Journal of Memory and Language 152 (2027) 104814

Contents lists available at ScienceDirect

Journal of Memory and Language

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

Tell me what you read, and I will tell you what you remember: Evidence for personalized embeddings in memory modelling

Dominic Guitard [a,*,1], Randall K. Jamieson [b], Jean Saint-Aubin [c], Brendan T. Johns [d]

[a] School of Psychology, Cardiff University, United Kingdom
[b] Department of Psychology, University of Manitoba, United Kingdom
[c] École de psychologie, Université de Moncton, United Kingdom
[d] Department of Psychology, McGill University, United Kingdom

ARTICLE INFO

Keywords:
Cued recall
Distributional semantics
Individual differences
Instance-based memory models
Reading experience

ABSTRACT

Computational models of memory have achieved considerable success by formalizing how traces are encoded, cued, and retrieved. However, the role that individual linguistic experience plays in shaping representational structure has been relatively understudied. Where the issue has been examined, most models assume that a common population-level semantic space suffices, effectively treating individual variation in language experience as noise rather than signal. We test the theoretical adequacy of this assumption. In a large-scale experiment ($N = 478$), participants completed a cued recall task in which identical target words were paired with cues drawn from eight different genre-specific semantic spaces (mystery, fantasy, horror, literary fiction, romance, science fiction, thriller, and non-fiction). Participants reported their reading habits across the eight genres, and personalized semantic representations were constructed by weighting genre-specific corpora proportionally to each participant's reading profile. Two complementary analyses were conducted. In a model-free analysis, cosine similarity computed within each participant's personalized semantic space predicted recall and omission outcomes more reliably than similarity derived from generic semantic spaces. In a computational analysis, substituting personalized representations into the embedded Computational Framework of Memory improved cue–target-level predictions over generic alternatives, accounting for approximately 4 percentage points more explained variance. These findings provide proof of concept that variation in individual linguistic experience shapes semantic structure in ways that are both behaviourally detectable and computationally tractable. More broadly, they suggest that the predictive limits of memory models may reside not only in their retrieval mechanisms but also in the fidelity of the representations they assume.

Introduction

Computational models of memory have made remarkable progress over the past four decades. Across paradigms ranging from serial recall to cued recall and recognition memory, these models can account for a long list of benchmark phenomena by specifying how traces are formed, cued, and reconstructed (e.g., Hintzman, 1986, 1988; Howard & Kahana, 2002; Polyn et al., 2009; Raaijmakers & Shiffrin, 1981; Shiffrin & Steyvers, 1997). One dimension of memory, however, has received comparatively less investigation: the contribution of individual experience to the representational structure over which encoding, storage, and retrieval operate. This omission matters. A foundational principle in cognitive science holds that intelligent behaviour cannot be understood by examining internal processes alone but emerges from the interaction between a processing system and the structure of its environment (Johns et al., 2023; Simon, 1956, 1990). Chase and Simon's (1973) work on chess memory provides a vivid illustration: experts remembered briefly presented board positions better than novices not because they possessed a larger general-purpose memory store, but because chess experience had given them larger, meaningful chunks over which memory could operate. The representational structure of experience was doing the work. If this principle applies to memory in general, then models that treat all individuals as sharing the same representational space may leave meaningful individual variation in

* Corresponding author at: School of Psychology, Cardiff University, Tower Building, 70 Park Place, Cardiff CF10 3AT, United Kingdom.
E-mail address: guitardd@cardiff.ac.uk (D. Guitard).
[1] https://orcid.org/0000-0002-4658-3585

https://doi.org/10.1016/j.jml.2026.104814
Received 30 May 2026; Received in revised form 1 August 2026; Accepted 10 September 2026

behaviour unaccounted for, misattributed to shortcomings in our accounts of how memory operates rather than the knowledge on which those accounts are applied.

People do not acquire language and knowledge from identical experiential environments. A reader of fantasy fiction encounters words such as *dragon*, *falcon*, and *sorcery* in rich, genre-specific contexts that differ from those experienced by a reader of crime fiction where words such as *evidence*, *motive*, and *forensic* carry different experientially-specific meanings. According to distributional theories of meaning, words experienced in different co-occurrence environments come to occupy different relations in semantic space (Günther et al., 2019; Harris, 1954; Landauer & Dumais, 1997). Recent work has put this prediction on firmer quantitative footing. For example, Johns (2024) constructed personalized semantic representations from the language production histories of more than 500 individuals and showed measurable differences in word meanings across users. Similarly, Wulff et al. (2019) showed that distributional representations are updated systematically across the lifespan to track historical language change. The implication for memory research is direct. Individuals with different linguistic histories carry different semantic geometries into the laboratory; if retrieval depends on semantic similarity, those geometries would produce different patterns of remembering (and misremembering).

However, existing models supply all simulated participants with the same population-level representational space, and the question of whether retrieval mechanisms are limited by representation is overshadowed. A retrieval process can be formally well specified yet empirically inadequate if the similarity relations it refers to do not correspond to an individual's knowledge. This is not a measurement concern; it is a theoretical one. Memory models are, implicitly, theories of representation: they specify what is stored, how similarity between cue and target is computed, and how prior knowledge constrains the reconstruction of the past (Hintzman, 1986; Kahana, 2020; Nosofsky, 1986). Population-level approximations of semantic space may, therefore, obscure the very individual variation that drives item-level differences in recall (Cowan et al., 2024; Healey & Kahana, 2014; Unsworth et al., 2019). Fortunately, the long history of work and deep databases of measurements on memory for words presents a tractable basis on which to examine the issue.

This prediction, though theoretically natural, is often overlooked and thus under-formalised in computational models of memory. Population-level and individualized representations are best understood not as competing approximations of the same target but as descriptions pitched at different levels and, thus, suited to different questions. When the goal is to characterize shared structure in semantic memory or to predict average performance across participants, a population-level space is appropriate and has proven highly effective. When the goal is to explain systematic deviations from that shared structure, why a given person remembers this pair and forgets that one, the relevant variation is precisely what the population-level space averages away.

Decades of work in computational linguistics, cognitive psychology, and machine learning have established that semantic structure can be learned from distributional experience with language. Across multiple formal approaches, distributional semantic models have demonstrated an impressive capacity to construct word meaning representations from lexical experience (Griffiths et al., 2007; Jones & Mewhort, 2007; Landauer & Dumais, 1997; Mikolov et al., 2013; Pennington et al., 2014). These models differ considerably in their architectures, from matrix factorisation methods to predictive neural embeddings to holographic distributed representations. Yet, they converge on a single conclusion: the geometry of semantic space reflects a distilled record of experience with language, and the nature of that experience shapes the structure of the resulting space.

In memory modelling, this insight has yet to be fully integrated. Most models are supplied with a single semantic space derived from a broad general corpus, which is treated as a universal approximation of what every individual knows. In some cases, the simplification is taken further with lexical items assigned random vectors that sever any connection between structured representational geometry and the structure of natural language. Such choices carry theoretical commitments. Random representations impose similarity structures that depart systematically from those emerging from natural language experience (Johns & Jones, 2010), and corpus-based spaces derived from general text serve to average over genre and author-level variation. A growing body of work documents that linguistic experience varies meaningfully across individuals, genres, authors, and sociolinguistic communities, and that these differences leave detectable traces in cognition and language processing (Johns, 2021, 2024; Johns et al., 2019; Johns & Jamieson, 2018, 2019; Mar & Rain, 2015; Martí et al., 2023; Thompson et al., 2020; Wang & Bi, 2021). The implication for memory modelling is direct: if individual semantic representations diverge as a function of linguistic experience, predictions derived from a shared population-level space will capture what is common across individuals but systematically miss the variation that drives person specific differences in item-level remembering.

This reflects a structural asymmetry in what aggregation does to the two components of a memory model. The process architecture, how traces are encoded, cued, and reconstructed, is assumed to be shared, and averaging across individuals therefore preserves it; this is why group-level data have been such a productive instrument for investigating fundamental memory mechanisms. Representational structure is acquired individually, and averaging cancels it. Because many different individual configurations are consistent with the same aggregate mean, the group level constrains the individual only weakly, even as the individuals fully determine the group (Ashby et al., 1994; Estes, 1956). Population-level analyses are thus well matched to the invariant component of the system and structurally uninformative about the variable one. Chase and Simon's (1973) chess players make the same point concretely: what distinguished experts from novices was not the memory system operating over the board but the representational structure that experience had built for them to operate over.

