Putting the testing effect to the test in the wild: Retrieval enhances real-world memories and promotes their semantic integration while preserving episodic integrity

Lauren A. Homann[1], Mursal Jahed[1], and Morgan D. Barense[1,2]

[1] Department of Psychology, University of Toronto

[2] Rotman Research Institute, Baycrest

Author Note:

Correspondence concerning this article should be addressed to Lauren Homann, Department of Psychology, University of Toronto. Contact: lauren.homann@mail.utoronto.ca

Abstract

Retrieval practice—actively recalling information—is an established memory-strengthening technique. However, understanding how retrieval transforms memory requires examining its effects on memories that evolve across multiple episodic and semantic dimensions, as is typical of real-world events. Thus, we investigated how repeatedly *retrieving* event details without feedback versus *restudying* the same details influenced memory for an episodically-rich and meaningful staged event after 14 days (n = 26 per group). Retrieval enhanced retention of successfully-reviewed content, providing the first testing effect demonstration for real-world events. Retrieval also increased the incorporation of pre-existing semantic information into recall narratives, suggesting enhanced event integration with pre-existing knowledge, perhaps via co-activation of semantically-related content during retrieval. However, this semantic integration did not enhance—or impair—broader episodic memory beyond successfully-reviewed content. These findings suggest that retrieval reshapes memories by integrating recalled content into semantic knowledge networks—a mechanism that may underlie the testing effect—while preserving the overall integrity of episodic representations.

Research Transparency Statement

General Disclosures

Conflicts of interest: All authors declare no conflicts of interest. Funding: This research was supported by the Natural Sciences and Engineering Research Council of Canada (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. Artificial intelligence: ChatGPT created line drawings used in figures, made minor grammatical and phrasing suggestions for the manuscript text, and troubleshooted analysis code. Ethics: This research received approval from an ethics board at the University of Toronto (ID: 39014).

Study Disclosures

Preregistration: No aspects of the study were preregistered. Materials: All study materials are publicly available on OSF (https://osf.io/t6kc3/?view_only=8b8fc90c2f8a4cb6af5671e1ddafbb78). Data: Most primary data are publicly available on OSF (https://osf.io/t6kc3/?view_only=8b8fc90c2f8a4cb6af5671e1ddafbb78). The tour narrative interview data cannot be shared because it contains identifying information which cannot be removed; however, it is available upon request to the corresponding author. Analysis scripts: All analysis scripts are publicly available on OSF (https://osf.io/t6kc3/?view_only=8b8fc90c2f8a4cb6af5671e1ddafbb78).

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.

Putting the testing effect to the test in the wild: Retrieval enhances real-world memories and promotes their semantic integration while preserving episodic integrity

Retrieval is not merely a means of indexing stored knowledge, but a powerful memory reactivator that can reshape memory itself—a notion exemplified by a cornerstone finding in memory research: the *testing effect*. This is the phenomenon that *retrieval practice* (i.e., actively recalling information) improves memory for reviewed content more than *restudy* (i.e., reactivation via passive review; e.g., Roediger & Karpicke, 2006). As a memory-enhancing technique, retrieval practice has yielded some of the most robust and generalizable results in the field, improving retention across diverse materials and testing formats (see Karpicke, 2017 for a review). Retrieval, then, is uniquely powerful in enhancing memory.

Can retrieval do more than simply strengthen retrieved information? To answer this, we must look beyond basic testing effects and examine how retrieval shapes memories across multiple dimensions—something the simple lab-based stimuli traditionally studied cannot capture but real-world event memories embody. Yet, testing effect research has largely neglected retrieval's impact on real-world memories, which differ fundamentally from lab-based stimuli by encompassing a broader range of rich, interrelated sensory details and event-specific information embedded in spatial and temporal contexts, as well as intricately linked pre-existing semantic knowledge. This research gap is not merely a practical concern—given the frequent retrieval of real-world memories in daily life—but also a fundamental limitation in our understanding of retrieval-driven memory dynamics. Fully understanding retrieval's impact requires examining its influence not only on directly-reviewed episodic content, but also on unreviewed details, relationships within and across episodes, and semantic transformation of episodic information. Real-world event memories thus offer an ideal context for studying these complex, multidimensional effects simultaneously.

Retrieval practice may reinforce episodic content by promoting integration of information—both into broader semantic memory networks and among elements within the

episode itself. Unlike restudy, retrieval is thought to co-activate related semantic knowledge supported by the neocortex, promoting integration between more transient, hippocampus-dependent episodic memory and more stable, distributed neocortical representations (Antony et al., 2017). For example, recalling a dinner with a friend may co-activate semantically-related knowledge about that friend, embedding the event within neocortical systems by strengthening links between episodic details and semantic memory (Ritvo et al., 2019, 2024). This integration process has been proposed as a mechanism underlying the testing effect, as retrieval renders fleeting episodic content more robust by anchoring it to a more stable neocortical trace (Antony et al., 2017). In contrast, restudy primarily reactivates target content, limiting such representational change. Theoretically, representational change can result from co-activation of any mnemonic elements, including episodic content (Ritvo et al., 2019, 2024). For example, recalling a dinner conversation may also co-activate non-target contextual details, like the restaurant's appearance. If sufficiently reactivated, these non-target details may be strengthened and integrated with co-active conversation details, yielding a richer, more integrated episodic dinner memory. Compared to restudy, retrieval may more effectively promote such co-activation of non-target episodic details, increasing opportunities for representational change within the episodic memory itself.

Counterintuitively, retrieval practice may not always enhance memory accuracy—especially for narratively-coherent real-world memories—given the inherently reconstructive, error-prone nature of retrieval and its potential to induce representational changes that introduce distortion. First, retrieval is constructive: rather than accessing a complete, veridical record, we reassemble episodic fragments under the influence of schemas, making memories vulnerable to distortion (Reagh & Ranganath, 2023; Spens & Burgess, 2024). Retrieval practice might introduce and reinforce such distortions in both directly-retrieved content *and* co-activated information. Second, co-activation of thematically-related content may paradoxically lead to inaccuracies if these elements become integrated. For example, recalling

one conversation may co-activate a thematically-related conversation from 30 minutes later, leading to temporal disorganization if time-separated events are misremembered as consecutive, or content errors if conversation details are recombined erroneously. These nuanced dynamics underscore the importance of examining retrieval's effects on real-world memory accuracy, as they are difficult to capture with simpler, decontextualized, and less semantically-rich laboratory-based stimuli—conditions that likely constrain co-activation and representational change.

Here, to investigate how retrieval shapes memory across multiple dimensions, we explored how retrieval practice, beyond mere reexposure through restudy, modifies complex real-world memories. Participants experienced an immersive staged tour, then either practiced retrieval of tour details without feedback (*retrieve group*) or passively reviewed the same details (*restudy group*) across three review sessions over five days. Memory was assessed after an eight-day delay following the final review session (i.e., 14 days post-tour) using multifaceted measures permitting a more comprehensive investigation of retrieval-driven memory dynamics than is possible for laboratory-based stimuli. First, we investigated whether the testing effect extends to real-world memories, predicting better retention of successfully-reviewed content in the retrieve group (versus restudy) after the eight-day delay. Second, we explored whether retrieval promotes integration of the event with semantic knowledge structures, anticipating stronger links to pre-existing (i.e., acquired before the tour) factual knowledge through co-activation of semantically-related information during review. Third, we examined whether retrieval influences broader episodic memory. Because retrieval, unlike restudy, may holistically co-activate contextually- and thematically-related non-target episodic elements, we predicted the retrieve group would show strengthened memory for non-target episodic content overall. However, because reactivation is inherently error-prone, it may also lead to reduced accuracy. Moreover, given that tour events were spatially segregated (likely reducing co-activation of temporally adjacent episodes) retrieval may instead promote co-activation of

thematically-related content, leading memory structure to favor thematic over chronological organization. Together, these changes may contribute to shifts in subjective memory phenomenology.

Methods

Participants

Participants were randomly assigned to retrieve (retrieve group; *n* = 26; 13 male, 13 female) or restudy tour details (restudy group; *n* = 26; 21 male, 5 female; Table 1 displays sample statistics). A between-subjects design avoided spillover effects across conditions. Groups were matched on episodic memory ability using pre-tour Survey of Autobiographical Memory scores (Palombo et al., 2013). Power analysis indicated that 24 participants per group were sufficient for detecting predicted effects on internal-episodic detail count (details and exclusion criteria reported in supplement [‘Participants’]). This research was approved by a university ethics board.

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

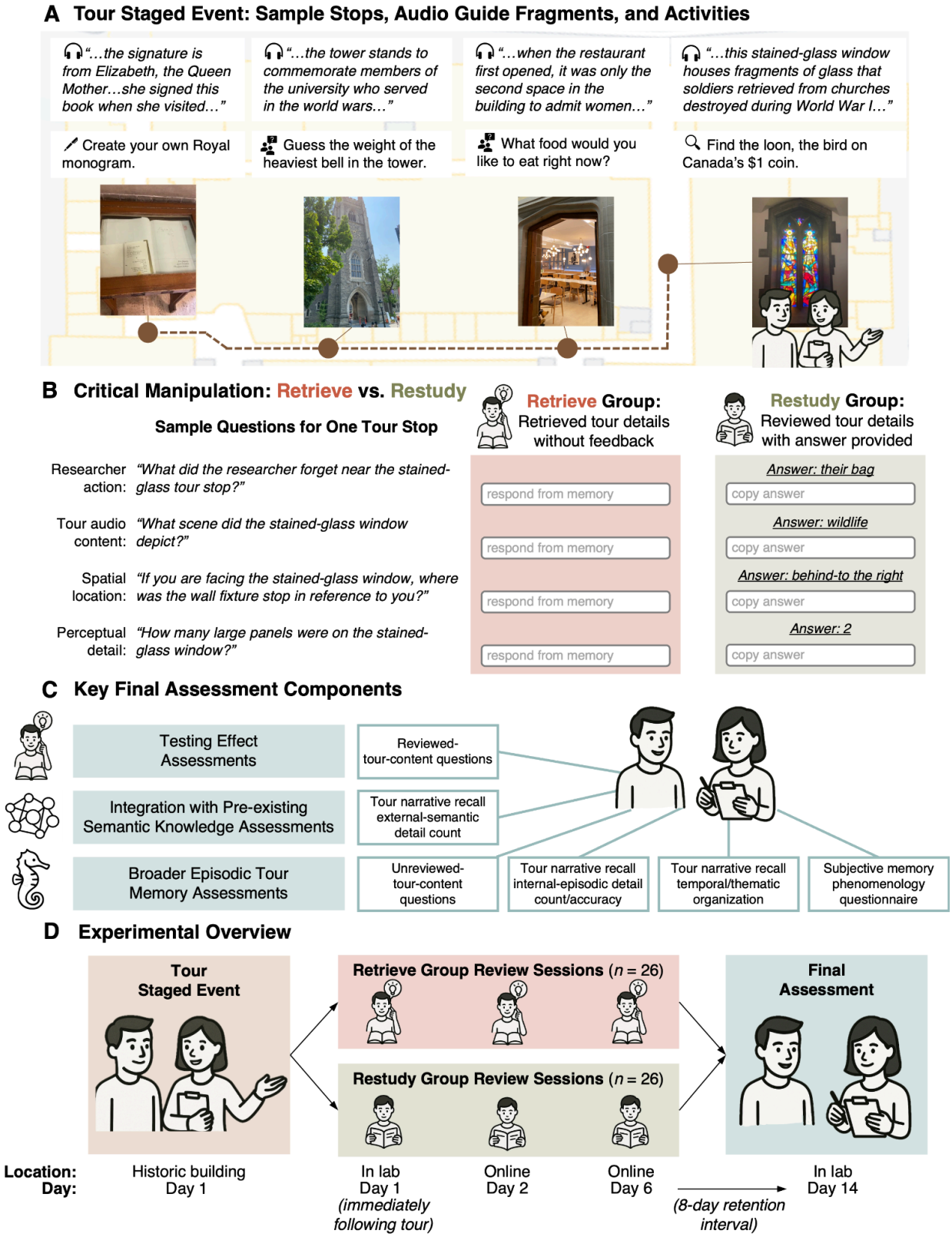
A Tour Staged Event: Sample Stops, Audio Guide Fragments, and Activities
"…the signature is from Elizabeth, the Queen Mother…she signed this book when she visited…"
"…the tower stands to commemorate members of the university who served in the world wars…"
"…when the restaurant first opened, it was only the second space in the building to admit women…"
"…this stained-glass window houses fragments of glass that soldiers retrieved from churches destroyed during World War I…"
Create your own Royal monogram.
Guess the weight of the heaviest bell in the tower.
What food would you like to eat right now?
Find the loon, the bird on Canada's $1 coin.
B Critical Manipulation: Retrieve vs. Restudy
Sample Questions for One Tour Stop
Researcher action: "What did the researcher forget near the stained-glass tour stop?"
Tour audio content: "What scene did the stained-glass window depict?"
Spatial location: "If you are facing the stained-glass window, where was the wall fixture stop in reference to you?"
Perceptual detail: "How many large panels were on the stained-glass window?"
Retrieve Group: Retrieved tour details without feedback
respond from memory
respond from memory
respond from memory
respond from memory
Restudy Group: Reviewed tour details with answer provided
Answer: their bag
copy answer
Answer: wildlife
copy answer
Answer: behind-to the right
copy answer
Answer: 2
copy answer
C Key Final Assessment Components
Testing Effect Assessments
Integration with Pre-existing Semantic Knowledge Assessments
Broader Episodic Tour Memory Assessments
Reviewed-tour-content questions
Tour narrative recall external-semantic detail count
Unreviewed-tour-content questions
Tour narrative recall internal-episodic detail count/accuracy
Tour narrative recall temporal/thematic organization
Subjective memory phenomenology questionnaire
D Experimental Overview
Tour Staged Event
Retrieve Group Review Sessions (n = 26)
Restudy Group Review Sessions (n = 26)
Final Assessment
Location:
Day:
Historic building
Day 1
In lab
Day 1
(immediately following tour)
Online
Day 2
Online
Day 6
(8-day retention interval)
In lab
Day 14

Figure 1. Study methodology and procedure overview. (A) Participants experienced an interactive, factually-detailed researcher- and audio-guided tour. **(B)** 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. **(C)** The study had three core aims—the testing effect, integration of the event with semantic knowledge structures, and broader episodic tour memory—each assessed with distinct measures. **(D)** Our key question was how the type of review session (retrieval vs. restudy) influenced staged-event memory after an 8-day delay.

The Tour Staged Event

Participants completed a 25-minute, 13-stop interactive walking tour of a historic building, combining audio- and researcher-guided components (Figure 1A). At each stop, participants listened to descriptive audio guides while the extensively-trained researcher facilitated the experience by asking self-referential questions, presenting props, guiding activities, stating facts, and performing staged actions (full experimenter script, audio guides, and testing materials available on OSF: https://osf.io/t6kc3/?view_only=8b8fc90c2f8a4cb6af5671e1ddafbb78).

Tour Review Sessions

Participants reviewed (retrieved or restudied) verifiable tour details across three review sessions (Figures 1B, 1D): immediately post-tour (in-lab), 24 hours post-tour (online), and five days post-tour (online). The expanding interval schedule was designed to balance retrieval effort and success (Toppino et al., 2018; compliance details in supplement [‘Review Session Timing Compliance’]).

