unconditionalized analysis: the restudy group showed numerically higher retention compared to the retrieve group for Retention of all reviewed-tour-content questions from the final review session to the final assessment (8-day delay), though this difference was not significant. All questions are included here (large dots = model estimates with 95% confidence intervals; small dots = participant scores). * $p < .05$. (Color figure online)

observed semantic integration may arise from co-activation of semantically related nontarget memories during retrieval, integrating episodes with semantic knowledge structures (Antony et al., 2017), and future work should examine how co-activation during retrieval versus restudy drives the representational changes that shape memory. Notably, some participants produced many external-semantic details, while several produced none, indicating the pattern we observed was not universal. Contextual factors (e.g., episodic-focused

recall instructions and recounting to a co-witness familiar with the event) likely discouraged the production of general or explanatory semantic information, thus reducing external-semantic detail production and potentially underestimating semantic integration. Future work using open-ended instructions and nonshared experiences may reveal stronger effects.

Second, unexpectedly, retrieval practice did not preferentially alter memory for unreviewed-tour-content questions, narrative accuracy, temporal and thematic

organization, or subjective phenomenology. Importantly, comparable group performance does not imply retrieval fails to alter broader episodic memory, but rather that it may not induce changes beyond those achieved by reexposure. This aligns with prior work showing retrieval and reexposure support comparable transfer of learning (Pan & Rickard, 2018; Pickering et al., 2021), and integration of more-naturalistic episodic memories (Homann & Barense, 2026). Perhaps both retrieval and restudy elicited comparable co-activation of nontarget episodic content, yielding similar representational outcomes. Notably, we observed higher re-experiencing phenomenological ratings following the restudy than retrieval review sessions. Although restudy is generally considered more passive and less effortful than retrieval practice (Bjork, 1994), the rich interconnections of real-world memories may enable spontaneous elaboration during restudy, potentially attenuating the typical retrieval advantage and explaining the null episodic findings. Alternatively, differing co-activation levels may have produced similar outcomes in the retrieve and restudy groups due to the proposed U-shaped relationship with memory change (Ritvo et al., 2019). Lastly, because the audio-guide-derived similarity matrix, while well-suited to semantic content, may not fully capture other experiential dimensions of meaning that participants used to cluster narratives (e.g., atmosphere, spatial scale), the null thematic clustering result may reflect the limits of this index rather than a true absence of clustering.

Third, extending prior laboratory and educational findings, we found some evidence for a testing effect in a real-world context: Retrieval practice enhanced retention of real-world event details, so long as they were successfully retrieved during retrieval practice, more than restudy. The testing effect persisted across diverse episodic details, suggesting retrieval practice benefits generalize and may serve as a scalable, nonpharmacological means to strengthen everyday memory. The testing effect may rely on integration, allowing retrieved details to draw on stable neocortical traces, especially important as hippocampal traces degrade (Antony et al., 2017). In an exploratory 12–18-month follow-up, we observed a small, nonsignificant numerical relationship between external-semantic detail counts on Day 14 and retention of reviewed-tour-content questions (Fig. S5). However, the small effect size (likely due to a restriction of range issue in the restudy group) precludes firm conclusions and adequately powered replication is needed.

This testing effect should, however, be interpreted cautiously. We conditionalized the analysis such that only successfully reviewed items were included for the retrieval group to avoid an artificial disadvantage from including retrieval failures. However, conditionalizing may select for well-encoded details that would have been retained regardless of retrieval practice. Four exposures per item in

the restudy group, along with one-on-one tour administration ensuring attention to all reviewed details, offer some reassurance against this concern, as even with this level of attention, restudy participants still forgot some details. Still, item selection remains a plausible alternative explanation the current data cannot rule out. We omitted corrective feedback, which could have offset low initial retrieval success, to avoid introducing additional mechanisms beyond direct retrieval-practice effects (e.g., prediction errors, reencoding) and to preserve ecological validity, as feedback is uncommon for the real-world experiences we wished to study. This, however, left retrieval success rate uncontrolled. Future work bolstering retrieval performance in an ecologically valid manner is needed to disentangle these effects and establish whether retrieval-practice benefits hold for real-world stimuli, especially given evidence that testing effects may not generalize well beyond simplified laboratory materials (Homann & Barense, 2026).

Conclusion

We harnessed the complexity of real-world event memories to reveal nuanced effects of retrieval practice versus restudy. Retrieval selectively promoted integration of episodes with preexisting semantic knowledge, and this integration supported rather than competed with episodic richness. These findings suggest retrieval shapes memory by embedding episodes into stable semantic networks—a process that may underlie the testing effect—and demonstrate that distinct memory systems can undergo parallel, complementary transformations during retrieval-based consolidation. Together, these findings highlight semantic integration as a potential mechanism through which retrieval stabilizes memories.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.3758/s13423-026-03003-5>.