Two recent theoretical developments make a personalized-embeddings test of these ideas timely. First, instance theory has been articulated as a domain-general framework for cognitive psychology (Jamieson et al., 2018, 2022). The embedded Computational Framework of Memory (eCFM; Guitard et al., 2025a, 2025b, Guitard et al., 2026) realizes this proposition by combining an instance-based process model for encoding, storage, and retrieval from MINERVA 2 (Hintzman, 1986, 1988) with structured semantic representations derived from distributional models. In doing so, it can make item-level predictions about words that people do and do not remember, as well as words they misremember. Second, item-level prediction has emerged as a meaningful evaluative criterion for memory models, distinct from aggregate fit. This criterion has been taken up across a range of modelling frameworks and paradigms (e.g., Chang et al., 2025; Cox et al., 2018; Cox, 2026; Gatti & Günther, 2026; Guitard et al., 2025a, 2025b, Guitard et al., 2026; Healey & Kahana, 2014; Osth et al., 2026; Petilli et al., 2026; Reid & Jamieson, 2023; Reid et al., 2026; Unsworth, 2019; Unsworth et al., 2019), and work has demonstrated that representational quality is a primary determinant of how well a model captures item-by-item variation. The natural next question is whether the same representational machinery, when individualized, can improve prediction of memory outcomes at the cue–target level, capturing not just what is remembered on average across people, but how patterns of remembering and forgetting vary based on what participants know based on their reading habits.

The present study presents a targeted proof of concept directed at this question. Although we cannot provide a complete account of how a lifetime of experience shapes semantic knowledge because personal histories must be guessed to some degree, we use reading-genre as a tractable and theoretically motivated clue for variation in linguistic experience (Acheson et al., 2008; Long & Freed, 2021; Lowder & Gordon, 2017; Schwering et al., 2021; Stanovich & West, 1989;

Vermeiren et al., 2023). To maximize the chance of detecting a personalization effect, we constructed stimulus materials spanning eight literary genres and derived personalized semantic embeddings by weighting genre-specific corpora in proportion to each participant's reported reading habits (Johns et al., 2019; Johns & Jamieson, 2018, 2019). Although the strategy involves some guesswork about people's language exposure, if even this coarse approximation of variation in individual language experience yields measurable gains in predictive accuracy, it provides principled motivation to invest in finer-grained characterizations of individual representational structure in future work on what people remember (and misremember).

To examine the issue, we apply two complementary strategies. Demonstration 1 is a model-free statistical test: we ask whether memory outcomes measured in the context of cued recall are better predicted when cue–target similarity is computed using personalized rather than generic semantic spaces. Demonstration 2 is a computational test: we ask whether personalized embeddings can be substituted into the eCFM and whether doing so improves cue–target level prediction (i.e., which pairs people remember better and worse conditional on their language experience and corresponding semantic spaces). Together, the two demonstrations advance a shared argument. The behaviour and accuracy of process models of memory is bounded not only by their accounts of encoding, storage, and retrieval but also by the specificity of the representational spaces over which those mechanisms operate. Thus, refining those spaces by grounding them in individual experience is not a peripheral improvement but, rather, a core theoretical priority.

General method

We collected data from a large sample of online participants to maximize the variability of reading histories represented in the sample. Each participant completed a reading-experience questionnaire and an online cued recall task in which the targets were held constant across participants, but the cues were deliberately selected based on word similarities derived from genre-specific semantic spaces. The reading-experience questionnaire was used to construct a personalized semantic space for each participant. The two demonstrations that follow use the strategy to test whether memory performance is better predicted when cue–target similarity is quantified from participant-specific spaces (i.e., personalized) rather than a shared generic one.

Ethical approval

All procedures were reviewed and approved by the School of Psychology Ethics Committee at Cardiff University (EC.22.09.20.6627G). Participants provided informed consent prior to participation and received a debrief upon completion of the study.

Data availability

The data, codes, and materials reported in our demonstrations are openly available on the Open Science Framework at OSF.

Participants

A total of 478 individuals took part in the experiment. Participants were recruited from two sources: 38 undergraduate students from Cardiff University who received course credit for their participation and 440 participants via Prolific (248 based in the United Kingdom and 192 in the United States) who were compensated at a rate of £9.00 per hour on a prorated basis. The sample had a mean age of 25.10 years (SD = 3.80). Of the 478 participants, 177 identified as male and 301 as female.

Eligibility was determined using the following criteria: native proficiency in English; residence in, or citizenship of, the United Kingdom or the United States; no self-reported history of language-related disorders, cognitive impairments, dementia, or literacy difficulties; age between 18 and 30 years; and normal or corrected-to-normal vision. Participants recruited via Prolific were additionally required to have a prior approval rate (i.e., an indirect index of participant quality) of 90% or better on the platform, as calculated by Prolific based on their history of prior study completions. All remaining eligibility criteria were assessed via self-report.

Materials

The experimental materials consisted of 144 cued recall pairs. Each pair comprised a target word coupled with a genre-specific cue word (Mystery, Fantasy, Horror, Literary Fiction, Romance, Science Fiction, Thriller, and Non-fiction). Every participant studied and was tested on all 144 target words. What varied across trials was the genre of the semantic space from which the cue had been derived.

On each cued-recall trial, participants were presented with six cue-target pairs, where the cues were drawn from one of the eight genre-specific distributional spaces. Over the full experiment, participants were tested on three trials from each of the eight literary genres: Mystery, Fantasy, Horror, Literary Fiction, Romance, Science Fiction, Thriller, and Non-fiction. The target words were therefore held constant across all participants and all genre conditions; only the cues changed as a function of genre. Thus, memory for targets was constant outside differences that might arise due to variation in item recallability reflecting relationships to paired cues.

Cues were selected algorithmically to maximize their associative relationship with the target within a given genre's distributional space, while minimizing their associative overlap with the same target in the remaining seven genres. This selectivity constraint ensured that each cue carried information that was diagnostically linked to its target in one genre but not in others, thereby amplifying the potential for individual differences in reading history and thus genre exposure to produce detectable variation in recall performance. Additional selection constraints required (a) a minimum within-genre cosine similarity of 0.15 between cue and target, (b) a maximum cross-genre cosine similarity of 0.08 with any other genre, (c) no morphological overlap between cue and target (Levenshtein distance greater than two), and (d) word lengths of four to eight characters with a maximum difference of two characters between cue and target. Candidate words were further restricted to nouns listed in the SUBTLEX-US database (Brysbaert & New, 2009), with similar frequency maintained across all eight genre corpora. All 1296 words in the final stimulus set were unique. Mean cosine similarity between cues and targets within each genre was relatively comparable across all eight genre conditions (means ranging from 0.176 to 0.200). The complete stimulus list is provided in Appendix A.

Procedure

The experiment was administered online via PsyToolkit (Stoet, 2010, 2017) in a single session lasting approximately 25 min. Participants first provided informed consent and then completed the reading-experience questionnaire described in the next section. They then received instructions explaining that they would study sequentially presented word pairs sequentially on a black background and that their memory for each pair would be tested immediately after a list of six pairs had been presented. Each pair consisted of one word displayed in yellow (i.e., the cue) and one displayed in white (i.e., the target); participants were told that at test they would see the yellow cue word and would need to recall and type the white target word with which it had been paired.

The experiment comprised 24 recall trials, each comprised of 6 cue-target pairs. Each of the eight genre conditions appeared exactly three times across the experimental session, with trial order randomised for each participant. Each trial began with a self-initiated start screen. During the study phase, six cue–target pairs were presented sequentially in the centre of the screen. The cue word appeared in yellow above the target word, which appeared in white, both in 20-point Arial font on a

black background. Each pair remained on screen for 2000 ms before the screen cleared. The six targets for a given trial were drawn at random without replacement from the pool of unstudied items, and the cue for each target was the word associated with it in that trial's genre-specific semantic space. Following the sixth pair, a blank screen was shown for 5000 ms before the test phase began.

At test, the six yellow cue words were presented one-at-a-time in a newly randomised order that differed from the study sequence; a feature that impeded participants from using serial position at study to aid retrieval. On each test screen, the cue word appeared in yellow at the centre of the display with a text entry field below and brief on-screen instructions reminding participants to type the missing word and then press Enter to continue. No feedback was provided at any point during testing. By the end of the session, each participant had studied 144 cue–target pairs exactly once, with each target word encountered in the context of a cue derived from one of the eight genre-specific distributional spaces.