Retrieve Group

The retrieve group answered 39 short-answer questions (Table S1) probing verifiable tour details without feedback; guessing was encouraged to promote engagement. Multiple questions per tour stop covered different aspects (e.g., tour audio content, perceptual details, researcher/participant actions, and spatial information). To promote context reinstatement, questions were grouped by tour stop and presented as blocks with questions for one stop appearing on the same page (not disclosed to participants). Questions spanning the tour (e.g.,

about the researcher) were also presented as a block. Each question was allotted 25 seconds, with an additional 5 seconds for pages with headers (pages advanced automatically). Questions were scored manually as correct or incorrect.

Restudy Group

The restudy group followed the same procedure as the retrieve group but viewed questions alongside correct answers and copied them into a text box (Figure 1B; restudy performance in supplement ['Restudy Performance']).

Tour Review Session Metacognition Questions

After each review session, participants answered three metacognition questions. The restudy group reported greater episodic reexperiencing and perceived review effectiveness than the retrieve group (*p*s < .05; Figures S2, S4), with no difference in anticipated future tour memory (*p* > .05; Figure S3; see supplement ['Tour Review Session Metacognition Questions'] for details).

Final Assessment Administration

Eight days after the final review session (two weeks post-tour), participants completed memory assessments in-lab (Figure 1C). Tour narratives involved verbally recalling all they remembered from the tour (adapted from 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 Diamond et al., 2020 and Levine et al., 2002; supplement provides scorer training details ['Training on Tour Narrative Scoring Protocol'] and final assessment measure administration ['Final Assessment']).

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 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. 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 created in-lab influenced later access to that event through the identical cue (Emmerdinger & Kuhbandner, 2018).

Integration with Pre-existing 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 world knowledge (e.g., “Gothic architecture originated in France”) or self-knowledge (e.g., “I enjoy Gothic architecture”; Figures 2A, 2B). 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 sufficiently novel and event-specific.

Broader Episodic Tour Memory

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.

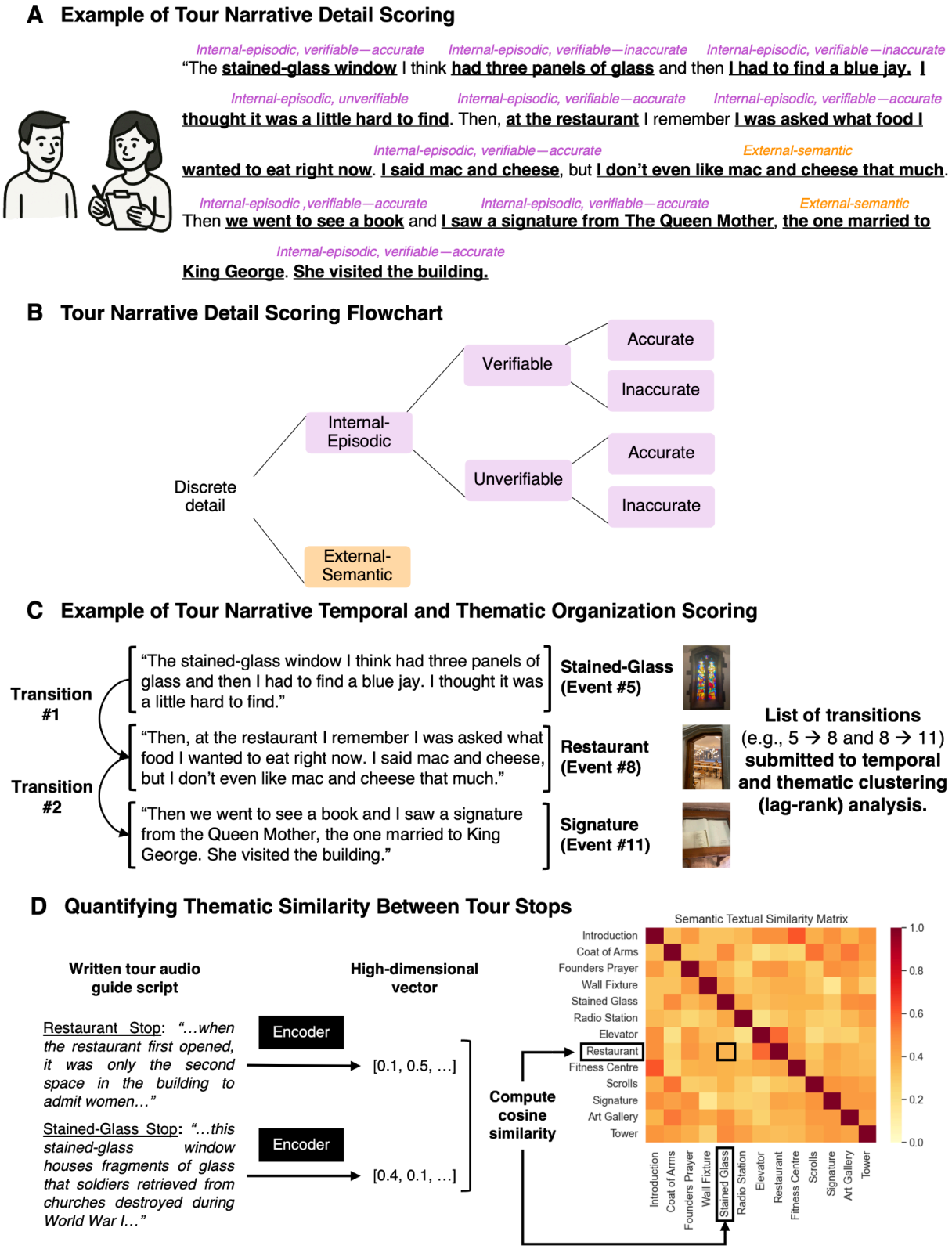

Figure 2. Tour narrative scoring procedures. (A, B) Tour narrative details were classified as internal-episodic (tour-specific; verifiable/unverifiable and correct/incorrect) or external-semantic

(facts not learned during the tour). **(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.

Internal-Episodic Details in Tour Narratives

Internal-episodic details were those that were tour-related, including factual information learned, 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 (Figures 2A, 2B).

Temporal and Thematic Organization in Tour Narratives

Recall order of broader tour events in narratives was scored (Figure 2C), counting each event as a single recall, irrespective of detail provided. 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. Thematic clustering scores (semantic lag-rank analysis) assessed recall organization by semantic relatedness, with scores above 0.5 indicating preferential recall of related stops. Semantic similarity of events was based on cosine similarities between vector embeddings of transcribed audio guides (computed using Universal Sentence Encoder, Cer et al., 2018; Figure 2D; more details in supplement ['Temporal and Thematic Organization Analyses']).

Subjective Memory Phenomenology

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

Statistical Analyses

Generalized-mixed, linear-mixed, and general models were chosen to best fit the data structure, with random intercepts included as appropriate. Key claims of no effect were

supported by Bayes factors comparing models with and without the effect in question (BF_{01}; supplement details model selection, priors, functions, and packages ['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 probes.

Results

The Testing Effect

Reviewed-Tour-Content Questions

During review sessions, the retrieve group received no feedback and only reviewed details they could recall (Figure 3A shows review session retrieve-group performance; Figure S1 shows performance by question). In contrast, the restudy group reviewed all details three times, leading to a mismatch in what information was successfully reviewed by each group before the final assessment. Thus, given our key question—whether *successfully-retrieved* information would be better remembered than restudied information after an eight-day delay—we conditionalized each retrieve participant's final assessment performance on their review session performance. Specifically, we excluded questions not recalled across all three review sessions, ensuring all analyzed items had been reviewed equally across participants. A generalized mixed model (binomial distribution) predicted correct responses to reviewed-tour-content questions, with group (treatment coded) as a fixed effect and random intercepts for question and participant (adjusted ICC = 0.366). The retrieve group was more likely to answer correctly, $b = -0.63$, 95% CI [–1.26, –0.01], SE = 0.32, $z = -1.99$, $p = .046$ (Figure 3B). Thus, retrieval practice strengthened successfully-reviewed tour content more than restudy.

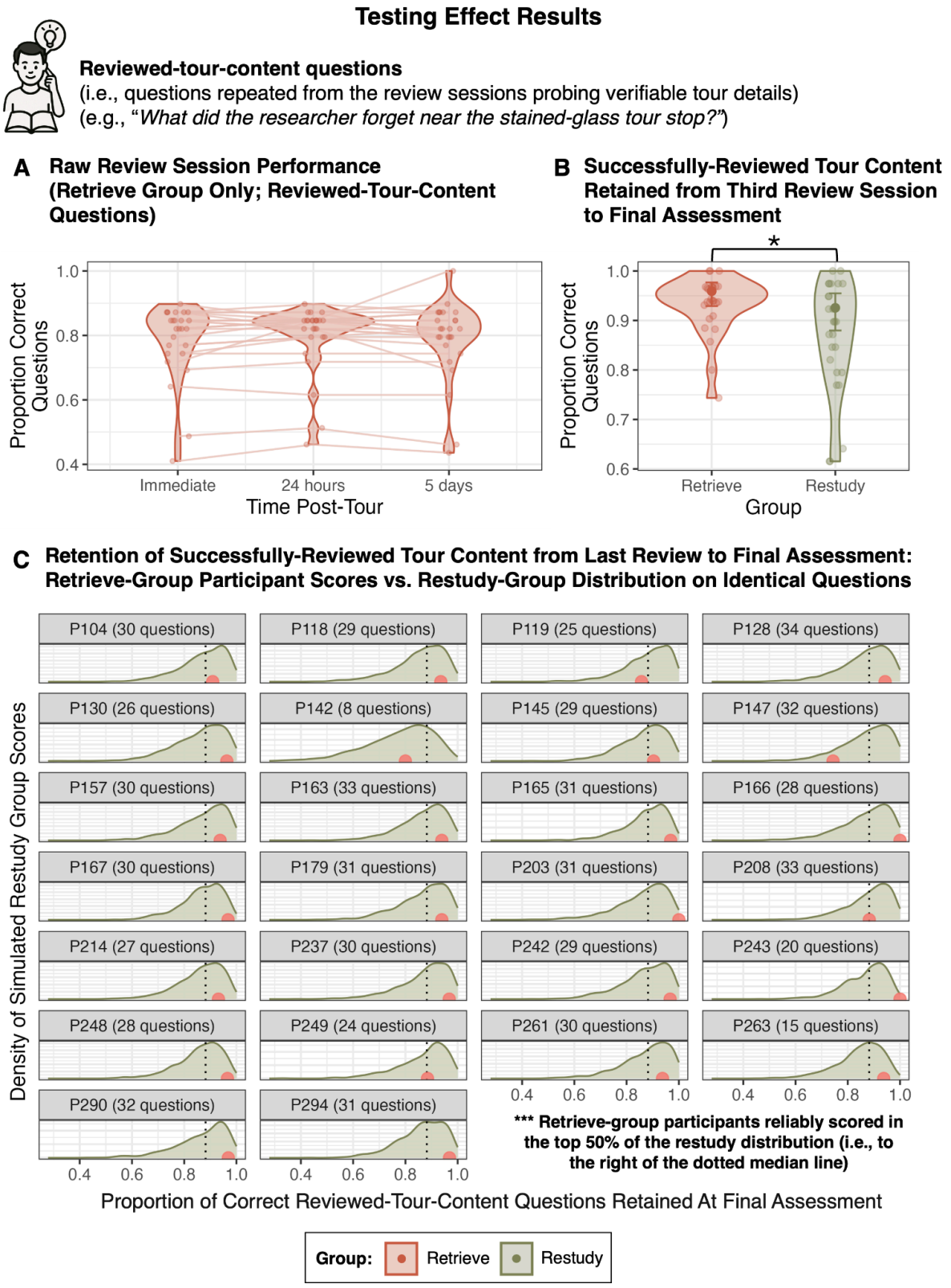

Figure 3. The testing effect results. (A) Retrieve-group performance during the review sessions was generally consistent across all three sessions. **(B)** The retrieve group outperformed the restudy group in retaining successfully-reviewed tour content from the final

review session to the final assessment (eight-day delay). To match groups, only reviewed-tour-content questions that were answered correctly in all three review sessions were included (large dots = model estimates with 95% confidence intervals). **(C)** Retrieve-group participants' retention scores for successfully-reviewed tour content (large dots) consistently ranked within the top 50% of the simulated restudy-group retention-score distributions when matched on identical questions, suggesting that the observed testing effect in (B) was not due to item effects. Beta distributions were simulated from restudy group scores (dotted line = median). The panel header shows the number of questions correctly answered by retrieve-group participants across all three sessions (39 total questions). *p < .05. ***p < .001.

To address the possibility that the retrieve group's advantage resulted from excluding more difficult questions, we compared retrieve participants' conditionalized reviewed-tour-content scores to restudy participants' scores on the same question subset. If retrieval truly strengthened memory, retrieve participants should outperform most restudy participants on the same questions; if the effect reflects item difficulty, removing harder questions would elevate restudy participants' scores, placing retrieve participants more evenly across the restudy range. For each retrieve-group participant, we calculated the proportion of reviewed-tour-content questions answered correctly at the final assessment (excluding items not recalled across all three review sessions). We then computed the same proportion for each restudy-group participant using the identical question subset for each retrieve-group participant. These restudy-group values proportions were used to simulate expected restudy performance distributions via beta distributions, creating one restudy distribution for each of the 26 retrieve-group participants. Cumulative probabilities in the lower tail of each distribution indicated the likelihood that a restudy participant would score at or below the corresponding retrieve participant. An exact binomial test showed that ~85% of retrieve-group participants ranked in the top 50% of their matched restudy-group distribution (95% CI [0.68, 1.00]), a result highly unlikely by chance (p < .001; Figure 3C), reinforcing that retrieval strengthened reviewed tour content more effectively than restudy, even when matched on question content.

For completeness, although retrieve-group participants are unlikely to recall items they missed during review—making these items effectively unanswerable at the final

assessment—we also examined unconditionalized performance across all questions. Unsurprisingly, the restudy group correctly answered more reviewed-tour-content questions correctly overall ($p < .01$; Figure S5).

Integration with Pre-existing Semantic Knowledge

External-Semantic Details in Tour Narratives

A generalized mixed model (negative binomial distribution) predicted external-semantic detail count (i.e., facts learned outside the tour context), with group (treatment coded) as a fixed effect. The retrieve group included significantly more external-semantic details in their tour narratives, $b = –0.84$, 95% CI [–1.66, –0.03], $SE = 0.41$, $z = –2.04$, $p = .041$, indicating enhanced integration of the tour with pre-existing knowledge structures (Figure 4A; Figure S6 displays external detail counts categorized by subtype [i.e., personal knowledge vs. general knowledge]).