Acknowledgements We thank Brian Levine and Bradley Buchsbaum for insightful discussions, and Jia Gu, Simran Grewal, Jessica Sun, Shelby Davies, Amir Samadi, and Janice An for assisting with data collection and scoring.

Funding This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC; Discovery Grant RGPIN-2020-05747 to M.D.B.), the James S. McDonnell Foundation (Scholar Award to M.D.B.), and the Canada Research Chairs Program (M.D.B.). L.A.H. was supported by NSERC Canada Graduate Scholarships at both the master's and doctoral levels.

Data availability Materials and most primary data are publicly available on OSF (https://osf.io/t6kc3/overview?view_only=c031116695c44853a4ec7aab9d563636). The tour narrative interview data cannot be shared because it contains identifying information which cannot

be removed; however, it is available upon reasonable request to the corresponding author.

Code availability Analysis scripts are publicly available on OSF (https://osf.io/t6kc3/overview?view_only=c031116695c44853a4ec7aab9d563636).

Declarations

Ethics approval This research was approved by the University of Toronto Office of Research Ethics (protocol number 39014) and was performed in accordance with the ethical standards as laid down in the 1964 Declaration of Helsinki.

Consent Informed consent was obtained from all individual participants included in the study.

Consent for publication Not applicable.

Conflicts of interest The authors have no relevant financial or non-financial interests to disclose.

Open practices statement Materials, most primary data, and analysis scripts are publicly available on OSF (https://osf.io/t6kc3/overview?view_only=c031116695c44853a4ec7aab9d563636). The tour narrative interview data cannot be shared because it contains identifying information which cannot be removed; however, it is available upon reasonable request to the corresponding author. The experiment was not preregistered.

References

- Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language*, 49(4), 415–445. <https://doi.org/10.1016/j.jml.2003.08.006>
- Antony, J. W., Ferreira, C. S., Norman, K. A., & Wimber, M. (2017). Retrieval as a fast route to memory consolidation. *Trends in Cognitive Sciences*, 21(8), 573–576. <https://doi.org/10.1016/j.tics.2017.05.001>
- Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing*. MIT Press. <https://doi.org/10.7551/mitpress/4561.001.0001>
- Bridge, D. J., & Paller, K. A. (2012). Neural correlates of reactivation and retrieval-induced distortion. *Journal of Neuroscience*, 32(35), 12144–12151. <https://doi.org/10.1523/JNEUROSCI.1378-12.2012>
- Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(6), 1563–1569. <https://doi.org/10.1037/a0017021>
- Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. S., . . . Kurzweil, R. (2018). Universal Sentence Encoder. *ArXiv Preprint*. <https://doi.org/10.48550/arXiv.1803.11175>
- Chan, J. C. K., McDermott, K. B., & Roediger, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, 135(4), 553–571. <https://doi.org/10.1037/0096-3445.135.4.553>
- Denis, D., DiPietro, C., Spreng, R. N., Schacter, D. L., Stickgold, R., & Payne, J. D. (2024). Sleep and retrieval practice both strengthen and distort story recollection. *Sleep Advances*, 5(1), Article zpae083. <https://doi.org/10.1093/sleepadvances/zpae083>
- Diamond, N. B., & Levine, B. (2020). Linking detail to temporal structure in naturalistic-event recall. *Psychological Science*, 31(12), 1557–1572. <https://doi.org/10.1177/0956797620958651>
- Emmerdinger, K. J., & Kuhbandner, C. (2018). Testing memories of personally experienced events: The testing effect seems not to persist in autobiographical memory. *Frontiers in Psychology*, 9, Article 810. <https://doi.org/10.3389/fpsyg.2018.00810>
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41, 1149–1160.
- Ferreira, C. S., Charest, I., & Wimber, M. (2019). Retrieval aids the creation of a generalised memory trace and strengthens episode-unique information. *NeuroImage*, 201, Article 115996. <https://doi.org/10.1016/j.neuroimage.2019.07.009>
- Gilboa, A., & Moscovitch, M. (2021). No consolidation without representation: Correspondence between neural and psychological representations in recent and remote memory. *Neuron*, 109(14), 2239–2255. <https://doi.org/10.1016/j.neuron.2021.04.025>
- Homann, L. A., & Barense, M. D. (2026). Event memory is strengthened and integrated by review: Retrieval practice offers inconsistent additional benefit. *PsyArXiv Preprint*. https://doi.org/10.31234/osf.io/tgm3w_v1
- Jonker, T. R., Dimsdale-Zucker, H., Ritchey, M., Clarke, A., & Ranganath, C. (2018). Neural reactivation in parietal cortex enhances memory for episodically linked information. *Proceedings of the National Academy of Sciences of the United States of America*, 115(43), 11084–11089. <https://doi.org/10.1073/pnas.1800006115>
- Karpicke, J. D. (2017). Retrieval-based learning: A decade of progress. In J. H. Byrne (Ed.), *Learning and memory: A comprehensive reference* (2nd ed., pp. 487–514). Academic. <https://doi.org/10.1016/B978-0-12-809324-5.21055-9>
- Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, 65(2), 85–97. <https://doi.org/10.1016/j.jml.2011.04.002>
- Levine, B., Svoboda, E., Hay, J. F., Winocur, G., & Moscovitch, M. (2002). Aging and autobiographical memory: Dissociating episodic from semantic retrieval. *Psychology and Aging*, 17(4), 677–689. <https://doi.org/10.1037/0882-7974.17.4.677>
- Lifanov, J., Linde-Domingo, J., & Wimber, M. (2021). Feature-specific reaction times reveal a semanticisation of memories over time and with repeated remembering. *Nature Communications*, 12, 3177. <https://doi.org/10.1038/s41467-021-23288-5>
- Liu, X. L., & Ranganath, C. (2021). Resurrected memories: Sleep-dependent memory consolidation saves memories from competition induced by retrieval practice. *Psychonomic Bulletin & Review*, 28, 2035–2044. <https://doi.org/10.3758/s13423-021-01953-6>
- Luchetti, M., & Sutin, A. R. (2016). Measuring the phenomenology of autobiographical memory: A short form of the Memory Experiences Questionnaire. *Memory*, 24(5), 592–602. <https://doi.org/10.1080/09658211.2015.1031679>
- Palombo, D. J., Williams, L. J., Abdi, H., & Levine, B. (2013). The survey of autobiographical memory (SAM): A novel measure of trait mnemonics in everyday life. *Cortex*, 49(6), 1526–1540. <https://doi.org/10.1016/j.cortex.2012.08.023>

- Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, 144(7), 710–756. <https://doi.org/10.1037/bul0000151>
- Pickering, J. S., Henderson, L. M., & Horner, A. J. (2021). Retrieval practice transfer effects for multielement event triplets. *Royal Society Open Science*, 8(11), 1–19. <https://doi.org/10.1098/rsos.201456>
- Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, 330(6002), 335–335. <https://doi.org/10.1126/science.1191465>
- Reagh, Z. M., & Ranganath, C. (2023). Flexible reuse of cortico-hippocampal representations during encoding and recall of naturalistic events. *Nature Communications*, 14, Article 279. <https://doi.org/10.1038/s41467-023-36805-5>
- Ritvo, V. J. H., Turk-Browne, N. B., & Norman, K. A. (2019). Non-monotonic plasticity: How memory retrieval drives learning. *Trends in Cognitive Sciences*, 23(9), 726–742. <https://doi.org/10.1016/j.tics.2019.06.007>
- Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3), 249–255. <https://doi.org/10.1111/j.1467-9280.2006.01693.x>
- Roediger, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(5), 1155–1159. <https://doi.org/10.1037/0278-7393.31.5.1155>
- Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, 140(6), 1432–1463. <https://doi.org/10.1037/a0037559>
- Rowland, C. A., & DeLosh, E. L. (2014). Benefits of testing for non-tested information: Retrieval-induced facilitation of episodically bound material. *Psychonomic Bulletin & Review*, 21, 1516–1523. <https://doi.org/10.3758/s13423-014-0625-2>
- Rowland, C. A., & DeLosh, E. L. (2015). Mnemonic benefits of retrieval practice at short retention intervals. *Memory*, 23(3), 403–419. <https://doi.org/10.1080/09658211.2014.889710>
- Spens, E., & Burgess, N. (2024). A generative model of memory construction and consolidation. *Nature Human Behaviour*, 8(3), 526–543. <https://doi.org/10.1038/s41562-023-01799-z>
- Toppino, T. C., Phelan, H.-A., & Gerbier, E. (2018). Level of initial training moderates the effects of distributing practice over multiple days with expanding, contracting, and uniform schedules: Evidence for study-phase retrieval. *Memory & Cognition*, 46, 969–978. <https://doi.org/10.3758/s13421-018-0815-7>
- Wimber, M., Alink, A., Charest, I., Kriegeskorte, N., & Anderson, M. C. (2015). Retrieval induces adaptive forgetting of competing memories via cortical pattern suppression. *Nature Neuroscience*, 18, 582–589. <https://doi.org/10.1038/nn.3973>
- Winocur, G., & Moscovitch, M. (2011). Memory transformation and systems consolidation. *Journal of the International Neuropsychological Society*, 17(05), 766–780. <https://doi.org/10.1017/S1355617711000683>

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.