Memory scoring

All typed responses were screened for spelling errors prior to scoring. Responses that differed from the target word by a Levenshtein distance of one were corrected to the intended target and treated as correct. A Levenshtein distance of one corresponds to a string match different by no more than one character-level operation: an insertion, a deletion, or a substitution. For example, if the target word was *drummer* and a participant typed *drumer*, the missing letter represents a single deletion and the response was accepted as correct. Responses differing from any target by a Levenshtein distance greater than 1 were not corrected.

Three outcome variables were then computed for each trial. A correct recall was recorded when the participant's response, following spelling correction, exactly matched the target word that had been paired with the cue at study (21,207 responses, 30.8%). An omission was recorded when the participant failed to produce a scoreable response, including cases where participants typed a non-response such as “don't know,” “unsure,” “skip,” or left the field blank (18,881 responses, 27.4%). An intrusion was recorded when the participant produced a word that was present in the full lexicon of the distributional memory model, which comprised 80,838 words, but that was neither the correct target nor any cue or target appearing on the current study list (13,010 responses, 18.9%). Because these intrusions are drawn from the broader lexical space over which the distributional representations are defined, they are the error type on which the representational comparison bears most directly, and they serve, with correct recall and omissions, as the key dependent variables in both demonstrations.

For transparency, we also report the remaining response types. These are not used as dependent variables in Demonstrations 1 or 2. Two of these categories involved words that had appeared on the current study list. Repetitions, in which a participant produced a target they had already recalled correctly at another test position within the same trial, accounted for 6204 responses (9.0%); these reflect the continued availability of a recently produced response, which the model addresses through its response suppression mechanism (described below). Binding errors, in which a participant produced a studied target that they had not otherwise produced on that trial, accounted for 4904 responses (7.1%) and reflect competition among the six pairs studied together. Responses reproducing a cue word accounted for a further 1104 responses (1.6%). Finally, 3522 responses (5.1%) did not correspond to any entry in the model's lexicon and could not be scored; these comprised misspellings differing from the intended target by more than a single character operation, common English words absent from the lexicon, and non-lexical entries. Because the model can only produce responses that exist in its lexicon, these responses have no counterpart in the simulations and were excluded from analysis rather than assigned to an outcome category. Across the 478 participants, the seven categories exhaust the 68,832 responses collected.

Reading-experience questionnaire

Prior to the memory task, participants completed a reading-experience questionnaire designed to index their reading history across the eight genres from which the experimental stimulus materials were drawn. The questionnaire was intended as an approximate but tractable measure of individual differences in genre exposure rather than a precise record of cumulative reading history. It comprised two sections. The first assessed general reading habits, including reading frequency (six options ranging from every day to rarely), typical session duration (five options ranging from about 15 min to more than two hours), and total annual book volume across all genres (five options ranging from none to more than 20). The second section assessed genre-specific reading. Participants rated their preference for each of the eight genres on an 8-point scale ranging from 1 (least preferred) to 8 (most preferred) and reported separately for each genre how many books of that type they read per year (five options: none, fewer than 2, 2 to 5, 6 to 10, or more than 10). The full text of the questionnaire is reproduced in Appendix B.

Descriptive statistics for general reading habits and genre-specific reading volume are presented in Tables 1 and 2, respectively. As shown in Table 1, the sample was composed primarily of regular readers: approximately 66% of participants reported reading at least twice per week and nearly 80% reported reading for at least 30 min per session. The majority of participants read between 5 and 20 books per year across all genres combined (60.1%), with only 1.7% reporting no reading at all.

Table 2 presents the distribution of annual reading volume for each genre separately. Mystery and Fantasy were the most widely read genres, with only 15.7% and 18.0% of participants, respectively, reporting no books in those genres per year. Horror showed the most restricted readership, with 41.8% of participants reporting no Horror books annually. Across genres, the modal response was typically in the lower reading volume categories, reflecting the broad but shallow genre coverage characteristic of general adult reading. This distributional pattern matters for corpus construction: participants with concentrated genre preferences received personalized corpora that differed substantially from the generic all-books corpus, whereas participants with more uniform reading profiles received personalized corpora that approximated it. This variability across participants is the individual-level signal embedded within shared language experience that the present

Table 1
Reading Habits of Participants: Frequency, Duration, and Annual Volume.

Source	*n*	%
Reading frequency		
Every day	94	19.7%
At least 4 times per week	103	21.5%
At least 2 times per week	117	24.5%
About once per week	75	15.7%
Less than once per week	59	12.3%
Rarely	30	6.3%
Reading time per session		
About 15 min	23	4.8%
About 30 min	185	38.7%
About 1 h	182	38.1%
About 2 h	60	12.6%
More than 2 h	28	5.9%
Books read per year (all genres combined)		
None	8	1.7%
Fewer than 5	87	18.2%
Between 5 and 10	149	31.2%
Between 10 and 20	138	28.9%
More than 20	96	20.1%

Note. $N = 478$. Reading frequency, reading time per session, and annual reading volume (all genres combined) are reported as counts (n) and percentages of the total sample. Percentages are rounded to one decimal place and may not sum to exactly 100%.

Table 2
Annual Reading Volume by Genre: Distribution of Participants Across Response Categories.

Genre	None	< 2	2–5	6–10	> 10
Fantasy	86 (18.0%)	184 (38.5%)	126 (26.4%)	51 (10.7%)	31 (6.5%)
Horror	200 (41.8%)	178 (37.2%)	78 (16.3%)	16 (3.3%)	6 (1.3%)
Literary Fiction	146 (30.5%)	181 (37.9%)	99 (20.7%)	35 (7.3%)	17 (3.6%)
Mystery	75 (15.7%)	203 (42.5%)	149 (31.2%)	34 (7.1%)	17 (3.6%)
Non-fiction	111 (23.2%)	157 (32.8%)	130 (27.2%)	56 (11.7%)	24 (5.0%)
Romance	142 (29.7%)	128 (26.8%)	115 (24.1%)	55 (11.5%)	38 (7.9%)
Science Fiction	151 (31.6%)	185 (38.7%)	101 (21.1%)	37 (7.7%)	4 (0.8%)
Thriller	115 (24.1%)	180 (37.7%)	117 (24.5%)	39 (8.2%)	27 (5.6%)

Note. $N = 478$. Each cell reports the number of participants (n) and the corresponding percentage of the total sample selecting each annual reading volume category for that genre. Column headers refer to the number of books read per year within the genre (< 2 = fewer than 2; 2–5 = between 2 and 5; 6–10 = between 6 and 10; > 10 = more than 10). Percentages are rounded to one decimal place and may not sum to exactly 100%.

study was designed to detect.

Personalized corpus construction

Genre-specific reading counts were used to construct a personalized corpus for each participant. The five response options for each genre were assigned integer scores of 1 through 5, corresponding to none, fewer than 2, 2 to 5, 6 to 10, and more than 10 books per year, respectively. The lowest category, reporting no books in a given genre, received a score of 1 rather than 0. A total reading score was then computed for each participant by summing the eight genre scores. Genre proportions were obtained by dividing each genre score by the participant's total, yielding a weight vector of eight values that summed to 1.0. Because every category contributed a non-zero score, every genre received a positive weight for every participant.

This design has an important consequence. Participants who reported reading no books at all, or who read uniformly across genres, received approximately equal weights of $1/8 = 0.125$ across all eight genres. For these participants, the personalized corpus closely approximates the generic all-books corpus. Participants with strong genre preferences received a personalized corpus dominated by their preferred genres. To construct a participant's corpus based on their self-reported reading pattern, sentences were sampled from each genre-specific subcorpus at a rate proportional to the participant's genre weight until a personalized corpus of approximately 80 million words was assembled. The comparison between the personalized and the generic representational spaces is therefore conservative: for participants with undifferentiated reading habits, the two spaces are nearly identical, and any advantage of the personalized representation must be driven by participants with distinctive genre-specific reading histories.