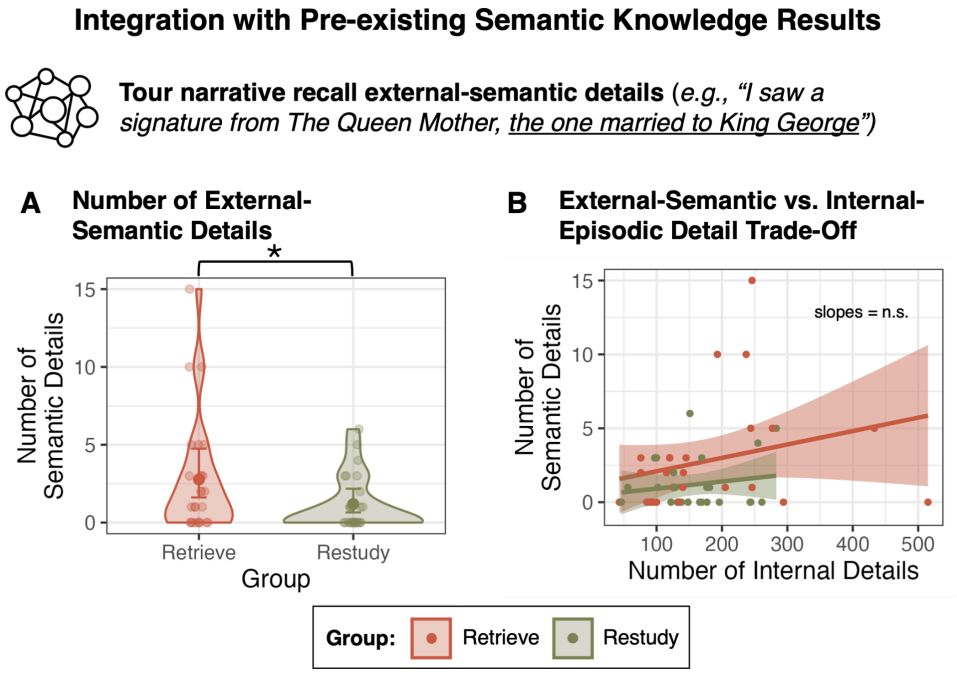

Figure 4. Integration with pre-existing semantic knowledge results. The retrieve group included more external-semantic details (i.e., non-tour-acquired personal/general facts) in tour narratives, without losing episodic detail. Panel **(A)** shows the number of external-semantic details in tour narratives (large dots = model estimates with 95% confidence intervals). Panel

(B) illustrates the relationship between internal-episodic and external-semantic detail counts (lines = slope estimates; ribbons = 95% confidence intervals). * $p < .05$.

To determine whether greater external-semantic recall reduced internal-episodic detail recall, we correlated these detail types in both groups. Although there was a numerical trend toward higher external-semantic counts being linked to higher internal-episodic counts (Figure 4B), correlations were non-significant: retrieve group, $r = 0.28$, t(24) = 1.40, 95% CI [–0.13, 0.59], $p = 0.174$; restudy group, $r = 0.17$, t(24) = 0.85, 95% CI [–0.23, 0.52], $p = 0.404$. Thus, increased external-semantic detail production did not come at the expense of episodic detail.

Broader Episodic Tour Memory

Here, we examined whether review activity (i.e., retrieval or restudy) affected broader event memory beyond reviewed details. Since retrieval may evoke holistic recollection beyond the specific content probed, even when exact answers are incorrect, controlling for review performance on the short-answer questions was neither appropriate nor feasible for the analyses below, given that it is impossible to determine what non-target information was correctly activated during review, and therefore, what should be disqualified.

Unreviewed-Tour-Content Questions

A generalized mixed model (binomial distribution) predicted correct responses to unreviewed-tour-content questions (i.e., probing tour content not covered during review sessions), with group (treatment coded) as a fixed effect and random intercepts for both question and participant (adjusted ICC = 0.393). The group effect was non-significant, b = –0.03, 95% CI [–0.28, –0.21], SE = 0.13, z = -0.26, $p = .797$ (moderate evidence for the null; $BF_{01} = 5.29$), indicating no strengthening of non-target episodic details (Figure 5A; Table S2 displays performance by question).

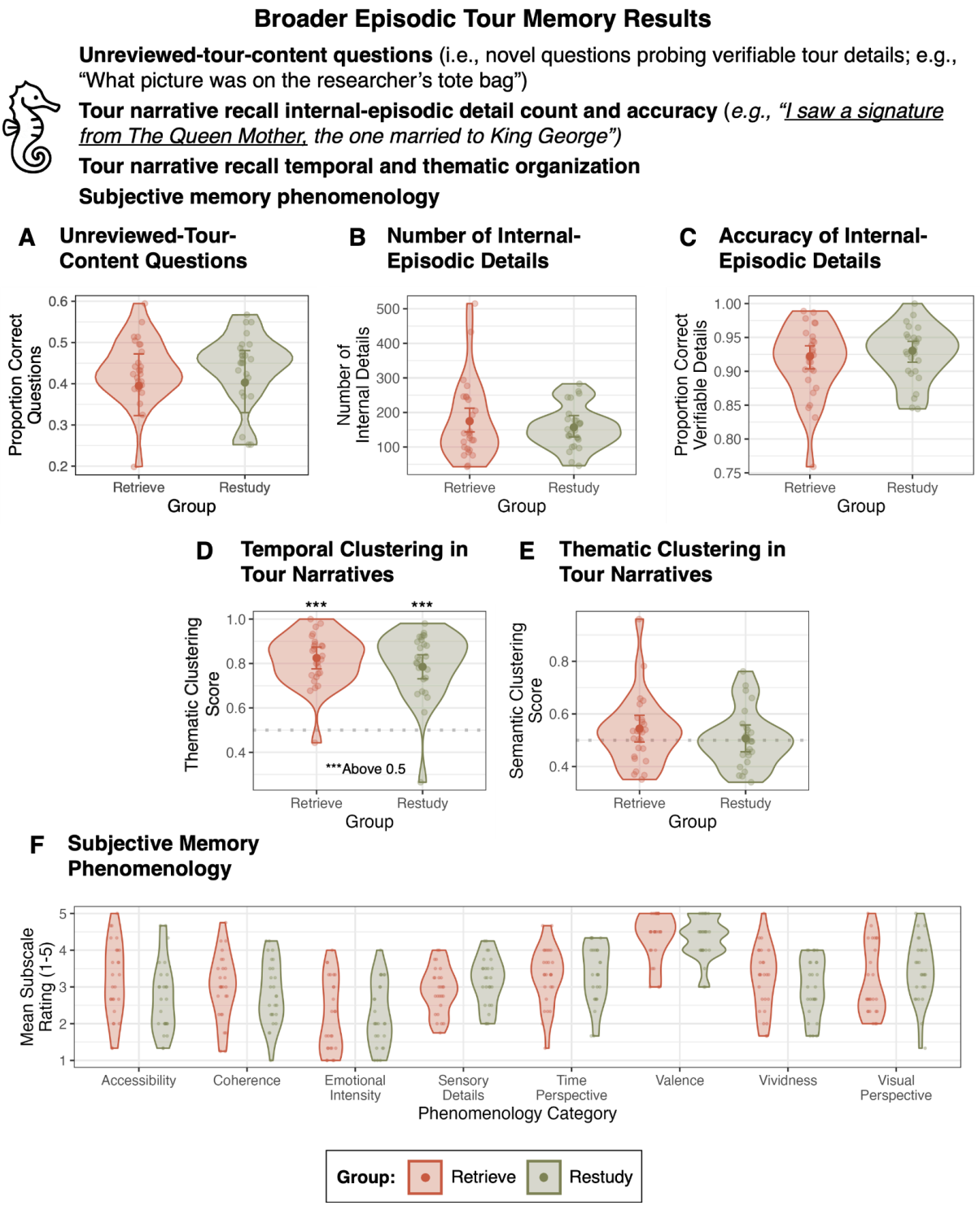

Figure 5. Broader episodic tour memory results. Overall, there was no significant difference in broader episodic tour memory between the retrieve and restudy groups (large dots = model estimates with 95% confidence intervals). Panel **(A)** shows the proportion of unreviewed-tour-content questions answered correctly. Panel **(B)** shows the number of internal-episodic details in tour narratives. Panel **(C)** shows the proportion of internal-episodic,

verifiable details that were accurate (vs. inaccurate). Panels **(D)** and **(F)** show temporal and thematic lag rank values, with chance clustering corresponding to 0.5. Temporal (but not thematic) lag rank values were significantly above 0.5 for both groups. Panel (**F)** shows average Memory Experiences Questionnaire scores for each subscale. *** p < .001.

Internal-Episodic Details in Tour Narratives

A negative binomial regression predicted internal-episodic detail count (i.e., episodic details pertaining to the tour event), with group (treatment coded) as a fixed effect. The group effect was non-significant, b = 0.11, 95% CI [–0.17, 0.38], SE = 0.14, z = 0.75, p = .452 (moderate evidence for the null; BF_{01}= 3.66), indicating retrieval did not enrich episodic tour recall (Figure 5B; Figures S8, S9, and S10 display internal-episodic detail counts by subtype [e.g., event, perceptual, thought/emotion, place, time], verifiability, and whether they were reviewed in tour sessions).

A generalized mixed model (binomial distribution) predicted internal-episodic detail accuracy, with group (treatment coded) as a fixed effect and a random intercept for participant (adjusted ICC = 0.071). The group effect was non-significant, *b* = –0.12, 95% CI [–0.45, 0.21], *SE* = 0.17, *z* = –0.71, *p* = .479 (moderate evidence for the null; BF_{01}= 3.30), indicating retrieval did not modulate overall episode accuracy (Figure 5C; Figure S11 displays error counts by subtype).

Temporal and Thematic Organization in Tour Narratives

A generalized regression model (beta distribution) predicted temporal clustering value using group (effect coded) as a fixed effect. The intercept was significant, *b* = 1.42, 95% CI [1.18, 1.67], *SE* = 0.12, *z* = 11.47, *p* < .001, indicating clustering was above chance, averaging 0.81 across groups. The group effect was non-significant, *b* = –0.13, 95% CI [–0.35, 0.10], *SE* = 0.11, *z* = –1.12, *p* = 0.264 (anecdotal evidence for the null; BF_{01}= 2.29), indicating that while participants clustered their narratives temporally, retrieval did not promote temporal disorganization (Figure 5D; lag-conditional response probability curves [Figure S12] and explicit temporal sequencing task results [Figure S13] corroborate these findings).

A generalized regression model (beta distribution) predicted thematic clustering value using group (effect coded) as a fixed effect. The intercept was non-significant, b = 0.10, 95% CI [–0.04, 0.25], SE = 0.07, z = 1.39, p = .165, indicating clustering was not above chance, averaging 0.53 across groups. The group effect was non-significant, b = –0.07, 95% CI [–0.22, 0.07], SE = 0.07, z = –1.01, p = 0.312 (moderate evidence for the null; BF_{01}= 3.37), indicating that participants did not cluster their narratives thematically, and retrieval did not promote thematic clustering (Figure 5E).

Subjective Memory Phenomenology

Linear mixed models predicted mean MEQ-SF subscale scores, using group (treatment coded) as a fixed effect, for the constructs of accessibility, coherence, emotional intensity, sensory details, time perspective, valence, vividness, and visual perspective. The group effect was non-significant in all models (ps < .083; anecdotal to moderate support for the null model for all categories [$BF_{01} \geq 1.75$] except accessibility [$BF_{01} = 0.89$], indicating anecdotal support for the alternative model; Table S3 presents all model results). Thus, retrieval did not influence subjective memory phenomenology (Figure 5F).

Discussion

Here, we leveraged the inherent multidimensionality of real-world memories to investigate how retrieval practice fundamentally shapes and reorganizes memory. Participants first experienced a real-world staged event (an interactive guided tour of a historic building). They subsequently completed three review sessions over five days, involving either repeatedly retrieving or restudying event details. Memory was assessed eight days after the final review session. Using diverse analytic approaches, three key findings emerged. First, to our knowledge, we provide the first classic testing effect demonstration for real-world event memory: event details successfully retrieved during review were remembered better after an eight-day delay than those that were restudied. Second, retrieval enhanced the incorporation of pre-existing semantic knowledge into event recall, supporting theories that retrieval embeds

episodes within neocortical semantic systems more effectively than simple reexposure (Antony et al., 2017). Third, this semantic integration did not correspond with enhancement—or impairment—of broader episodic memory: retrieval had no effect on episodic richness, accuracy, temporal or thematic organization, or subjective phenomenology. This suggests that, compared to restudy, retrieval does not restructure memory at a within-episode level. Overall, by assessing retrieval's impact on real-world memories using nuanced measures that capture mnemonic changes across multiple dimensions, our study not only demonstrates the testing effect in a naturalistic context but also reveals that retrieval shapes long-term memory by integrating episodes with neocortical semantic networks, without degrading broader episodic content.

First, we demonstrate a testing effect in a novel context: retrieval practice enhanced retention of successfully-reviewed real-world event details more than restudy. This practically-significant finding extends prior evidence from laboratory and educational settings (Karpicke, 2017) to complex, immersive experiences. The testing effect's persistence in this naturalistic context—robust across diverse episodic details like observed/completed actions, perceptual features, spatial information, and newly-acquired facts—reinforces that retrieval benefits transcend stimulus type.

Second, we show retrieval practice promotes semantic integration of episodes: participants incorporated more pre-existing (i.e., not acquired during the tour) semantic information into recall narratives following retrieval than restudy. This supports theories suggesting that retrieval—similarly to sleep—embeds initially-hippocampus-dependent episodic memories into stable neocortical semantic networks (Antony et al., 2017). 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), markers of semantic integration. Such integration may underlie the testing effect by enabling retrieved details to rely on more stable neocortical traces, especially important at longer delays

when hippocampal traces may have degraded (Antony et al., 2017). Our exploratory analysis (requiring replication) offers preliminary support for this idea: semantic integration at Day 14 marginally predicted performance on “testing effect” questions at a 12–18 month delay, when neocortical reliance likely predominates (Fig S7). Together, these findings highlight semantic integration as a key potential mechanism through which retrieval stabilizes memories.

The observed semantic integration effects may arise from more extensive co-activation of related memories during retrieval compared to restudy (Antony et al., 2017). Mechanistically, retrieval begins with an incomplete cue that triggers holistic pattern completion of the episode in the hippocampus, activating not only the target memory but also related non-target memories across hippocampal–neocortical circuits (Antony et al., 2017; Horner et al., 2015). Consequently, both target and co-activated memories are modified: strongly co-activated memories may be strengthened and integrated (representations become more similar), while moderately co-activated memories may be weakened and differentiated (representations diverge; Ritvo et al., 2019). Here, retrieving reviewed episodic tour content may have strongly co-activated semantically-related non-target memories, thereby integrating the episode with broader semantic knowledge structures. Unlike retrieval, restudy is thought to strongly activate the target memory but suppresses co-activation of related non-targets via lateral inhibition, limiting such representational changes (Ritvo et al., 2019). Although our behavioral study could not directly assess these mechanisms, future research should investigate co-activation patterns during retrieval versus restudy, their relationship to neural indices of representational integration and differentiation, and whether representational changes predict long-term retention of retrieved content. Furthermore, the dependence of semantic integration effects on sleep is unresolved. Sleep may moderate retrieval effects such that retrieved memories are tagged for prioritized replay during sleep, potentially broadening retrieval practice benefits by more readily reactivating content from temporally-separated episodes, and thereby promoting more thorough integration of semantically-related experiences across time than retrieval alone (Liu et al., 2024;

Liu & Ranganath, 2021). Ultimately, understanding the neurocomputational dynamics at play and retrieval's interaction with sleep will be crucial for uncovering the mechanisms by which retrieval shapes memory.

Third, beyond strengthening reviewed content, retrieval practice did not alter memory at a within-episode level: memory for unreviewed event details, overall accuracy, event organization (temporal and thematic), and subjective phenomenology were comparable across conditions. One possibility is that both retrieval and restudy elicited comparable pattern completion and co-activation of non-target episodic content (i.e., co-activation was similarly sparse or similarly holistic), resulting in similar representational outcomes. Alternatively, co-activation may have differed, but due to the proposed U-shaped, nonmonotonic relationship between co-activation strength and memory change (Ritvo et al., 2019), similar outcomes could have been observed in the retrieve and restudy conditions, despite differing levels of co-activation of episodic details. However, without neural data, we cannot adjudicate between these explanations—reinforcing the need to examine how co-activation patterns of within-episode content differ between retrieval and restudy. Importantly, comparable group performance does not imply that retrieval fails to induce change in broader episodic memory; rather, it suggests that retrieval may not enhance such changes beyond what thorough re-exposure achieves. Without a no-review control condition, however, it remains unclear whether either form of review produced memory modifications relative to a no-review baseline. Finally, our findings that retrieval promotes stronger semantic connections without a corresponding loss of episodic detail challenge standard consolidation accounts. Instead, they support the idea that distinct memory representations (episodic and semantic) can dynamically coexist and be expressed concurrently, rather than one replacing the other (Gilboa & Moscovitch, 2021; Winocur & Moscovitch, 2011).