The three representational spaces used in the demonstrations that follow (i.e., personalized, all-books, and Wikipedia) were constructed using the BEAGLE distributional learning algorithm (Jones & Mewhort, 2007) applied to their respective corpora. Specifically, the sparse BEAGLE method of Recchia et al. (2015) was used in order to refine the representations further using the Global Negative and Distribution of Associations transformations (Johns et al., 2019); a method that mimics aspects of the negative training procedures used by neural embedding models. The next section provides a detailed description of the distributional model.

Distributional model

The sparse implementation of the model (Recchia et al., 2015), combined with the matrix transformation techniques described by Johns et al. (2019), is utilized here. In its standard form, BEAGLE constructs semantic representations using both contextual information (word co-occurrence within sentences) and order information (the relative positions of words within sentences). In the present work, however, only contextual information is used, as it is assumed to capture the associative and relational structure underlying word meanings. This model has previously been used in the false memory model of Chang et al. (2025), demonstrating its suitability for being used in a process model of episodic memory.

BEAGLE develops representations through a process of random vector accumulation that is consistent with the principle of Hebbian Learning (i.e., words that fire together wire together). Each time a word occurs in a sentence, the corresponding contextual information, encoded as a context vector c_i, is added to that word's memory vector m_i. Before learning begins, every word is assigned a random environmental vector e_i, intended to approximate its underlying perceptual or experiential features. These environmental vectors (and, by extension, the resulting memory vectors) are *N*-dimensional and contain four non-zero elements, with values of -1 and 1 sampled in equal proportion.

The context vector, c_i, for a word, i, is computed as the sum of the environmental vectors of all other words that occur in the same sentence. Formally, the context vector is defined as the sum of the environmental vectors for the remaining words in the sentence, where n denotes the total number of words in that sentence (see Eq. 1):

$$\mathrm{c_i} = \sum_{j=1}^{n} e_j, \textit{where } i \neq j \tag{1}$$

After the context vector is computed, the memory vector for the target word is updated by superimposing the current context vector onto it (see Eq. 2):

$$\mathrm{m}_i' = \mathrm{m}_i + c_i \tag{2}$$

where m_i' denotes the updated memory vector. With repeated exposure to sentences, each word's memory vector incrementally aggregates contextual information, forming a cumulative record of its co-occurrence patterns across the corpus. Once learning is complete, the memory vectors for all words are collected into a single word-by-feature matrix M, in which each row i corresponds to the memory vector $\mathbf{m}_i$ of one word in the lexicon and each column *j* corresponds to one of the N vector dimensions. For the corpora used here, M is therefore an 80,838 $\times$ *N* matrix. The transformations described next are applied to this matrix.

Crucially, these representations also encode latent semantic structure beyond direct co-occurrence. Because semantically related words tend to occur in similar contexts, their context vectors share many of the same environmental patterns, leading their accumulated memory vectors to converge in similarity.

To improve the quality of the semantic representations, the BEAGLE vectors were transformed using a combination of the global negative (GN) and distribution of associations (DOA) procedures, following Johns et al. (2019). Together, the GN and DOA transformations approximate a mechanism analogous to the negative sampling procedure used in word2vec (Mikolov et al., 2013). Negative sampling involves training a word's representation to be unpredictive with respect to unrelated words by incorporating randomly selected "negative" examples, sampled according to word frequency. This process reduces spurious associations by actively suppressing connections between a target word and irrelevant items, in conjunction with actively sharpening connections between a target word and relevant items. Johns et al. (2019) showed that negative sampling plays a critical role in word2vec's strong performance on semantic tasks and its success was not due to honing a prediction

mechanism alone, but, in addition, facilitates the extraction of distinctive co-occurrence patterns while controlling for baseline differences in word frequency. However, explicitly sampling unrelated words during each learning iteration is not a cognitively plausible mechanism whereas the GN and DOA transformations approximate the functional role of negative sampling without requiring an explicit sampling process (see also Goldberg & Levy, 2014). When applied to BEAGLE and other distributional models, these transformations yield improvements comparable in magnitude to those achieved with negative sampling (see also Johns, 2021, 2023, 2024).

The GN transformation incorporates base-rate co-occurrence information directly into the representations, operating in a manner analogous to the negative sampling procedure used in neural embedding models such as word2vec. In contrast, the DOA transformation emphasizes distinctive associations by amplifying co-occurrence patterns that differentiate items within the matrix. These transformations can also be applied jointly (GN + DOA), leveraging their complementary strengths (interested readers can find further technical details in Appendix C.).

Corpora. The main corpora that we used to construct distributional models was a book set used previously by Johns and Jamieson (2018, 2019), Johns and Dye (2019), and Johns et al. (2021). The details of the properties of the books are contained in Table 3, which shows that across the 8 genres tested here there were over 20,000 books utilized from 2500 authors and the total number of tokens was approximately 1.7 billion words. In addition, a comparison corpus of Wikipedia articles was used, which was capped at 80 million tokens.

To construct the personalized representation for each participant, as mentioned, their genre ratings were used to fill in a corpus of 80 million words, sampled by sentences. Each genre contributed sentences to these personalized corpora proportionally to their self-report ratings of how likely they are to read a certain genre. For example, if a participant reported that they read 40% mystery novels, 40% fantasy novels, and 20% literature novels, then their corpus would consist of 32 million words from mystery novels, 32 million words from fantasy novels, and 16 million words from literature novels. The corpora were filled by randomly sampling sentences from the respective genre corpora.

Demonstration 1: Personalized Representations Predict Memory Performance.

The first demonstration we present tests whether personalized semantic representations predict individual differences in memory more precisely than using generic representations. To test the proposition, we evaluated (a) if semantic similarity between a cue and its paired target improves the probability of successful retrieval (Howard & Kahana, 2002; Polyn et al., 2009) and (b) if that similarity is predictable based on a participant's self-reported reading history (Johns, 2024; Johns & Jamieson, 2018). If true, word representations derived from a corpus reflecting a participant's reading preferences should account for variation in correct cued recall as well as patterns of omissions and intrusions more accurately, compared to simulations with word representations derived from a generic corpus. To do that, we computed predictions directly from the word representations alone, independent of any commitment to a process model of memory that articulates and applies mechanisms for and assumptions about encoding, storage, and retrieval. Based on the strategy, the approach asks if the representational space matters independent of memory function and if predictions improve given the representational space corresponds to an individual's genre preferences?

Three representational spaces were used. The first was the personalized space, constructed for each participant from their personalized and genre-weighted corpus as described above. The second was the generic all-books space, derived from the evenly genre-weighted 80 million token sample from the full 1.73-billion-word corpus spanning all eight genres without any participant-specific weighting. The third was the generic Wikipedia space, derived from approximately 80 million tokens of encyclopaedic text and intended to represent non-literary and general world knowledge (i.e., internet language exposure). All three spaces were constructed with the same BEAGLE+GN + DOA pipeline. Semantic similarity between each cue–target pair was quantified using cosine similarity computed on the resulting memory vectors over the three conditions.

Table 3
Characteristics of Books Across the Eight Genres.

Genre	# of Authors	# of Books	Words per Book	Total # of Words
Fantasy	264	2231	98,972	220,806,532
Horror	75	728	86,202	62,755,056
Literature	357	2740	88,729	243,117,460
Mystery	321	2505	72,928	182,684,640
Non-fiction	248	1189	105,745	125,730,805
Romance	615	5004	74,835	374,474,340
Science Fiction	455	5336	75,225	401,400,600
Thriller	169	1245	99,249	123,565,005
Total/ Average	2504	20,978	87,736	1,734,534,438

Note. Counts and totals reflect the corpora used to derive the BEAGLE representations for the all-books and personalized conditions. The Wikipedia comparison corpus, not shown, comprising of 80 million tokens.

Statistical model

Cue–target similarity was quantified in each of the three representational spaces, and the three resulting measures were moderately intercorrelated across the tested pairs (personalized–generic-all-books $r = 0.37$; generic-all-books– generic-Wikipedia $r = 0.34$; personalized–generic-Wikipedia $r = 0.12$). The modest size of these correlations is a consequence of the stimulus construction: because cues were selected to be associated with their target within one genre space (cosine $\geq$0.15) and unassociated in the remaining seven (cosine $\leq$0.08), the similarity measures are substantially decoupled on the pairs actually tested, notwithstanding that all three spaces were built by the same algorithm from partly overlapping corpora.

To evaluate the three representational spaces, we estimated each in a separate model. For each of the three outcomes (correct recall, omission, and intrusion) we fitted three single-predictor Bayesian mixed-effects logistic regressions, one per representational space. Each model included a random intercept for participant to account for baseline differences in overall performance, and the outcome was modelled as a Bernoulli-distributed binary response with a logit link. Estimating one predictor per model, rather than all three simultaneously, has two advantages for present purposes: it removes any dependence among the predictors from the estimation, so that each coefficient reflects the total predictive relationship between that space and the outcome; and it allows the three spaces to be compared directly by Bayes factor.

Similarity predictors were standardized prior to estimation. Because Bayes factors require proper priors, we specified weakly informative priors on all parameters: normal (0,1) on the standardized slope, normal (0, 1.5) on the intercept, and a half-Student-t (3,0,1) prior on the participant random-effect standard deviation. Models were estimated with the brms package (Bürkner, 2017) in R (R Core Team, 2026) following the principled Bayesian workflow of Schad et al. (2021), using four Markov Chain Monte Carlo chains of 8000 iterations each, of which 4000 were discarded as warm-up, yielding 16,000 posterior draws per model. Convergence was assessed using the Gelman–Rubin statistic (all R-hat values were 1.000).

Posterior summaries are reported as median estimates with 95% highest density intervals (HDI) and the probability of direction (pd). For readers less familiar with these measures: the posterior median divides the posterior distribution in half; the 95% HDI is the narrowest interval containing 95% of the posterior probability mass, so that every value inside it is more credible than any value outside it; and the pd. is the proportion of the posterior sharing the sign of the point estimate,

interpretable as the certainty that the effect is positive or negative and bounded between 50% and 100%. An effect was considered credible when the 95% HDI excluded zero and the pd. exceeded 95%.

The central question of this demonstration, which of the three representational spaces provides the better account of memory performance, is a question of model comparison rather than of parameter estimation, and we address it directly with Bayes factors computed from the marginal likelihood of each model, estimated by bridge sampling (Gronau et al., 2017) as implemented in the bridge sampling package in R. A Bayes factor expresses the relative evidence the data provide for one model over another; a value of 3 indicates moderate evidence and a value of 10 strong evidence, with values below 1 favouring the comparison model (Jeffreys, 1961). We report the pairwise comparisons between representational spaces, which are the comparisons of theoretical interest and which, because they contrast models of identical structure, are largely insensitive to the scale of the prior placed on the slope. To confirm this, we repeated the full analysis across a range of prior scales on the standardized slope (SD = 2, 1, 0.5, 0.25, 0.1, and 0.05), refitting all models and recomputing all marginal likelihoods at each scale; the resulting Bayes factors are reported in Table 5.

Finally, we note that the models are deliberately minimal. Because every participant studied the same 144 targets and cue words were matched at selection for length, morphological overlap, and comparable frequency across the eight genre corpora, targets and thus item-level properties are held constant across the three similarity measures being compared and cannot differentially favour any one of them; only the genre-specificity of the cue varied. The comparison is moreover between a predictor that varies across participants for a given item (personalized similarity) and two that do not (all-books and Wikipedia similarity), so the additional predictive power of the personalized measure is by construction attributable to person-specific rather than item-level variation.

Results

The posterior estimates are summarised in Table 4 and Fig. 1. Because the three representational spaces differ in the variability of their similarity values, coefficient magnitudes are not directly comparable across spaces; thus, the comparison between representations is addressed by the Bayes factors reported below rather than by the relative size of these estimates.

Table 4
Posterior Summaries for Single-Predictor Bayesian Mixed-Effects Logistic Regression Models Predicting Trial-Level Memory Outcomes.

Predictor	b	95% HDI	pd	R-hat	ESS
Correct recall					
Personalized	0.54	[0.15, 0.92]	99.69%	1	33,723
Generic - All Books	0.37	[0.07, 0.65]	99.31%	1	33,131
Generic - Wikipedia	0.25	[−0.05, 0.56]	94.58%	1	34,069
Omission					
Personalized	−0.55	[−0.99, −0.09]	99.16%	1	29,637
Generic - All Books	−0.29	[−0.63, 0.04]	95.28%	1	18,939
Generic - Wikipedia	−0.33	[−0.70, 0.05]	95.64%	1	21,754
Intrusion					
Personalized	−0.16	[−0.62, 0.31]	74.86%	1	23,749
Generic - All Books	0.04	[−0.32, 0.39]	58.58%	1	32,602
Generic - Wikipedia	−0.3	[−0.66, 0.09]	93.64%	1	33,617

Note. Estimates are log-odds coefficients from separate Bayesian mixed-effects logistic regression models, one per predictor per outcome, each with a random intercept for participant. Predictors were standardized for estimation and coefficients back-transformed to the raw cosine scale, so *b* is expressed per unit of cosine similarity; the standard deviations used were 0.047 (personalized), 0.062 (all books), and 0.057 (Wikipedia). HDI = highest density interval; pd. = probability of direction; ESS = bulk effective sample size, which is estimated from the autocorrelation of the chains and can therefore exceed the number of posterior draws (Vehtari et al., 2021). Bold indicates estimates whose 95% HDI excluded zero and whose pd. exceeded 95%. N observations = 68,832; N participants = 478.

For correct recall, cue–target similarity in the participant's personalized space was positively associated with the probability of successful retrieval ($b = 0.54$, 95% HDI [0.15, 0.92], pd. = 99.69%). Similarity in the all-books space was also credibly associated with correct recall ($b = 0.37$, 95% HDI [0.07, 0.65], pd. = 99.31%), whereas the Wikipedia predictor did not meet the criterion ($b = 0.25$, 95% HDI [−0.05, 0.56], pd. = 94.58%). The pattern was complementary for omissions, where greater cue–target similarity was associated with a lower probability of retrieval failure: the personalized predictor yielded a credible negative estimate ($b = -0.55$, 95% HDI [−0.99, −0.09], pd. = 99.16%), while neither the all-books predictor ($b = -0.29$, 95% HDI [−0.63, 0.04], pd. = 95.28%) nor the Wikipedia predictor ($b = -0.33$, 95% HDI [−0.70, 0.05], pd. = 95.64%) had an HDI excluding zero. For intrusions, no representational space met both criteria: the personalized (pd = 74.86%) and generic all-books (pd = 58.58%) predictors were far from directional certainty, and although the generic Wikipedia predictor approached it (pd = 93.64%), its HDI included zero.

When each space is evaluated on its own, then, the generic all-books space also carries predictive signal for correct recall, and the estimates for the three spaces are of broadly similar magnitude. The relevant question is therefore not whether generic representations are informative (they are), but whether the personalized representations provide a better account as measured by improved predictive accuracy. This is a comparison the individual estimates cannot settle, and we turn to it next.

Model comparison

Bayes factors comparing the three representational spaces are reported in Table 5, both at the primary prior scale and across the full sensitivity range. For correct recall, the data favoured the personalized space over the generic all-books (BF = 4.57) and Wikipedia spaces (BF = 24.78), constituting moderate and strong evidence respectively. For omissions the ordering was the same, with weaker but still positive evidence (BF = 2.10 over all-books; BF = 2.89 over Wikipedia). For intrusions no space was favoured: the personalized space was marginally preferred to the all-books space (BF = 1.60) and marginally disfavoured relative to the Wikipedia space (BF = 0.55), values that provide no meaningful evidence in either direction.

These pairwise comparisons were stable across the range of prior scales examined (Table 5). For correct recall, the personalized space was favoured over the all-books (BF 1.16–4.57) and Wikipedia (BF 6.42–28.22) spaces at every scale. For omissions the same held (BF 1.48–13.87 over all-books; BF 2.86–7.46 over Wikipedia). For intrusions, no consistent ordering emerged at any scale (BF 0.59–2.16 over all-books; BF 0.31–0.75 over Wikipedia). Because the pairwise comparisons contrast models of identical structure, the influence of the prior largely cancels, and the conclusion that the personalized representation provides the better account of correct recall and omissions does not depend on the particular prior adopted.

Cue–target similarity computed in a participant's personalized semantic space predicted correct recall and omissions credibly, and predicted them better than similarity computed in either generic space with the advantage holding across a range of prior scales. Intrusions were not reliably predicted by any representational space, nor did the spaces differ in their account of them.