Conclusion

Here, we harnessed the complexity of real-world event memories to reveal nuanced effects of retrieval practice versus restudy. Compared to restudy, retrieval enhanced memory for reviewed content—demonstrating a novel testing effect for real-world event memories—and selectively promoted integration of the episode with pre-existing semantic knowledge structures, without altering the strength, accuracy, or organization of broader episodic content. These findings suggest that retrieval shapes long-term memory by embedding retrieved episodes into stable semantic knowledge networks, a process that may underlie the testing effect.

References

Antony, J. W., Catarina S. Ferreira, 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

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., & Kurzweil, R. (2018). *Universal Sentence Encoder* (arXiv:1803.11175). arXiv. https://doi.org/10.48550/arXiv.1803.11175

Diamond, N. B., Armson, M. J., & Levine, B. (2020). The truth is out there: Accuracy in recall of verifiable real-world events. *Psychological Science*, *31*(12), 1544–1556. https://doi.org/10.1177/0956797620954812

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

Ferreira, C. S., Charest, I., & Wimber, M. (2019). Retrieval aids the creation of a generalised memory trace and strengthens episode-unique information. *NeuroImage*, *201*, 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

Horner, A. J., Bisby, J. A., Bush, D., Lin, W.-J., & Burgess, N. (2015). Evidence for holistic episodic recollection via hippocampal pattern completion. *Nature Communications*, *6*, Article 7462. https://doi.org/10.1038/ncomms8462

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 Press. https://doi.org/10.1016/B978-0-12-809324-5.21055-9

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*, Article 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

Liu, X. L., Ranganath, C., & O'Reilly, R. C. (2024). A complementary learning systems model of how sleep moderates retrieval practice effects. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-024-02489-1

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

Reagh, Z. M., & Ranganath, C. (2023). Flexible reuse of cortico-hippocampal representations during encoding and recall of naturalistic events. *Nature Communications*, *14*, 1279. https://doi.org/10.1038/s41467-023-36805-5

Ritvo, V. J. H., Nguyen, A., Turk-Browne, N. B., & Norman, K. A. (2024). A neural network model of differentiation and integration of competing memories. *eLife*, *12*, Article RP88608. https://doi.org/10.7554/eLife.88608

Ritvo, V. J. H., Turk-Browne, N. B., & Norman, K. A. (2019). Nonmonotonic 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

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

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

Supplemental material for “Putting the testing effect to the test in the wild: Retrieval enhances real-world memories and promotes their semantic integration while preserving episodic integrity”

This file contains supplementary text, Figures S1 to S15, and Tables S1 to S3.

Other supplementary materials for this manuscript include study materials, anonymized data (excluding tour narrative recordings and transcripts), and code provided in an Open Science Framework repository:
https://osf.io/t6kc3/?view_only=8b8fc90c2f8a4cb6af5671e1ddafbb78

See the Table of Contents (next page) for more information.

Table of Contents

Supplemental Methods

Participants

Recruitment

Participants were recruited from the local community through a mix of social media (e.g., Facebook, Reddit, and Instagram) and snowball sampling methods.

Power 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 (Diamond et al., 2020). We increased this value by 1.5, as our tour contained more content compared to 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 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 conditions). 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.

Exclusion Criteria

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 memory of the building (one participant in each group). 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 (see OSF for statistical output files). Three participants were excluded before statistical analyses due to not attempting to recall the tour during the final assessment (n = 2 in the retrieve group), and (ii) due to a different experimenter conducting the final assessment (i.e., for all other participants, the same researcher conducted both the tour and the final assessment session; n = 1 in the restudy group).

Review Session Timing Compliance

All participants completed all components of the study; however, some did not complete the review sessions on time (i.e., they completed them after 2 AM the following day). All participants completed session one on time. Three participants (all from the retrieve group) completed session two late, three participants (two from the retrieve group and one from the restudy group) completed session three late, and two participants (both from the retrieve group) completed the follow-up late. On average, session two was completed 1.15 days (SD = 0.46) after the scheduled time in the retrieve group and 1.00 day (SD = 0) in the restudy group. Session three was completed 5.19 days (SD = 0.49) after the scheduled time in the retrieve group and 5.19 days (SD = 0.80) in the restudy group. The follow-up occurred 13.12 days (SD =

0.43) post-tour in the retrieve group and 13.00 days (SD = 0) in the restudy group, or 7.92 days (SD = 0.63) and 7.81 days (SD = 0.80) after session three, respectively. These participants were included in the final analyses, although the pattern of results remained unchanged regardless of their inclusion (see OSF for statistical output files).

Restudy Performance

Errors during restudy (e.g., missing a question or copying an answer incorrectly) were extremely rare. Three participants made a single error in either Session 2 or Session 3. One participant made one error in both Session 2 and Session 3, but on different questions. Notably, all errors across participants occurred on different questions.

Final Assessment

Measures were administered in the following order: (1) tour narrative free recall and general probe sections, in which participants verbally described everything they could remember from the tour (adapted from the Autobiographical Interview; Levine et al., 2002), (2) memory phenomenology questionnaire, (3) metacognition questions probing perceived review session efficacy and confidence in memory for the tour event (see Figures S11 and S12), (4) explicit temporal sequencing task, (5) tour narrative specific probing section (originally administered to procure additional information participants had left out of their narratives, although these data are not discussed further), and (6) reviewed-tour-content (testing effect) questions intermixed with unreviewed-tour-content questions. All measures, except for tour narratives, were administered using Qualtrics.

Training on Tour Narrative Scoring Protocol

All four research assistants were trained under the supervision of the first author, who created the scoring manual. Training began with pilot study tour memories, which included a slightly different tour and review questions. Each scorer and the first author compared internal and external detail counts for five narratives, calculating intraclass correlation coefficients (ICCs)

using the icc() function from the psych package (Revelle, 2024). ICCs (ICC2) were computed between the first author and each scorer separately. Other detail types (e.g., verifiability, reviewed vs. unreviewed, errors, and internal/external detail subtypes) were examined for consistency across scorers. Scorers met with the first author to resolve discrepancies before rescoring the five initial transcripts along with two additional pilot narratives. If ICCs for internal and external details met the adequacy threshold (≥0.8; all did after review), scorers proceeded to score two transcripts from the actual dataset. The first author then reviewed every detail of these transcripts, discussed discrepancies, and ensured accurate application of the scoring protocol. Throughout scoring, scorers met regularly with the first author to clarify any uncertainties.

Temporal and Thematic Organization Analyses

Lag-rank analyses were conducted using the Psifr package (Morton, 2020) in Python. 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. 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.

Statistical Analyses

All statistical analyses were conducted using R (R Core Team, 2013; version 4.4.0). Package version information is included in output files on OSF.

For continuous outcomes, either a linear model was fitted using the lm function from the *stats* package (R Core Team, 2013), or a linear mixed model was fitted using the lmer function from the *lme4* package (Bates et al., 2015). For binary outcomes, generalized linear mixed

models with a binomial distribution and a logit link function were fitted using the glmer function from the *lme4* package. For count data, a Poisson regression model was initially fitted using the glmer function from the *lme4* package or the glm() function from the *stats* package. However, if overdispersion was detected using the check_overdispersion function from the *performance* package (Lüdecke et al., 2021), indicating that a Poisson regression would be too liberal, a negative binomial regression model was instead fitted to account for the excess variance using either the glm.nb function from the *MASS* package (Venables & Ripley, 2002) or the glmer.nb function from the *lme4* package. For data bound by 0 and 1, to ensure that predictions line up with the constraints of the data, a beta regression model was fitted using the betareg function from the *betareg* package (Cribari-Neto & Zeileis, 2010) using a logit link function. Any values of one were adjusted to 0.9999, since data could not take on values of 0 or 1. Correlation tests were conducted using the cor.test function from the *stats* package. For ordinal outcomes, cumulative link models were fitted using either the clm() function or the clmm() function from the *ordinal* package (Christensen, 2023) using adaptive Gauss-Hermite quadrature approximation with 10 quadrature points. Wald 95% confidence intervals were calculated for regression parameters using the confint() function from the *stats* package. Adjusted intraclass correlation coefficients (ICC's) for all mixed models are reported, computed using the ICC function from the *performance* package. Polynomial contrasts were fitted using the contr_code_anova function from the *faux* package (DeBruine, 2025). Estimated marginal means were computed using the emmeans function from the *emmeans* package (Lenth et al., 2023).

Beta distributions were fit to restudy-group scores, and shape parameters were obtained, using the fitdist() function from the fitdistrplus package (Delignette-Muller & Dutang, 2015). Beta distributions were simulated for each set of restudy data using the rbeta() function (with 1000 samples) from the rBeta2009 package (Cheng et al., 2024). Cumulative probability values in the lower tail of the beta density distribution were extracted using the pbeta() function from the fitODBOD package (Mahendran & Wijekoon, 2024). An exact binomial test was

conducted to compare the observed number of successes with what would be expected based on a binomial distribution, using the binom.test() function from the stats package.

Bayes factors were computed using the bayes_factor function from the *brms* package (Bürkner, 2017). The brm function from the *brms* package was used to fit bayesian models, using four Markov chain Monte Carlo chains, each running for 10000 iterations (unless more iterations were required for the model to converge). Priors chosen depended on the nature of the model, the link function used, and the parameter in question. For mixed models, priors for random effects standard deviations were set to exponential(1), reflecting positive and typically small variations. All models converged (Rhats = 1). Values between 1–3 were interpreted as anecdotal evidence for the null. Values between 3 and 10 were interpreted as moderate evidence for the null. Values above 10 were interpreted as strong evidence for the null.

Bayesian Models and Priors

Unreviewed-Tour-Content Questions. Bayesian models with a Bernoulli family (logit link) were fitted, with and without the fixed effect of group. The prior for the intercept was set to normal(0, 1.5), reflecting a baseline response probability of approximately 50% ($p = e^0/(1+e^0) = 0.5$) for the baseline (restudy) group, with plausible values primarily ranging from approximately 18% to 83%. The prior for the group fixed effect (*b*) was set to cauchy(0, 0.5), indicating no prior preference for direction while placing most of the prior mass within a ±12% probability range, but allowing for larger effects through its heavier tails.

Number of Internal Details. Bayesian models with a negative binomial family (log link) were fitted, with and without the fixed effect of group. The prior for the intercept was set to normal(4.94, 0.25), reflecting the belief that there would be a baseline of approximately 140 details (exp(4.94)) in the baseline (retrieve) group, with plausible values primarily ranging from approximately 109 to 179. The prior for the group fixed effect (*b*) was set to cauchy(0, 0.2), indicating no prior preference for direction while placing most of the prior mass within ±0.2 log-units, but allowing for larger effects through its heavier tails. The default prior for the shape parameter inv_gamma(0.4, 0.3) was used

Reviewed vs. Unreviewed Internal Details. Bayesian models with a Bernoulli family (logit link) were fitted, with and without the fixed effect of group. The prior for the intercept was set to normal(–1.74, 1.5), reflecting a baseline response probability of approximately 15% ($p = e^0/(1+e^0) = 0.5$) for a detail being classified as reviewed (vs. unreviewed) ($p = e^0/(1+e^0) = 0.5$) for the baseline (restudy) group, with plausible values primarily ranging from approximately <1% to 44%. The prior for the group fixed effect (*b*) was set to cauchy(0, 0.5), indicating no prior preference for direction while placing most of the prior mass within a ±12% probability range, but allowing for larger effects through its heavier tails.

Accuracy of Internal Details. Bayesian models with a Bernoulli family (logit link) were fitted, with and without the fixed effect of group. The prior for the intercept was set to normal(2.75, 1), reflecting a baseline response probability of approximately 94% ($p = e^0/(1+e^0) = 0.94$) for a detail being classified as accurate (vs. inaccurate) for the baseline (restudy) group, with plausible values primarily ranging from approximately 85% to 98%. The prior for the group fixed effect (*b*) was set to cauchy(0, 0.5), indicating no prior preference for direction while placing most of the prior mass within a ±12% probability range, but allowing for larger effects through its heavier tails.

Temporal Clustering in Tour Narratives. Bayesian models with a Beta family (log link) were fitted, with and without the fixed effect of group. The prior for the intercept was set to normal(-0.16, 0.5), reflecting a baseline temporal clustering of approximately 0.85 (exp(0.16)) on average across groups, with plausible values primarily ranging from approximately .52 to .99. The prior for the group fixed effect (b) was set to cauchy(0, 0.3), indicating no prior preference for direction while placing most of the prior mass within ±0.3 log-units, but allowing for larger effects through its heavier tails. The default prior for the shape parameter gamma(0.01, 0.01) was used.

Semantic Clustering in Tour Narratives. Bayesian models with a Beta family (log link) were fitted, with and without the fixed effect of group. The prior for the intercept was set to

normal(-0.51, 0.5), reflecting a baseline temporal clustering of approximately 0.6 (exp(0.51)) on average across groups, with plausible values primarily ranging from approximately .36 to .99. The prior for the group fixed effect (*b*) was set to cauchy(0, 0.3), indicating no prior preference for direction while placing most of the prior mass within ±0.3 log-units, but allowing for larger effects through its heavier tails. The default prior for the shape parameter gamma(0.01, 0.01) was used.

Temporal Sequencing Task. Bayesian models with a Beta family (log link) were fitted, with and without the fixed effect of group. The prior for the intercept was set to normal(-0.16, 0.5), reflecting a baseline correlation approximately 0.85 (exp(0.16)) for the baseline (restudy??) group, with plausible values primarily ranging from approximately .52 to .99. The prior for the group fixed effect (*b*) was set to cauchy(0, 0.3), indicating no prior preference for direction while placing most of the prior mass within ±0.3 log-units, but allowing for larger effects through its heavier tails.

Subjective Memory Phenomenology. Bayesian models with a Gaussian family were fitted, with and without the fixed effect of group. The prior for the intercept was set to normal(3, 2) with upper and lower bound of 1 and 5, reflecting baseline mean subscale rating of 2.5 in the baseline group, with plausible values for the whole range of the scale (values could take 1-5). The prior for the group fixed effect (*b*) was set to cauchy(0, 0.5), indicating no prior preference for direction while placing most of the prior mass within ±0.5 units, but allowing for larger effects through its heavier tails.

External Semantic vs. Internal Detail Trade-Off. Bayesian models with a Gaussian family were fitted separately for the retrieve and restudy groups, with external semantic detail count as a predictor of internal detail count. The null model excluded external semantic detail count as a predictor, as the goal was to assess whether there is support for the claim that no trade-off occurred. The prior for the intercept was set to Normal(140,40), reflecting a baseline group mean of 140 with plausible values primarily ranging from 100 to 180. The prior for the group fixed effect (b) was set to Cauchy(0,0.5), indicating no prior preference for direction, while

placing most of the prior mass within ±0.5 units but allowing for larger effects due to its heavier tails.

Tour Review Session Results

Retrieve-Group Performance During Tour Review Sessions (Figure S1 and Table S1)

For the retrieve group only, a generalized mixed effects model with a binomial distribution and a logit link function was fitted to determine whether the outcome of responding correctly to a review session question changed across review sessions or to the final assessment. The fixed effect of session (i.e., timing of the review session question administration; four levels) was treatment coded, using session one as the reference level. Random intercepts were included for the crossed variables of review session question and participant (adjusted ICC for the model = 0.426). There was no evidence that the likelihood of responding correctly to a review session question changed from session one to session two (b = 0.09, 95% CI [–0.16, 0.33], SE = 0.13, z = 0.70, p = .482), from session one to session three (b = 0.02, 95% CI [–0.22, 0.27], SE = 0.12, z = 0.19, p = .849), or from session one to the final assessment (b < 0.01, 95% CI [–0.24, 0.24], SE = 0.12, z = 0.00, p > .999). Performance was generally high; the estimated marginal means for the probability of answering a review session question correctly across the four sessions were as follows: Session 1 (M = 0.87, 95% CI [0.79, 0.92]), Session 2 (M = 0.88, 95% CI [0.80, 0.93]), Session 3 (M = 0.87, 95% CI [0.79, 0.92]), and final assessment (M = 0.87, 95% CI [0.79, 0.92]). Thus, the likelihood of responding correctly to a review session question did not change across review sessions.

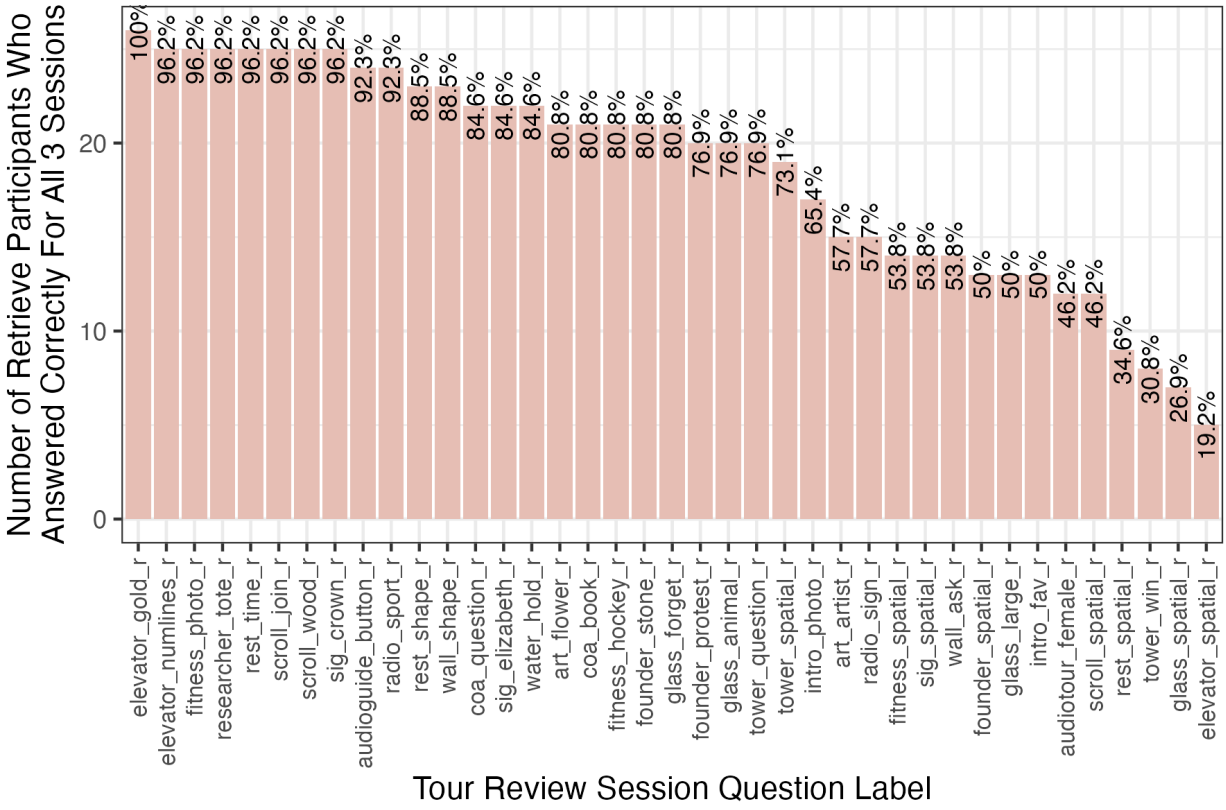

Figure S1. Number of participants who answered a given tour review session question correctly across all three review sessions. Percentages indicate the proportion of retrieve-group participants who consistently responded correctly across all three review sessions. See Table S1 for question text. Some reviewed-tour-content questions were consistently answered correctly by nearly all participants in the retrieve-group across all three review sessions, while others had lower accuracy. However, every question was answered correctly by at least five participants in each session. Certain types of questions were more difficult than others. Questions probing the spatial location of a tour stop or item in reference to another location or item were among the most challenging (e.g., elevator_spatial_r, glass_spatial_r, rest_spatial_r, scroll_spatial_r, founder_spatial_r, sig_spatial_r, fittness_spatial_r). Questions probing more incidental perceptual tour details were also particularly challenging, for instance, memory for the number of large windows on the tower (tower_win_r), memory for the number of large panes of glass on the stained-glass window (glass_large_r), or memory for the colours present on the sign for the radio station (radio_sign_r). Similarly, questions involving incidental tour facts were also among the most difficult, for instance, memory for the name of the female guide from the audio guide (audio_female_r) as well as memory for the guide's favourite things about the building (intro_fav_r). Conversely, certain types of questions were answered more accurately. Questions about salient perceptual features from tour items were answered correctly by many participants, such as the colour of the elevator doors (*elevator_gold_r*), the colour of the researcher's tote bag (*researcher_tote_r*), the material composition of the scroll frames (*scroll_wood_r*), the content of the photo at the fitness centre (fitness_photo_r), or the image on the piece of paper they were required to draw on (*sig_crown_r*). Information about activities the participants engaged in was similarly answered with high accuracy, such as the number of lines of Braille they read (*elevator_numlines_r*) or the content of the question they were asked at the scrolls

stop (*scroll_join_r*). Additionally, certain facts from the tour audio were responded to with high accuracy, such as the time women were allowed in the restaurant during a period of gender segregation in the building (r*est_time_r*).

Table S1. Tour review session questions with corresponding answers provided to the restudy group. Questions were crafted to elicit short, unambiguous responses, to best control for content equivalence between retrieve and restudy groups. A pilot study using a modified tour helped refine questions for clarity and difficulty.

Question Label	Block	Question Text	Restudy Answer
researcher_tote_r	1 (Overall Tour)	What colour was the researcher's tote bag?	Yellow
audioguide_button_r	1 (Overall Tour)	On what side of the audio device was the button you used to check the time? [Options: front, back, left, right, top, or bottom]	Left
audiotour_female_r	1 (Overall Tour)	What was the name of the female tour guide from the tour audio device?	Annie
intro_photo_r	2 (Introduction Stop)	There was a photo on the index card attached to the audio device. What did the photo portray?	The exterior of Hart House
intro_fav_r	2 (Introduction Stop)	What aspect of Hart House did the audio tour guides like the most?	The pool
coa_question_r	3 (Coat of Arms Stop)	At this stop, the researcher asked: "If you were to design your own coat of arms, what _____ would you include?"	Animal
coa_book_r	3 (Coat of Arms Stop)	How many books were on the coat of arms that you examined?	Two
founder_protest_r	4 (Founders Prayer Stop)	Why was there a protest during the visiting debater's visit?	Women could not attend the debate

Question Label	Block	Question Text	Restudy Answer
founder_stone_r	4 (Founders Prayer Stop)	What material was used for the Founders' Prayer?	Stone
founder_spatial_r	4 (Founders Prayer Stop)	If you are facing the Founders' Prayer, where was the door mentioned at the coat of arms stop in reference to you? [Options: behind–to the right, behind–to the left, same wall–to the left, same wall–to the right]	Same wall-to the left
wall_ask_r	5 (Wall Fixture Stop)	At this stop, the researcher asked: "Do you _____?"	Believe in ghosts
wall_shape_r	5 (Wall Fixture Stop)	What shape was the wall fixture?	Rectangular
glass_animal_r	6 (Stained-Glass Stop)	What scene did the stained-glass window depict?	Canadian wildlife
glass_large_r	6 (Stained-Glass Stop)	How many large panels were on the stained-glass window?	Two
glass_forget_r	6 (Stained-Glass Stop)	What did the researcher forget near this tour stop?	Their tote bag
glass_spatial_r	6 (Stained-Glass Stop)	If you are facing the stained-glass window, where was the wall fixture stop in reference to you? [Options: behind–to the right, behind–to the left, same wall–to the left, same wall–to the right]	Behind–to the right
radio_sign_r	7 (Radio Station Stop)	The colours on the radio station sign were blue and _____.	Orange

Question Label	Block	Question Text	Restudy Answer
radio_sport_r	7 (Radio Station Stop)	After the tour audio, the researcher asked you: “What is your favourite _____?”	Genre of music
elevator_gold_r	8 (Elevator Stop)	What colour were the elevator doors?	Gold
elevator_numlines_r	8 (Elevator Stop)	How many lines of braille did you read?	Two
elevator_spatial_r	8 (Elevator Stop)	If you are facing the elevator, where was the radio station stop in reference to you? [Options: behind–to the right, behind–to the left, same wall–to the left, same wall–to the right]	behind–to the left (multiple choices given)
rest_shape_r	9 (Restaurant Stop)	What shape were the lightbulbs on the chandeliers in the restaurant?	Spherical
rest_time_r	9 (Restaurant Stop)	After what time could women enter the restaurant when it opened?	3 pm
rest_spatial_r	9 (Restaurant Stop)	If you are facing the door to the restaurant, where was the water fountain in reference to you? [Options: behind–to the right, behind–to the left, in front–to the left, in front–to the right]	Behind–to the right
water_hold_r	10 (Water Fountain Event)	What did you hold close to the water fountain?	A clipboard
fitness_photo_r	11 (Fitness Centre Stop)	In the photograph you examined, what were the people doing?	Running

Question Label	Block	Question Text	Restudy Answer
fitness_hockey_r	11 (Fitness Centre Stop)	The person you read about joined a boy's _____ team.	Hockey
fitness_spatial_r	11 (Fitness Centre Stop)	If you are facing the fitness centre hallway, where was a set of stairs in reference to you? [Options: forward, behind, left, right]	Left
scroll_wood_r	12 (Scrolls Stop)	What material were the scroll handles made of?	Wood
scroll_join_r	12 (Scrolls Stop)	At this stop, the researcher asked you: “Which club _____”	Would you choose to join
scroll_spatial_r	12 (Scrolls Stop)	If you are facing the scroll that you pulled out, where was the window in reference to you? [Options: in front–to the left, in front–to the right, behind–to the left, behind–to the right]	In front–to the right
sig_elizabeth_r	13 (Signature Stop)	Type out the signature as you viewed it in the guestbook.	Elizabeth R
sig_crown_r	13 (Signature Stop)	What image was printed on the sheet of paper you drew on?	A crown
sig_spatial_r	13 (Signature Stop)	If you are facing the signature that you examined, where was the information desk in reference to you? [Options: behind–to the right, behind–to the left, same wall–to the left, same wall–to the right]	Same wall–to the right

Question Label	Block	Question Text	Restudy Answer
art_flower_r	14 (Art Gallery Stop)	What was the colour of the flower the little girl was holding?	White
art_artist_r	14 (Art Gallery Stop)	The artist of the mural was the first official Canadian _____.	War artist
tower_question_r.	15 (Tower Stop)	At this stop you were asked to guess the _____	Weight of the heaviest bell in the tower
tower_win_r	15 (Tower Stop)	How many large windows did you see on the tower?	Two
tower_spatial_r	15 (Tower Stop)	If you are facing the tower, where was the art gallery stop in reference to you? [Options: in front, left, right, behind]	Left

Tour Review Session Metacognition Questions

At the end of each review session, participants answered three metacognition questions. The first question probed the perceived degree of episodic reexperiencing (Figure S2), the second question probed participants' perceptions of how well they anticipate remembering the tour (Figure S3)**,** and the third question probed the perceived efficacy of the review session (Figure S4). Due to the ordinal nature of the dependent variable, cumulative link mixed models were fitted including fixed effects of group (restudy vs. retrieve; treatment coded) and session (three review sessions; linear and quadratic terms), as well as a random intercept for participant.

Reliving Ratings (Figure S2)

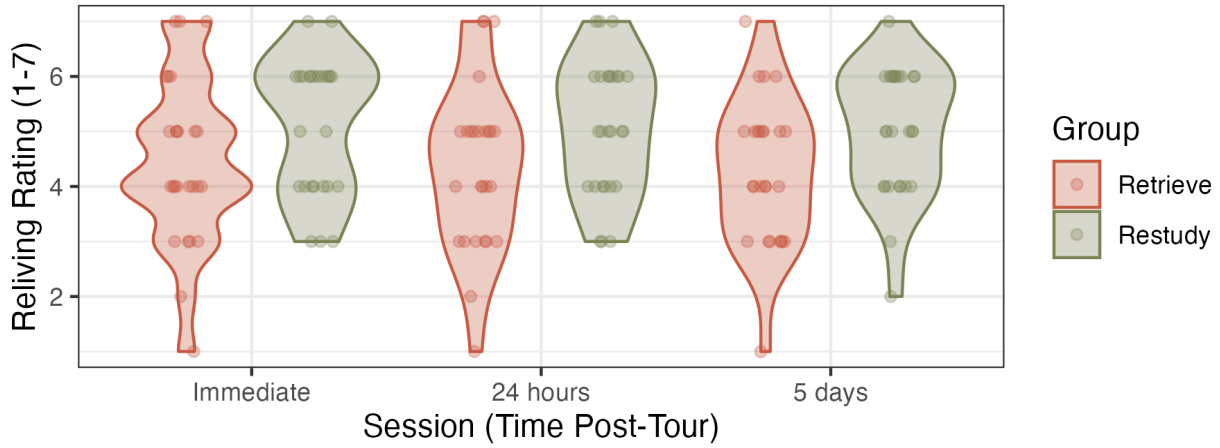

Figure S2. Self-reported episodic reexperiencing during tour review sessions. Question read: "To what extent did you feel like you were reliving the tour when completing this review session?" on a scale from 1, *an extremely small extent*, to 7, *an extremely large extent*. The fixed effect for the group variable was significant, with participants in the restudy group reporting higher reliving ratings, b = 1.68, SE = 0.79, z = 2.12, p = .034). However, the linear (b = -0.24, SE = 0.27, z = -0.88, p = 0.378) and quadratic (b = 0.01, SE = 0.27, z = 0.02, p = 0.985) session terms did not significantly predict reliving ratings.

Anticipated Memory Ratings (Figure S3)

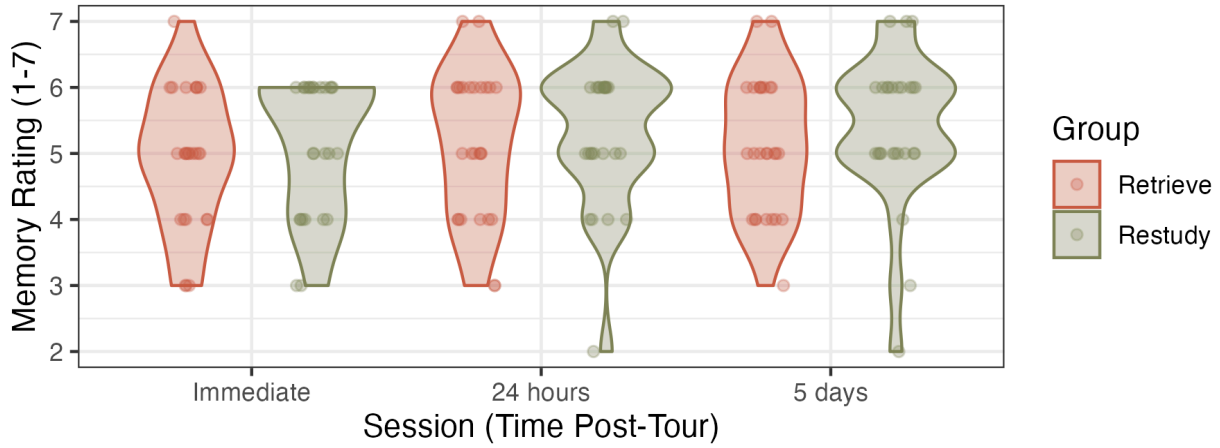

Figure S3. Anticipated subsequent memory for the tour for the tour as rated during tour review sessions. Question read: "Compared to a typical memory, how well do you think you will remember the tour?" on a scale from 1, *much worse*, to 7, *much better*. The fixed effect for the group variable was not significant, b = 0.68, SE = 0.89, z = 0.77, p = .442). The linear term was significant, b = 0.62, SE = 0.30, z = 2.09, p = .037, with anticipated memory rating increasing over time. The quadratic term was not significant, b = –0.28, SE = 0.29, z = –0.98, p = .327.