Note. Each panel displays the effect of cosine similarity on the predicted probability of the outcome variable, derived from a separate Bayesian mixed-effects logistic regression model containing that similarity measure as its only predictor (nine models in total). Rows correspond to the three outcome variables: correct recall (top), omission (middle), and intrusion (bottom). Columns correspond to the three representational spaces from which cosine similarity was computed: Wikipedia (red), All Books (green), and Personalized (blue). Solid lines represent the posterior median predicted probability as a function of cosine similarity; shaded regions correspond to the 95% credible interval. Predictors were standardized for estimation, and the horizontal axis

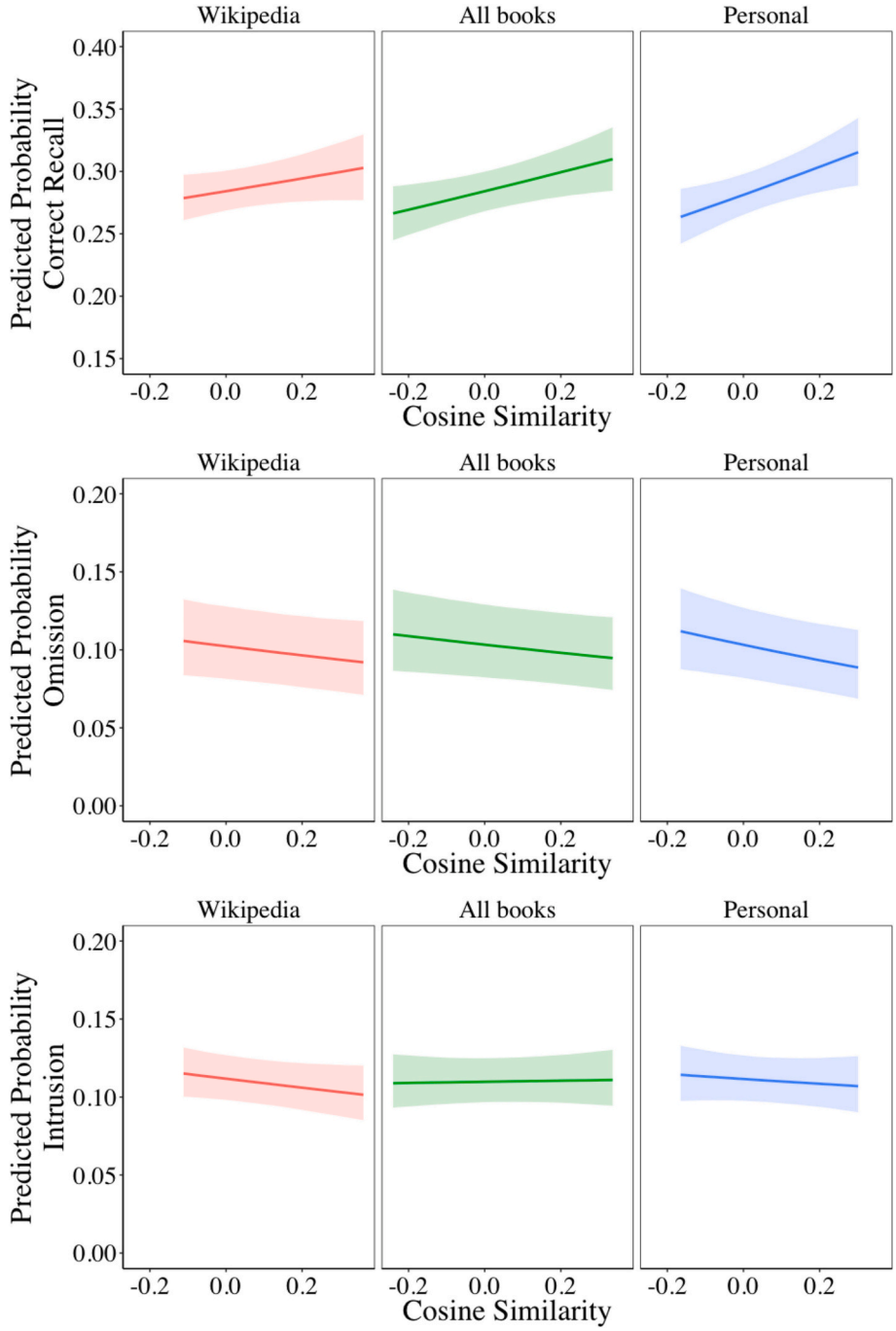

Fig. 1. Effects of Cosine Similarity on the Predicted Probability of Correct Recall, Omission, and Intrusion Across Three Representational Spaces.

has been back-transformed to the raw cosine scale. The personalized representational space produced the steepest positive slope for correct recall and the steepest negative slope for omission.

Discussion

These results support the central hypothesis of the study. The personalized representational space provided a more reliable account of correct recall and omissions than either generic space, despite the fact that all three spaces were built from the same distributional architecture applied to the corpora that differed in domain and genre composition, and independent of a theory for encoding, storage, and retrieval.

The advantage of the personalized space was evident not in its raw effect size relative to the generic spaces but in the reliability of its posterior distribution: only the outcomes using the personalized representations produced a posterior that excluded zero with sufficient consistency to meet the directional criterion for both correct recall and omissions. This result is noteworthy given the fact that language histories were only coarsely represented and only one aspect of an individual's language history was assessed. The personalized corpora were constructed from self-reported genre preferences, a necessarily coarse and imperfect proxy for actual reading experience (cf. Acheson et al., 2008; Vermeiren et al., 2023). Nevertheless, the fact that representational structure derived from even this coarse index of reading history yielded reliable differences in predictive power suggests that experiential variation in language experience that plays a role in determining a person's semantic space carries a detectable signal, one that a more precise characterization of individual linguistic history would be

Table 5
Bayes Factors Comparing Representational Spaces Across Prior Scales.

Prior SD	Correct recall	Omission	Intrusion
Personalized vs. Generic - All Books			
2	1.43	13.07	2.16
1	4.57	2.1	1.6
0.5	2.48	5.75	1.93
0.25	2.94	13.87	1.55
0.1	2.51	3.1	1.38
0.05	1.16	1.48	0.59
Range	*1.16–4.57*	*1.48–13.87*	*0.59–2.16*
Personalized vs. Generic - Wikipedia			
2	10.91	7.46	0.47
1	24.78	2.89	0.55
0.5	16.12	5.31	0.64
0.25	28.22	6.89	0.31
0.1	16.95	2.86	0.75
0.05	6.42	4.31	0.43
Range	*6.42–28.22*	*2.86–7.46*	*0.31–0.75*

Note. Values are Bayes factors favouring the personalized representation, computed from marginal likelihoods estimated by bridge sampling. Values above 1 favour the personalized space; conventionally, values above 3 constitute moderate and values above 10 strong evidence (Jeffreys,1961). Each row corresponds to a different standard deviation of the normal prior placed on the standardized slope; the SD = 1.00 row is the primary analysis reported in the text.

expected to amplify further.

Demonstration 2: Personalized Representations Improve Predictions in the eCFM

Demonstration 1 established that cue–target similarity in a personalized semantic space predicts memory outcomes better than similarity in a generic one. Despite the importance of this result, it does not address whether personalized representations are compatible with the mechanistic architecture of a formal memory model, or whether they are useful within such a model. Demonstration 2 takes that step. We ask whether the same personalized embeddings, when imported into the eCFM (an instance-based model of memory; Guitard et al., 2025a, 2025b, Guitard et al., 2026), improve cue–target-level prediction in cued recall. The eCFM is one instantiation of a broader theoretical movement toward grounding memory models in individually tailored representations, a direction that is being pursued across multiple labs and frameworks (Chang et al., 2025; Cox, 2026; Guitard et al., 2025a, 2025b, Guitard et al., 2026; Healey & Kahana, 2014; Osth et al., 2026; Petilli et al., 2026; Reid & Jamieson, 2023; Reid et al., 2026).

The question itself has deep theoretical roots. Simon (1956) argued that intelligent behaviour cannot be understood by examining internal processes alone but emerges from the interaction between a processing system and the structure of its environment. Nosofsky (1985, 1986) gave this principle a formal and empirical foundation in the domain of memory and categorization, showing that individualized representational spaces derived from participants' individualized perceptual judgements, combined with a retrieval and decision model, yielded high-precision fits to individual behaviour. The present work is theoretically and structurally continuous with that history of work and its fundamental argument. In fact, we think of the advance we offer here as an extension of rather than a departure from Nosofsky's framework. Where he derived individualized representations from confusion matrices over small well-defined perceptual stimulus sets, we derive them from the distributional statistics of large, naturalistic language corpora. If even a coarse genre-level approximation of individual linguistic experience yields measurable gains in predictive accuracy when substituted into the eCFM, the implication is significant: individual representational structure of word meaning is not merely correlated with memory behaviour but is an important and integral component of remembering, and, thus, a critical consideration for advancing the precision and scope of computational memory theory.