Perceived Efficacy Ratings (Figure S4)

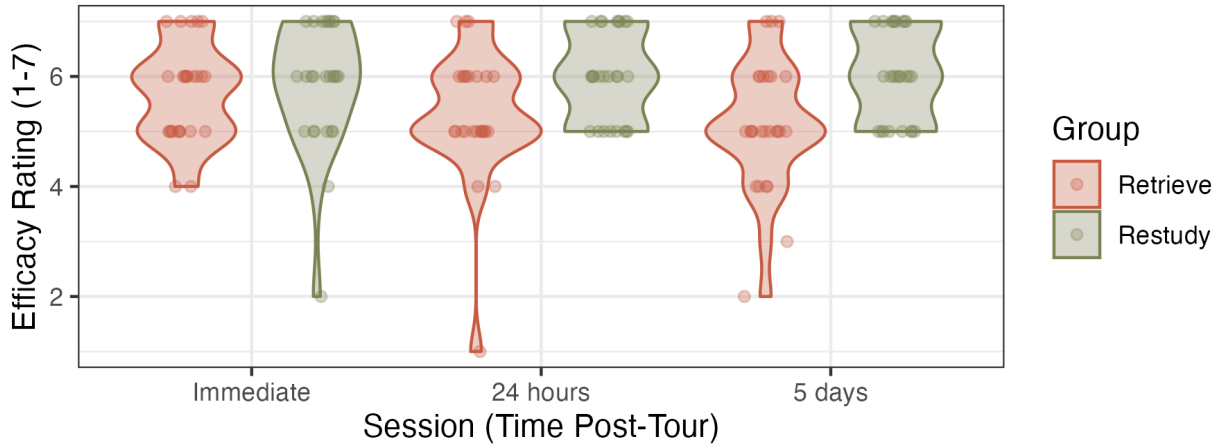

Figure S4. Perceived tour review session efficacy as reported during tour review sessions. Question read: "How effective do you think this review session was for improving your memory for the tour", on a scale from 1, *extremely ineffective*, to 7, *extremely effective*. The fixed effect for the group variable was significant, b = 1.60, SE = 0.61, z = 2.64, p = .008. The linear (b = –0.40, SE = 0.28, z = –1.40, p = .161) and quadratic (b = 0.06, SE = 0.28, z = 0.20, p = .40) session terms did not significantly predict efficacy ratings.

The Testing Effect Results (Figure S5)

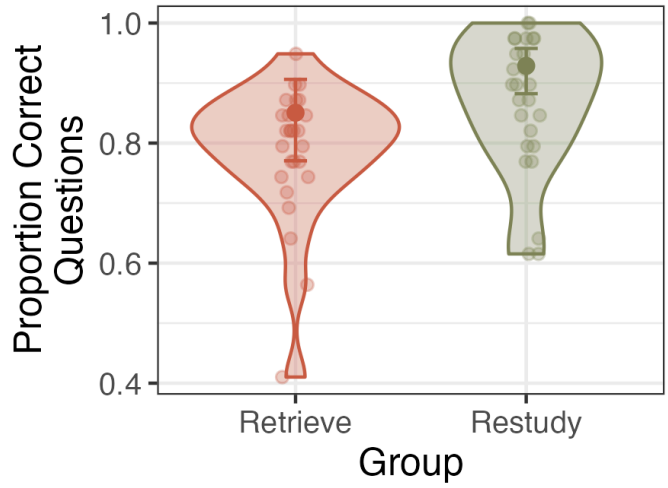

Figure S5. Raw proportion of testing effect (reviewed-tour-content) questions correct at the final assessment. A generalized mixed effects model with a binomial distribution and a logit link function was fitted to predict whether a testing effect question was correct or not, using group (treatment coded) as a fixed effect. Random intercepts were included for the crossed variables of question and participant (adjusted ICC for the model = 0.388). The restudy group was more likely to respond correctly to a given testing effect question, b = 0.83, 95% CI [0.29, 1.36], SE = 0.27, z = 3.04, p = .002. Thus, when the lack of feedback in the retrieve group was

not taken into account, the restudy group responded correctly to more reviewed-tour-content questions. Large dots represent model estimates with 95% confidence intervals

Integration with Pre-existing Semantic Knowledge Results

Number of External Semantic Details by Subtype (Figure S6)

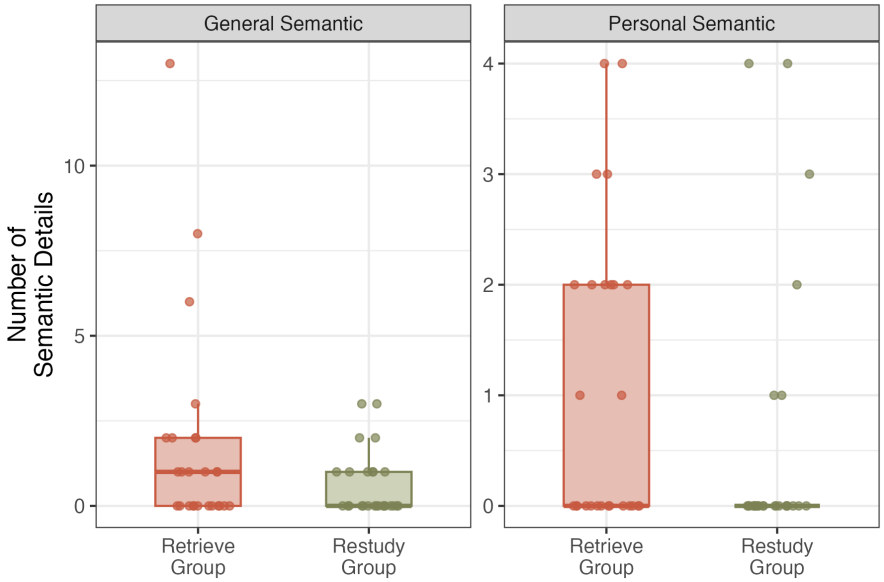

Figure S6. Number of external semantic details by subtype. External semantic details included in tour narratives were further classified according to their subtype as *personal* (i.e., tied to an individual's personal experiences, "I went to the University of X") or *general* (i.e., not personally tied to the individual, e.g., "the University of X is located in Germany") in accordance with Renoult et al., (2020). Visual inspection of the data suggested that the retrieve group included more semantic details for both subtypes.

Number of External Semantic vs. Review Session Question Performance at 12-18 Month Delay (Figure S7)

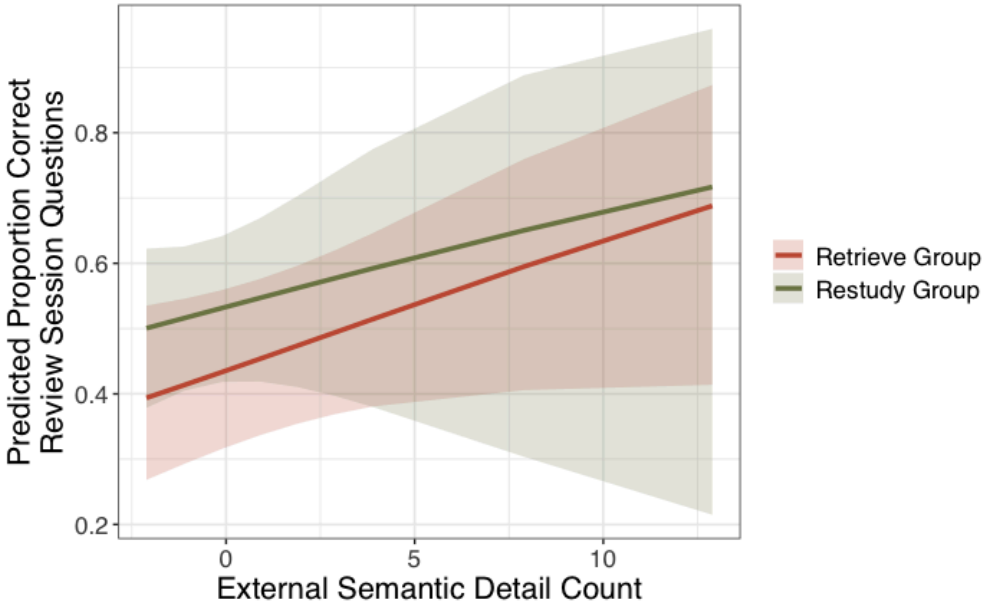

Figure S7. Exploratory analysis of semantic integration as a predictor of long-term retention of review session content (12–18 month delay). A generalized mixed-effects model (binomial family) tested whether integration with semantic knowledge (as indexed by external semantic detail count) predicted performance on review session questions after a 12–18 month delay (retrieve group n = 17; restudy group n = 21). External semantic detail count, group (treatment coded, restudy = 1, retrieve = 0), and their interaction were included as predictors, with random intercepts for participant and question. External detail count was a marginally significant predictor (b = 0.08, SE = 0.04, z = 1.81, p = .070), with a trend toward higher detail counts predicting better performance. Neither the effect of group (b = 0.39, SE = 0.25, z = 1.56, p = .118) nor the interaction (b = –0.02, SE = 0.09, z = –0.21, p = .831) was significant.

Broader Episodic Tour Memory Results

Unreviewed-Tour-Content Questions (Table S2)

Table S2. Proportion of participants who responded correctly to each unreviewed-tour-content question. Performance on unreviewed-tour-content questions was variable (see Table S2). Some of the worst-performing questions included incidental perceptual and spatial details, as well as certain facts presented in the audio guide about the building. Some of the best-performing questions included focal tour object details, tour activity and action details, and certain incidental spatial details.

Question Label	Block	Question Text	Restudy Group	Retrieve Group
researcher_cat_f	Researcher	What picture was on the researcher's tote bag?	0.12	0.15
researcher_clip_f	Researcher	What colour was the researcher's clipboard?	0.42	0.62
audioguide_how_f	Audioguide	How did you press the button to check the time?	0.42	0.38
audiotour_male_f	Audiotour	What was the name of the male tour guide from the tour audio device?	0.15	0.23
time_1_f	Checking Time Event #1	What was the time when you first checked it? (XX:XX PM/AM)	0.12	0.12
time_2_f	Checking Time Event #2	What was the time when you checked it the second time? (XX:XX PM/AM)	0.08	0.08
intro_location_f	Introduction Stop	Where was the photo located on the index card? (multiple choice)	0.46	0.50
intro_list_f	Introduction Stop	Besides the photo, what else was on the index card?	0.81	0.69
intro_colour_f	Introduction Stop	What colour was the background of the index card?	0.35	0.31
intro_mcphoto_f	Introduction Stop	Select the photo that was on the index card. (multiple choice)	0.88	0.73
intro_sport_f	Introduction Stop	What sport did the audio tour guides participate in?	0.27	0.50
coa_response_f	Coat of Arms Stop	At this stop, the researcher asked: “If you were to design your own coat of arms, what _____ would you include?” What was your response to the researcher’s question?	0.92	0.88
coa_gh_f	Coat of Arms Stop	What was the name of the room that was mentioned at this tour stop?	0.08	0.12

Question Label	Block	Question Text	Restudy Group	Retrieve Group
coa_largest_f	Coat of Arms Stop	The room that was mentioned at this tour stop has one of the largest collections of coats of arms in _____.	0.58	0.69
coa_war_f	Coat of Arms Stop	The coats of arms were painted in Hart House to commemorate _____.	0.15	0.00
coa_uni_f	Coat of Arms Stop	What was the name of the university for the coat of arms you examined?	0.73	0.58
coa_animal_f	Coat of Arms Stop	What animal was on the coat of arms you examined?	0.69	0.77
coa_crown_f	Coat of Arms Stop	Besides an animal and book(s), what other image was on the coat of arms you examined?	0.15	0.08
coa_gold_f	Coat of Arms Stop	The coat of arms you examined was painted with a layer of _____	0.23	0.08
coa_oak_f	Coat of Arms Stop	The coat of arms you examined was on a panel made of _____	0.31	0.27
coa_source_f	Coat of Arms Stop	Who narrated the COAT OF ARMS stop? (multiple choice)	0.54	0.54
founder_jfk_f	Founders Prayer Stop	Who was the notable visiting debater?	0.58	0.58
founder_state_f	Founders Prayer Stop	The visiting debater stated: “It’s a pleasure to be in a country where _____”	0.27	0.31
founder_topic_f	Founders Prayer Stop	You were asked if you would argue the affirmative or the opposition for a debate.What was the debate topic?	0.15	0.27
founder_deer_f	Founders Prayer Stop	What image was at the top of the Founders’ Prayer?	0.23	0.12

Question Label	Block	Question Text	Restudy Group	Retrieve Group
founder_tall_f	Founders Prayer Stop	Approximately how tall was the Founders' Prayer? Note: A baseball bat measures about 1 metre. (multiple choice)	0.50	0.65
founder_letter_f	Founders Prayer Stop	On the Founders' Prayer, the letter ___ looked like a different letter, exemplifying the gothic influence of the building.	0.46	0.46
founder_source_f	Founders Prayer Stop	Who narrated the Founders' Prayer stop? (multiple choice)	0.50	0.38
wall_ghost_f	Wall Fixture Stop	What did the researcher say might have saved Hart House from the fire?	0.65	0.46
wall_theatre_f	Wall Fixture Stop	Where was the fire at Hart House?	0.35	0.12
wall_fire_f	Wall Fixture Stop	What caused the fire at Hart House?	0.50	0.27
wall_time_f	Wall Fixture Stop	When was the fire at Hart House (i.e., what time of day)?	0.73	0.77
wall_use_f	Wall Fixture Stop	What were the wall fixtures used for?	0.65	0.77
wall_tall_f	Wall Fixture Stop	Approximately how tall was the wall fixture (what was the height)? Note: a pencil eraser is approximately 1 centimeter long. (multiple choice)	0.50	0.58
wall_door_f	Wall Fixture Stop	Approximately how far away was the wall fixture to the nearest doorway? Note: A baseball bat measures about 100 centimetres. (multiple choice)	0.15	0.12
wall_source_f	Wall Fixture Stop	Who narrated the wall fixture stop? (multiple choice)	0.46	0.42
glass_sun_f	Stained Glass Stop	What image was at the very top of the stained-glass window?	0.15	0.27

Question Label	Block	Question Text	Restudy Group	Retrieve Group
glass_loon_f	Stained Glass Stop	What did you look for on the stained-glass window?	0.65	0.73
glass_location_f	Stained Glass Stop	Where was the item you looked for located on the stained-glass window? (multiple choice)	0.77	0.85
glass_shape_f	Stained Glass Stop	What shape were the tops of the large stained-glass panels?	0.73	0.65
glass_church_f	Stained Glass Stop	The glass from the stained-glass windows came from destroyed _____	0.65	0.54
glass_seating_f	Stained Glass Stop	If you were facing the stained-glass window, where was the chapel seating in reference to you? (multiple choice)	0.81	0.92
glass_door_f	Stained Glass Stop	If you were facing the stained-glass window, where was the door to enter the chapel in reference to you? (multiple choice)	0.73	0.69
glass_source_f	Stained Glass Stop	Who narrated the stained-glass stop? (multiple choice)	0.65	0.38
radio_name_f	Radio Station Stop	The name of the radio station consisted of the letters CI__.	0.19	0.12
radio_num_f	Radio Station Stop	The name of the radio station consisted of the numbers __.5 FM.	0.08	0.04
radio_master_f	Radio Station Stop	The hip-hop show that was mentioned was called _____	0.00	0.00
radio_master1_f	Radio Station Stop	The hip-hop show that was mentioned is Canada's _____	0.08	0.15
radio_panels_f	Radio Station Stop	How many panels of glass did you see that led to the radio station studio?	0.15	0.08

Question Label	Block	Question Text	Restudy Group	Retrieve Group
radio_boarder_f	Radio Station Stop	What colour was the border around the glass panel(s) that led to the radio stationstudio?	0.23	0.19
radio_location_f	Radio Station Stop	If you were facing the radio station studio, where was the radio station sign located in reference to the panel(s) of glass leading to the radio studio? (multiple choice)	0.69	0.69
radio_outside_f	Radio Station Stop	If you were facing the radio station studio, where were the windows to the outside located in reference to you? (multiple choice)	0.85	0.69
radio_source_f	Radio Station Stop	Who narrated the radio station stop? (multiple choice)	0.73	0.46
elevator_interior_f	Elevator Stop	What did the walls inside the elevator look like?	0.31	0.62
elevator_h_f	Elevator Stop	What letter was repeated on the elevator doors?	0.73	0.62
elevator_deer_f	Elevator Stop	What picture was on the elevator doors?	0.27	0.31
elevator_theatre_f	Elevator Stop	The braille lines you read on the sign indicated the location of which specific room/place?	0.15	0.00
elevator_braille_f	Elevator Stop	The room/place that you read about on the braille sign was located on which floor?	0.00	0.00
elevator_design_f	Elevator Stop	Why did the elevator take so long to build?	0.65	0.54
elevator_year_f	Elevator Stop	What year did the elevator open?	0.08	0.15
elevator_access_f	Elevator Stop	What part of Hart House was completely inaccessible without using stairs (before the elevator was installed)?	0.15	0.12
elevator_source_f	Elevator Stop	Who narrated the elevator stop? (multiple choice)	0.42	0.65

Question Label	Block	Question Text	Restudy Group	Retrieve Group
rest_blue_f	Restaurant Stop	What colour was the back wall of the restaurant?	0.27	0.54
rest_reno_f	Restaurant Stop	In 2018, the restaurant was _____	0.65	0.62
rest_name_f	Restaurant Stop	What was the name of the restaurant?	0.00	0.04
rest_restrict_f	Restaurant Stop	Besides the timing restriction, what other restriction did women have to follow at the restaurant when it opened?	0.31	0.35
rest_eat_f	Restaurant Stop	At this tour stop, you were asked what you wanted to _____	0.73	0.81
rest_response_f	Restaurant Stop	What was your response to the question the researcher asked you?	0.85	0.85
rest_source_f	Restaurant Stop	Who narrated the restaurant stop? (multiple choice)	0.27	0.46
water_pen_f	Water Fountain Event	What did the researcher do at the water fountain?	0.69	0.77
water_atm_f	Water Fountain Event	What type of machine was next to the water fountain?	0.08	0.19
water_rbc_f	Water Fountain Event	What company owned the machine next to the water fountain?	0.00	0.08
water_fountains_f	Water Fountain Event	How many water fountains were there?	0.54	0.54
fitness_sign_f	Fitness Center Stop	What was the main background colour of the sign that you read (excluding the photo)?	0.35	0.27
fitness_sport_f	Fitness Center Stop	What sport was the person you read about doing when they got kicked out of Hart House?	0.38	0.46
fitness_dress_f	Fitness Center Stop	How did the person that you read about join the boys' sports team?	0.73	0.65
fitness_name_f	Fitness Center Stop	What was the first name of the person you read about?	0.15	0.08

Question Label	Block	Question Text	Restudy Group	Retrieve Group
fitness_season_f	Fitness Center Stop	During what season did the person you read about get kicked out of Hart House?	0.58	0.54
fitness_mcphoto_f	Fitness Center Stop	Who is the person that you read about (in the photo that you examined)? (multiple choice)	0.88	1.00
fitness_source_f	Fitness Center Stop	Who narrated the fitness centre stop? (multiple choice)	0.73	0.77
scroll_deer_f	Scrolls Stop	What image was at the top right of the scroll?	0.04	0.08
scroll_location_f	Scrolls Stop	Describe where the scroll you examined was located in the display case.	0.81	0.58
scroll_colour_f	Scrolls Stop	What was the colour of the first letter of each word on the scrolls?	0.27	0.19
scroll_listed_f	Scrolls Stop	What was the first organization listed on the scroll?	0.04	0.00
scroll_chess_f	Scrolls Stop	The Chess Club is the _____ chess club in Canada.	0.58	0.77
scroll_response_f	Scrolls Stop	What was your response to the researcher's question?	0.88	0.96
scroll_source_f	Scrolls Stop	Who narrated the scrolls stop? (multiple choice)	0.54	0.38
sig_loc_f	Signature Stop	Where was the signature that you viewed located in the podium? (multiple choice)	0.31	0.54
sig_letter_f	Signature Stop	What did the last letter in the signature stand for?	0.58	0.54
sig_lunch_f	Signature Stop	Regarding the person who signed the book, what did they do at Hart House?	0.31	0.23
sig_drew_f	Signature Stop	Type out what you drew at this tour stop.	0.65	0.35
sig_inside_f	Signature Stop	In relation to the image printed on the sheet, where did you draw? (multiple choice)	0.81	0.69

Question Label	Block	Question Text	Restudy Group	Retrieve Group
sig_er_f	Signature Stop	Type out the first letter that you saw on the monogram.	0.62	0.73
sig_image_f	Signature Stop	What image was at the top of the monogram that you saw?	0.38	0.19
sig_mcmono_f	Signature Stop	Which monogram did you look at? (multiple choice)	0.85	0.96
sig_source_f	Signature Stop	Who narrated the signature stop? (multiple choice)	0.31	0.46
art_dress_f	Art Gallery Stop	What was the colour of the dress that the woman holding the little girl was wearing?	0.12	0.27
art_angel_f	Art Gallery Stop	What was the image on the left side of the mural?	0.38	0.35
art_will_f	Art Gallery Stop	What was the first name of the mural artist?	0.04	0.04
art_war_f	Art Gallery Stop	What war did the artist of the mural fight in?	0.50	0.42
art_emotion_f	Art Gallery Stop	At this stop you were asked: “What is the first _____ that comes to mind?”	0.69	0.35
art_response_f	Art Gallery Stop	At this stop you were asked: “What is the first _____ that comes to mind?” What was your response to this question?	0.58	0.35
art_source_f	Art Gallery Stop	Who narrated the art gallery stop? (multiple choice)	0.46	0.42
tower_guess_f	Tower Stop	What was your response/guess to the question that the researcher asked you?	0.58	0.54
tower_answer_f	Tower Stop	What was the correct answer to the question that the researcher asked you?	0.38	0.42
tower_represent_f	Tower Stop	What did each bell in the tower represent?	0.27	0.27
tower_name_f	Tower Stop	What was the name of the tower?	0.04	0.15

Question Label	Block	Question Text	Restudy Group	Retrieve Group
tower_clock_f	Tower Stop	Other than window(s), what other large item decorated the face of the tower?	0.50	0.31
tower_arch_f	Tower Stop	What was the shape of the walkway under the tower?	0.62	0.85
tower_weather_f	Tower Stop	What was the weather like while you were looking at the tower?	0.73	0.85
tower_source_f	Tower Stop	Who narrated the Tower stop? (multiple choice)	0.50	0.69

Internal (Tour-Specific) Details from Tour Narratives

Internal Detail Subtypes (Figure S8)

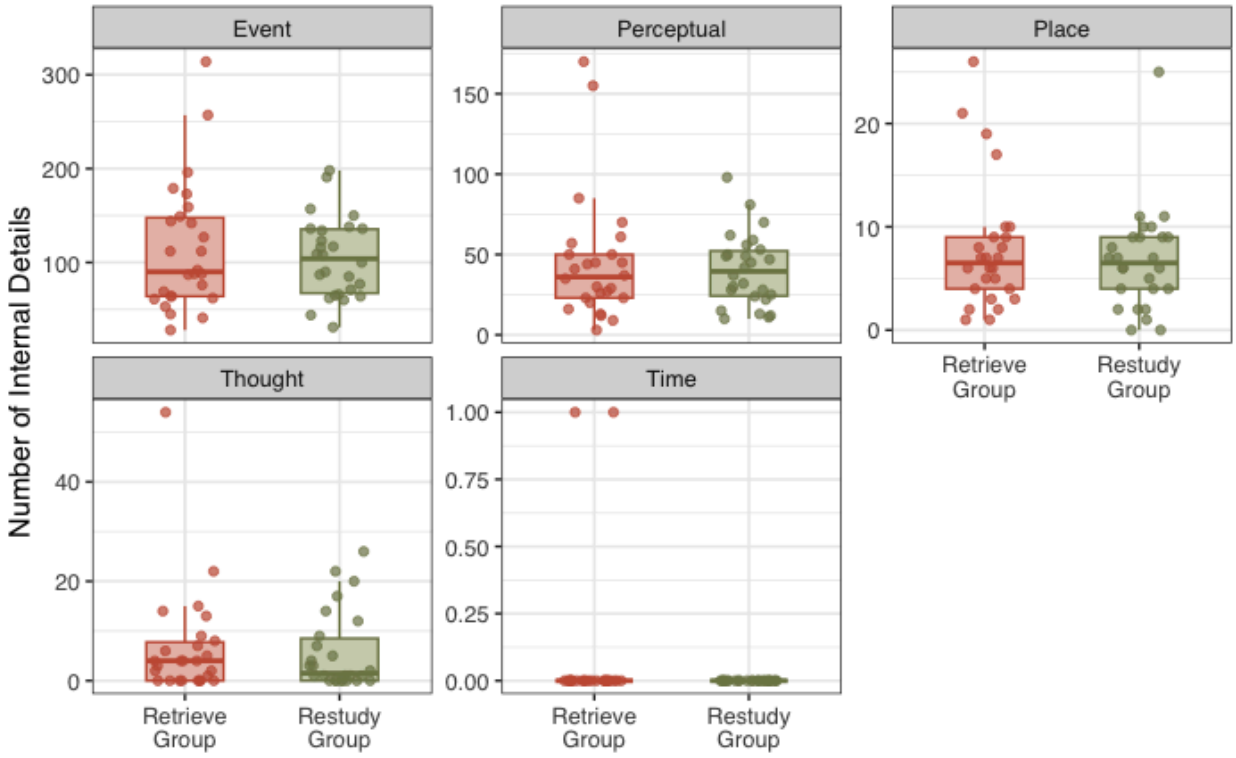

Figure S8. Number of internal details included in tour narratives by subtype. Internal detail subtype (e.g., event, place, time, perceptual, and emotion/thought) was not included as a factor in the model since we had no theory-driven predictions suggesting that group differences would vary by internal detail subtype. Consequently, including it as a factor was deemed unnecessary and would have complicated the model, risking overfitting without adding substantial value. A

visual inspection of the data indicated substantial variation in the number of internal details by subtype, yet the retrieve and restudy groups showed comparable performance across subtypes.

Internal Detail Verifiability (Figure S9)

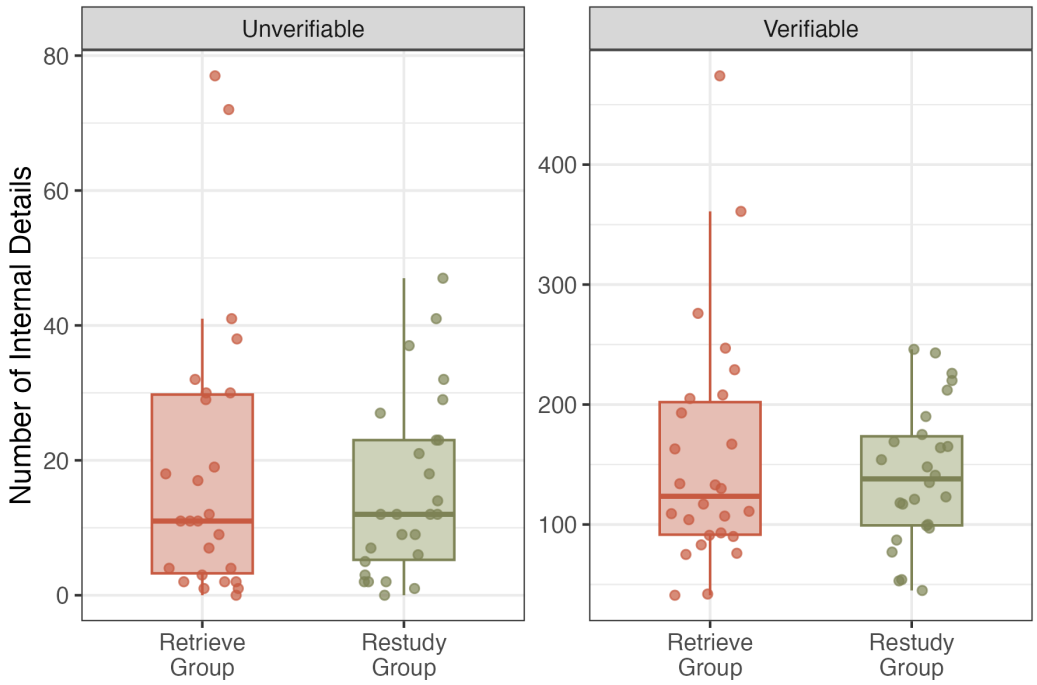

Figure S9. Number of verifiable and unverifiable internal details included in tour narratives. A visual inspection of the data indicated substantial variation in the number of internal details that were verifiable and unverifiable, such that participants primarily included verifiable tour details), yet the retrieve and restudy groups showed comparable performance across subtypes.

Internal Detail Reviewed vs. Unreviewed (Figure S10)

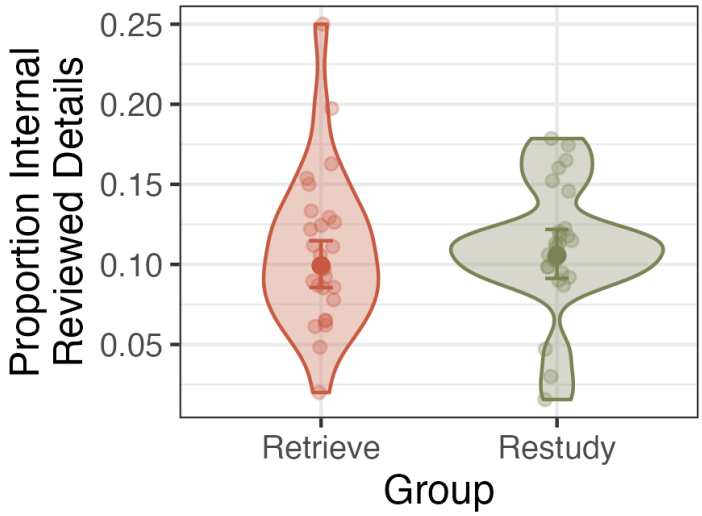

Figure S10. Proportion of internal details that were reviewed during tour review sessions. A generalized mixed model (binomial distribution) predicted whether an internal detail was classified as reviewed or unreviewed during tour review sessions, with group (treatment coded) as a fixed effect and a random intercept for participant (adjusted ICC = 0.030). The group effect was non-significant, b = –0.07, 95% CI [–0.30, 0.16], SE = 0.12, z = –0.59, p = 0.554, indicating that retrieval did not bias narratives toward reviewed content.