The eCFM

The model used in this demonstration is the eCFM (Guitard et al., 2025a, 2025b, Guitard et al., 2026), which integrates the encoding, storage, retrieval, and decision mechanisms of MINERVA 2 (Hintzman, 1984, 1986) with structured semantic representations derived from distributional models of language. The eCFM is best understood as a soft instance theory: it brings episodic (MINERVA) and semantic (e.g., BEAGLE) memory together under an integrated architecture while acknowledging that semantic representations and episodic retrieval operate by different principles; distinct from a harder instance-theoretic position in which semantic memory is a corollary of episodic storage and retrieval rather than a separately derived structure and system (Jamieson et al., 2018, 2022). The eCFM has accounted for serial recall, order reconstruction, and recognition performance across a wide range of materials and participant populations (Guitard et al., 2025a, 2025b, Guitard et al., 2026; Reid et al., 2026), and in the present study we extend it to cued recall. What makes the eCFM the right test bed for the present purposes is precisely its architecture: semantic representations enter the model as a structured lexicon that is separable from the encoding, retrieval, and decision components, meaning the lexicon can be swapped without altering any other aspect of the model. Three versions are run here, one using the personalized corpora, one using the generic all-books corpus, and one using the generic Wikipedia corpus.

Semantic representations

The semantic representations used here were the BEAGLE vectors with GN + DOA transformations already described and used in our first demonstration.[2] Thus, each word in the lexicon is represented as an n-dimensional vector. The full lexicon comprised 80,838 words.

Encoding and storage

In the cued recall simulations, each studied cue–target pair was encoded as a single memory trace formed by adding the cue vector C_i and its associated target vector T_i. The resulting memory trace M_i is therefore (see Eq. 3):

$$M_i = C_i + T_i \tag{3}$$

This summation approach differs from earlier eCFM implementations of serial recall, in which cues and items were concatenated into separate subspaces (e.g., C_i occupying dimensions 1–300 and T_i occupying dimensions 301–600; see Guitard et al., 2025a, 2025b, Guitard et al., 2026). However, summation is motivated here by two commitments of the framework itself. The first concerns what an instance-based model takes a memory trace to be. In models of the MINERVA family, a trace is not a stored associative bond between constituents but an undifferentiated record of what co-occurred on a study occasion; association is not a property of memory but a product of retrieval, emerging when a cue activates traces in proportion to their similarity to it and the echo aggregates the resulting activation (Hintzman, 1986, 1988; Jamieson et al., 2012, Jamieson et al., 2018, Jamieson et al., 2022). Superimposing the cue and target vectors is therefore not a claim that an association is a unique representation of the items associated, but an implementation of the more basic claim that the trace registers their co-occurrence, leaving associative structure to be recovered by the retrieval process. The second is a matter of internal consistency. Superposition is already the representational primitive of the model at the

[2] In the present simulations we focus on semantic representations. In other work we have included representations from the domains of orthography and phonology (see Guitard et al., 2025b, 2026; Reid et al., 2026).

semantic level: BEAGLE constructs word meanings by summing the environmental vectors of words that occur together within a sentence (Eqs. 1 and 2). Forming the episodic trace by the same operation applies that principle at the episodic level rather than introducing a second and different rule for events. The practical consequence for the present analyses is that summation embeds the cue–target relation within shared vector dimensions, so that retrieval is sensitive to the semantic overlap between the two words studied together, whereas concatenation confines cues and targets to distinct representational halves and so cannot express the within-pair semantic relatedness that the present comparisons turn on.

Each memory trace was encoded probabilistically, governed by a uniform learning parameter L that determined the proportion of feature values retained during encoding. When L is low, a greater proportion of feature values are replaced with zeros, simulating partial or degraded encoding; when $L = 1$, the trace represents a perfect encoding of the cue–target pair. Conceptually, L functions similarly to the encoding parameter used in prior recognition modelling work (Reid et al., 2026), capturing natural variability in encoding fidelity across items while maintaining representational consistency across conditions.

Retrieval and decision

Retrieval in the eCFM is guided by cue-based similarity, following the principles of MINERVA 2 (Hintzman, 1984, 1986) and other cue-driven memory models (e.g., Nairne, 1990; Saint-Aubin et al., 2021). At test, the cue vector, c, corresponding to the intact representation of a studied cue word, is presented to memory. This cue interacts in parallel with all stored traces in memory, M, activating each trace in proportion to its similarity with the cue. The resulting activation pattern produces an echo, e, that represents an activation-weighted sum of all traces in memory. Each trace contributes to the echo in proportion to its similarity with the cue. The echo serves as a retrieval signal from which the model attempts to reconstruct the associated target word. The retrieval process is summarised in Eq. 4:

$$\mathrm{e} = \sum_{i=1}^{i=m} \left(\frac{\sum_{j=1}^{j=n} c_j \times M_{ij}}{\sqrt{\sum_{j=1}^{j=n} c_j^2} \sqrt{\sum_{j=1}^{j=n} M_{ij}^2}} \right)^{\tau} \times M_{i\bullet} \tag{4}$$

In Eq. 4, c_j is the j-th feature of the cue vector, M_{ij} is the j-th feature of the i-th memory trace, n is the dimensionality of the representation, and m is the number of traces in memory (equal to the list length, here six). The parameter τ governs the selectivity of retrieval by scaling the contribution of each trace according to its similarity with the cue. Higher values of τ make retrieval more trace selective by suppressing activation of traces only moderately similar to the cue that, in turn, emphasizes traces that are highly similar to the cue, whereas lower values produce a more diffuse retrieval process. In all simulations $\tau = 3$, consistent with previous eCFM studies and with a long history of work using the MINERVA 2 model (Guitard et al., 2025a, 2025b, Guitard et al., 2026; Hintzman, 1984, 1986; Reid et al., 2026).

To determine the model's response, the echo is compared to every word in the lexicon of 80,838 words by cosine similarity. The model reports the word whose vector is most similar, provided that the similarity exceeds a decision threshold T. If the similarity between the echo and no word in the lexicon exceeds T, the model produces an omission. A global suppression mechanism was applied across the six test items within each trial. Each time a word was selected as a response, its cosine similarity to the echo was reduced by a fixed suppression parameter s for all subsequent retrievals within that trial. This reduction is cumulative: a word selected once has its similarity reduced by s, if selected again it would be reduced by a further s, and so on. Thus, the suppression mechanism does not prevent a word from being reported again but progressively reduces its likelihood of being reported, reflecting the decreasing probability of producing the same response to multiple different cues within the same recall trial (see Guitard et al., 2025a). To mirror the experimental instructions given to participants, the model was additionally constrained so that the cue word itself could not be selected as the response on the trial on which it served as the cue.

A response was scored as correct if it matched the studied target for that cue. It was scored as an intrusion if the reported word was present in the lexicon but was neither the correct target nor any other word appearing on the current study list. It was scored as an omission if no lexical entry exceeded the decision threshold T.

Simulation parameters

The eCFM was run with a single fixed parameter set applied identically across all three representational conditions: learning rate $L = 0.37$, retrieval sensitivity $\tau = 3$, response suppression $s = 0.01$, and decision threshold $T = 0.434$. Three of these four parameters were taken directly from our prior work (Guitard et al., 2026). The decision threshold T was the single exception. Because T controls the boundary between a recall response and an omission, its optimal value is sensitive to the specific materials and list length of a given experiment, making a direct importation of T from prior work inappropriate. We therefore identified the value of T that minimized the discrepancy between the model's aggregate predictions and the observed data across all three scoring categories simultaneously, that is, the value that brought the model's average proportion of correct responses, omissions, and intrusions into closest correspondence with the observed averages across participants. This optimization was performed once, collapsing across all three representational conditions, and the resulting value of $T = 0.434$ was then applied identically to all three conditions.