Error Subtypes (Figure S11)

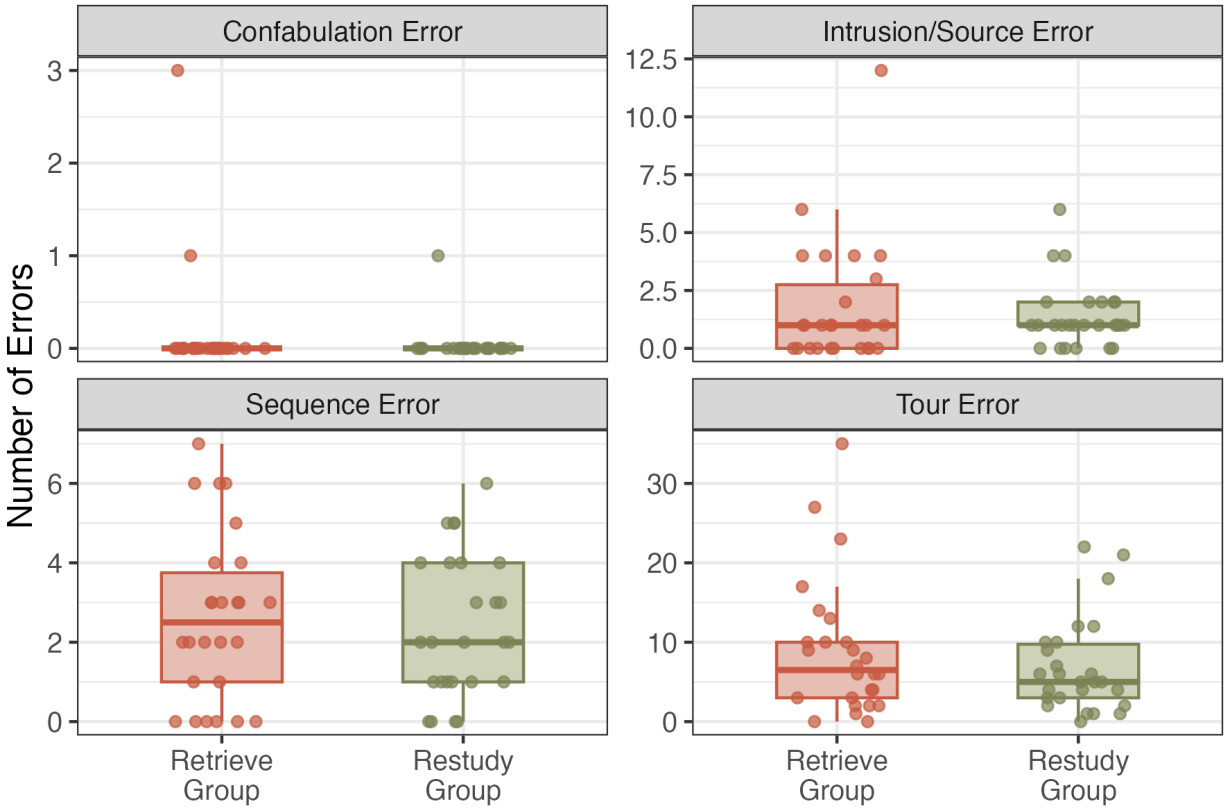

Figure S11. Number of errors included in tour narratives by subtype. If details were classified as incorrect, they were further categorized into one of five error types: *tour content errors* (incorrect details related to tour content), *intrusion/source errors* (incorrectly including elements from one tour event into the description of another event or misattributing the source of a piece of information), *confabulatory errors* (mentions of entire actions, objects, or facts that were false and unrelated to the tour content), *sequence errors* (explicit violations of the tour sequence), and *other errors* (any errors that did not fit into the aforementioned categories). A visual inspection of the data indicated variation in the number of errors in tour narratives by subtype, yet the retrieve and restudy groups showed comparable performance across subtypes ("other" errors are not depicted because none were scored).

Temporal Organization in Tour Narratives

Lag-Conditional Response Probability Curves (Figure S12)

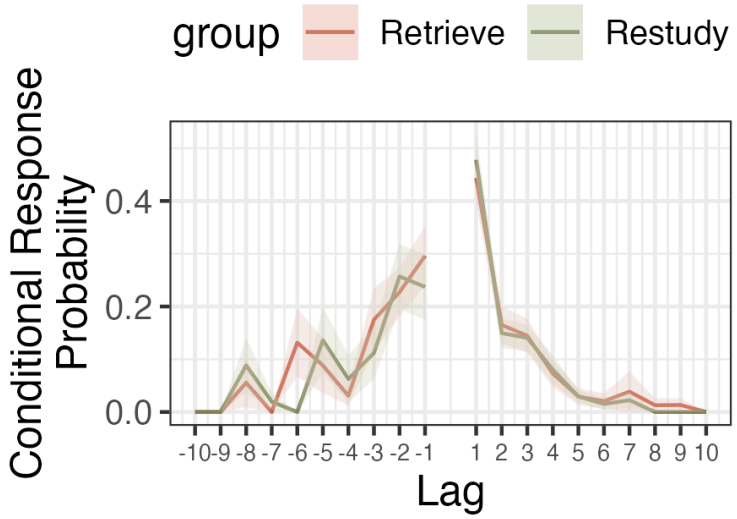

Figure S12. Lag conditional response probability curves. To further examine temporal organization, we computed temporal lag-conditional response probability, computed using the Psifr package (Morton, 2020). This metric quantifies the likelihood of transitioning between tour events based on time lag and relative order (e.g., a lag of +1 indicates a forward transition to the next tour stop in the original sequence). Visual inspection of the data demonstrated a typical temporal contiguity pattern for both groups, with participants preferentially making a transition to the next ordinal item in the forward direction. Slopes were fitted to the first five lags of each arm of the lag conditional response probability (lag-CRP) curve (i.e., in the positive and negative direction) for each participant (i.e., two slopes were fitted for each participant; as in Diamond & Levine, 2020). If a participant had less than two data points for an arm, these participants were excluded from the analysis, as slopes could not be estimated (this led to the exclusion of two participants in the retrieve group and two participants in the restudy group). Slope values were then submitted to a linear model using group (treatment coded) and curve direction (i.e., positive vs. negative lags) as fixed effects. A random intercept was initially included for participant, however the model did not converge due to low variance for this random effect so it was removed. The effect of group was not significant, b = –0.01, 95% CI [–0.08, 0.06], SE = 0.03, t = –0.31, p = 0.756. The effect of lag-CRP curve direction was significant, b = –0.15, 95% CI [–0.22, 0.08], SE = 0.03, t = –4.43, p < .001, with a more negative (i.e., extreme) slope for positive lags. The interaction between lag-CRP curve direction and group was not significant, b = 0.01, 95% CI [–0.09, 0.10], SE = 0.05, t = 0.13, p = 0.897.

Explicit Temporal Sequencing Task (Figure S13)

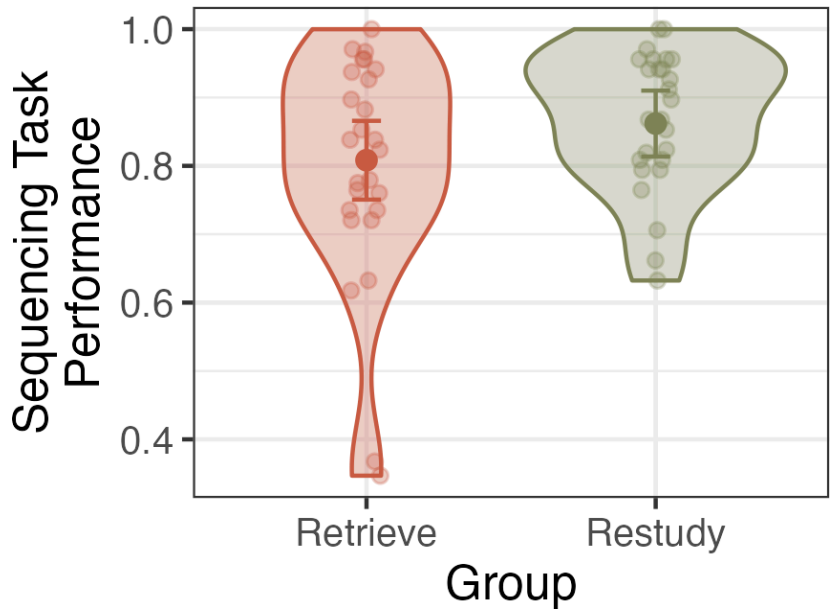

Figure S13. Explicit temporal sequencing task. Participants ordered tour stops by entering a number next to a label representing each tour event (events were presented in random order). For each participant, the correlation (Kendall's) between the actual stop order and their reported stop order was computed as a measure of performance (1 would indicate that they recalled perfectly). Then, these values were submitted to a generalized regression model with a beta distribution (logit link function) predicted performance using group (effect coded) as a fixed effect. Any values of 1 were altered to 0.9999. The group effect was non-significant, b = 0.39, 95% CI [–0.11, 0.90], SE = 0.26, z = 1.53, p = 0.127.

Subjective Memory Phenomenology (Table S3)

Table S3. Model results for memory phenomenology subscales. Linear mixed effects models were fitted to predict mean Memory Experiences Questionnaire–Short Form (MEQ-SF) subscale scores, using group (treatment coded) as a fixed effect, for the relevant categories of accessibility, coherence, emotional intensity, sensory details, time perspective, valence, vividness, and visual perspective. No models were significant (ps > .083).

Subscale	Estimate (b)	SE	df	t	p	BF_{01}
Accessibility	–0.49	0.28	50	–1.77	.083	0.89
Coherence	–0.14	0.24	50	–0.59	.558	2.64
Emotional intensity	0.03	0.26	50	0.10	.922	2.86
Sensory	0.23	0.18	50	1.27	.209	1.95

details						
Time perspective	0.06	0.22	50	0.29	.773	3.13
Valence	–0.04	0.17	50	–0.23	.823	3.90
Vividness	–0.22	0.23	50	–0.94	.350	2.21
Visual perspective	0.29	0.25	50	1.18	.244	1.76

Final Assessment Metacognition Question Results

Perceived Review Session Efficacy (Figure S14)

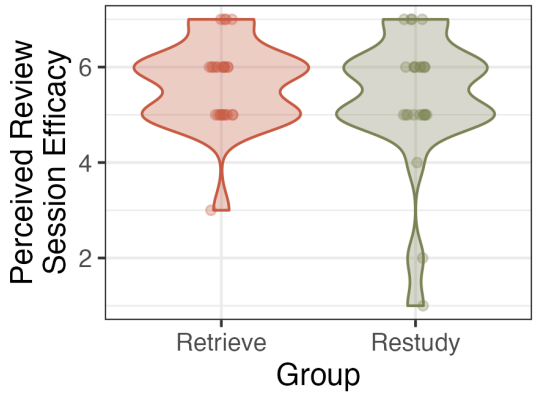

Figure S14. Perceived review session efficacy at the final assessment. Participants were asked, "How effective do you think the review sessions were for improving your memory of the Hart House Tour? (1 = extremely ineffective, 7 = extremely effective)." Due to the ordinal nature of the dependent variable, cumulative link models were fitted including group (retrieve vs. restudy; treatment coded) as a fixed effect. There was no evidence for a difference between groups, *b* = -0.33, SE = 0.51, z = -0.64, *p* = .523.

Confidence in Tour Memory Accuracy (Figure S15)

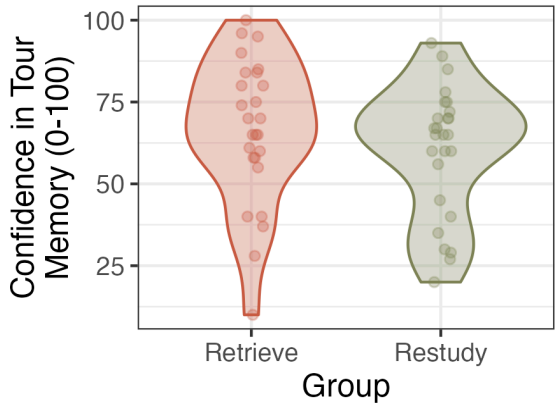

Figure S15. Confidence in tour memory accuracy at the final assessment. Participants were asked, "How confident are you in the accuracy of the memory that you just recounted to the researcher? (0% = not at all confident; 100% = completely confident)." A linear model was fitted with group (retrieve vs. restudy; treatment coded) included as a fixed effect. There was no evidence for a difference between groups, b = –6.04, SE = 5.80, t(50) = –1.04, p = .303.

Supplemental Material References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

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

Cheng, C.-W., Hung, Y.-C., & Balakrishnan, N. (2024). *rBeta2009: The Beta Random Number and Dirichlet Random Vector Generating Functions*. https://doi.org/10.32614/CRAN.package.rBeta2009

Christensen, R. H. B. (2023). *ordinal—Regression Models for Ordinal Data*. https://CRAN.R-project.org/package=ordinal

Cribari-Neto, F., & Zeileis, A. (2010). Beta Regression in R. *Journal of Statistical Software*, *34*(2), 1–24. https://doi.org/10.18637/jss.v034.i02

DeBruine, L. (2025). *faux: Simulation for Factorial Designs*. Zenodo. https://doi.org/10.5281/zenodo.2669586

Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus: An R Package for Fitting Distributions. *Journal of Statistical Software*, *64*(4), 1–34. https://doi.org/10.18637/jss.v064.i04

Diamond, N. B., Armson, M. J., & Levine, B. (2020). The truth is out there: Accuracy in recall of verifiable real-world events. *Psychological Science*, *31*(12), 1544–1556. https://doi.org/10.1177/0956797620954812

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.

Lenth, R. V., Bolker, B., Buerkner, P., Giné-Vázquez, I., Herve, M., Jung, M., Love, J., Miguez, F., Riebl, H., & Singmann, H. (2023). *emmeans: Estimated marginal means, aka least-squares means* (Version 1.8.5) [Computer software]. https://CRAN.R-project.org/package=emmeans

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

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, *6*(60), 3139. https://doi.org/10.21105/joss.03139

Mahendran, A., & Wijekoon, P. (2024). *fitODBOD: Fitting Discrete Univariate Binomial Mixture Distributions*. https://doi.org/10.32614/CRAN.package.fitODBOD

Morton, N. W. (2020). Psifr: Analysis and visualization of free recall data. *Journal of Open Source Software*, *5*(54), 2669. https://doi.org/10.21105/joss.02669

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Renoult, L., Armson, M. J., Diamond, N. B., Fan, C. L., Jeyakumar, N., Levesque, L., Oliva, L., McKinnon, M., Papadopoulos, A., Selarka, D., St Jacques, P. L., & Levine, B. (2020). Classification of general and personal semantic details in the Autobiographical Interview. *Neuropsychologia*, *144*, Article 107501. https://doi.org/10.1016/j.neuropsychologia.2020.107501

Revelle, W. (2024). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University. https://CRAN.R-project.org/package=psych

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth). Springer. https://www.stats.ox.ac.uk/pub/MASS4/