This procedure was deliberate. By deriving T from aggregate correspondence across all conditions jointly rather than separately for each kind of representation, we ensured that no representational condition received a parameter advantage over any other. All three representations therefore begin from the same aggregate baseline. Any differences in cue–target-level correspondence between empirical and simulated performance estimates that emerge across conditions cannot be attributed to differential parameter tuning and must instead reflect a genuine gain from the representational change itself.[3]

Simulations were run 20 times per participant per representational condition, yielding trial-level predictions for correct recall, omissions, and intrusions. Although 20 replications is modest, we settled on that decision because compute time was expensive and because, as we will show, results across 20 simulations along with differences were stable and decisive. A key advantage of the eCFM for the present purpose is that it operates over a structured lexicon of word representations (Guitard et al., 2025a, 2025b, Guitard et al., 2026), which means the model can be run on the exact same cue–target pairs that each participant studied. This matters because memory performance is well established to vary systematically across individual words, reflecting differences in their semantic properties, frequency, and neighbourhood structure (Guitard et al., 2018, 2019). By simulating the model on each participant's actual study list, we ensure that the model and the participant are processing the same material, making the comparison between

[3] As a supplementary check, the optimization was also performed separately for each representational condition. The resulting values were All Books $T = 0.433$, Wikipedia $T = 0.433$, and Personal $T = 0.438$, differing from the jointly optimised $T = 0.434$ by at most 0.004. Correlations at each condition's own best-fitting T were similar to those obtained under the joint T (Personal: $r = 0.497$ vs 0.502; All Books: $r = 0.463$ vs 0.462; Wikipedia: $r = 0.452$ vs 0.452), and the advantage of personalized representations remained significant by Williams's test in both cases ($p < .001$). The joint optimization approach is therefore not a source of representational advantage. Full results are available on the OSF page associated with this article.

simulated and observed performance empirically and methodologically accurate rather than merely approximate.

Model fit was assessed using Pearson and Spearman correlations (see Fig. 2) between simulated and observed proportions at the cue–target level, pooled across all three outcome measures. We focus on this level of analysis because it is the granularity at which the influence of individual reading experience is most likely to be detectable. If a participant's reading history has shaped the semantic space of a specific cue–target pair in a way that facilitates or impedes retrieval, this signal will be visible in cue–target-level variation and will tend to be obscured by aggregation to participant means.

Results

The results are presented in Fig. 2. At the cue–target level, model–data correspondence was assessed by correlating simulated and observed proportions across all 3456 cue–target × outcome combinations (144 cue–target pairs × 8 genre-specific cues × 3 outcome measures), with model predictions averaged across 20 simulation runs per participant per condition. The personalized representation spaces produced the strongest correspondence ($r = 0.515$), followed by the generic all-books ($r = 0.472$) and Wikipedia ($r = 0.469$) representations. In terms of explained variance, the personalized representation accounted for 26.5% of cue–target-level variability, compared with 22.3% for all-books and 22.0% for Wikipedia. Williams's (1959) test for dependent correlations confirmed that the personalized representation predicted cue–target-level outcomes reliably better than both the all-books, $t(3453) = 6.13$, $p < .001$, and the Wikipedia, $t(3453) = 5.36$, $p < .001$, representations.

To assess whether this advantage depended on the randomness of any single simulation, we computed the model–data correlation separately for each of the 20 runs, yielding a distribution of correlations for each representation. As shown in Fig. 3, the distributions were tightly concentrated and did not overlap. The mean per-run correlation was $r = 0.503$ ($SD = 0.006$) for the personalized representation, $r = 0.459$ ($SD = 0.004$) for all-books, and $r = 0.454$ ($SD = 0.005$) for Wikipedia; the personalized correlations ranged from 0.493 to 0.515 across runs, and neither generic representation exceeded 0.467 in any run. The personalized representations produced a higher correlation than both generic representations in all 20 runs, with a mean advantage of 0.044 over all-books (95% interval across runs [0.035, 0.056]) and 0.049 over Wikipedia, both paired comparisons $p < .001$. (The mean of the per-run correlations, 0.503, is marginally lower than the correlation computed on the run-averaged predictions reported above, 0.515, because averaging across runs removes simulation noise before the correlation is taken.) The small across-run variability confirms that both the model's predictions and the ordering of the three representations are stable.

Discussion

Overall, we observed a modest but empirically reliable and theoretically important improvement for personalized embeddings; though, the modest magnitude of the improvement deserves comment. The gain in explained variance at the cue–target level is real but not large, and its reliability is clear: across 20 independent simulation runs the personalized representation produced a better fit than both generic representations in every run, with non-overlapping distributions of model–data correlations. We regard the modest magnitude as expected rather than disappointing. The personalized corpora were constructed from self-reported genre preferences, a sufficient albeit pragmatically coarse index of individual reading history. Participants with undifferentiated reading profiles received personalized corpora that closely approximated the generic all-books space, diluting the manipulation for a substantial portion of the sample. However, the fact that a reliable and consistent gain emerged despite this conservatism was observed. It shows that even a rough genre-level approximation of individual experience is sufficient to improve the representational accuracy of the model in a way that participant-level analysis is unlikely to reveal. However, it is equally clear that our proof of concept based on people's self-reported genre preferences is insufficient to represent the detailed

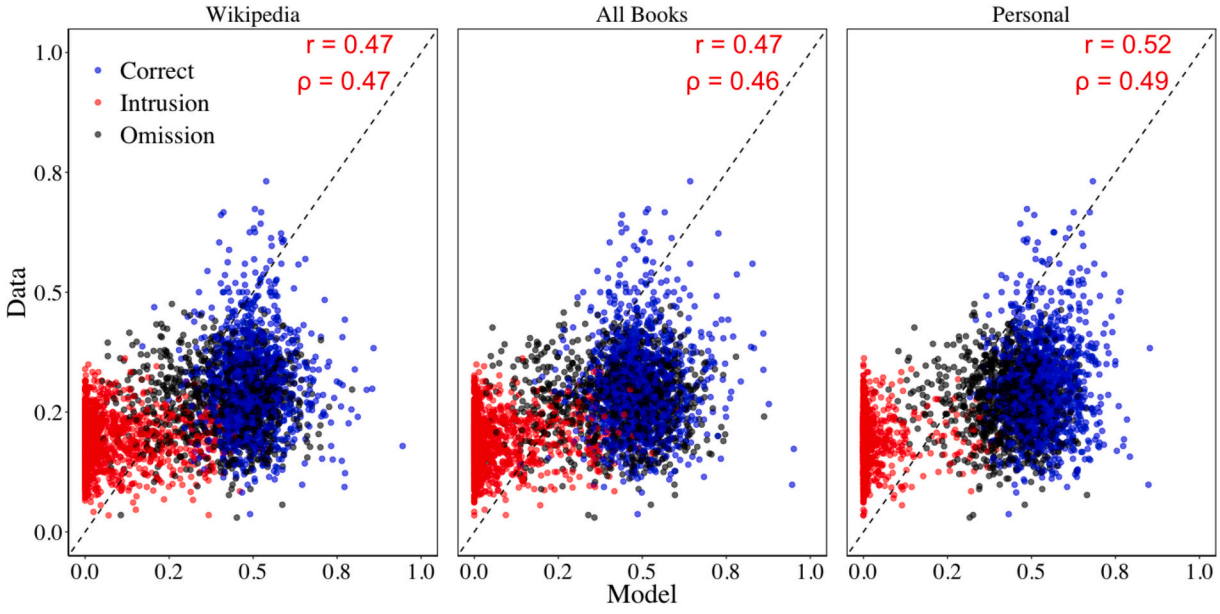

Fig. 2. Model–Data Correspondence for the eCFM Under Three Representational Conditions.
Note. Model-predicted proportions are averaged across 20 simulation runs per participant per representational condition. Pearson r and Spearman ρ shown within each panel are computed on these run-averaged predictions. Each panel displays observed versus model-predicted proportions for correct recall (blue), intrusions (red), and omissions (black) under each representational condition: Wikipedia (left), All Books (centre), and Personalized (right). Each data point represents one cue–target pair × outcome combination, averaged across participants. The dashed diagonal indicates perfect model–data agreement. Pearson r and Spearman ρ values are displayed within each panel. The eCFM was run with identical parameter values across all three representational conditions ($L = 0.37$, $\tau = 3$, $T = 0.434$, $s = 0.01$); only the lexical representation differed.

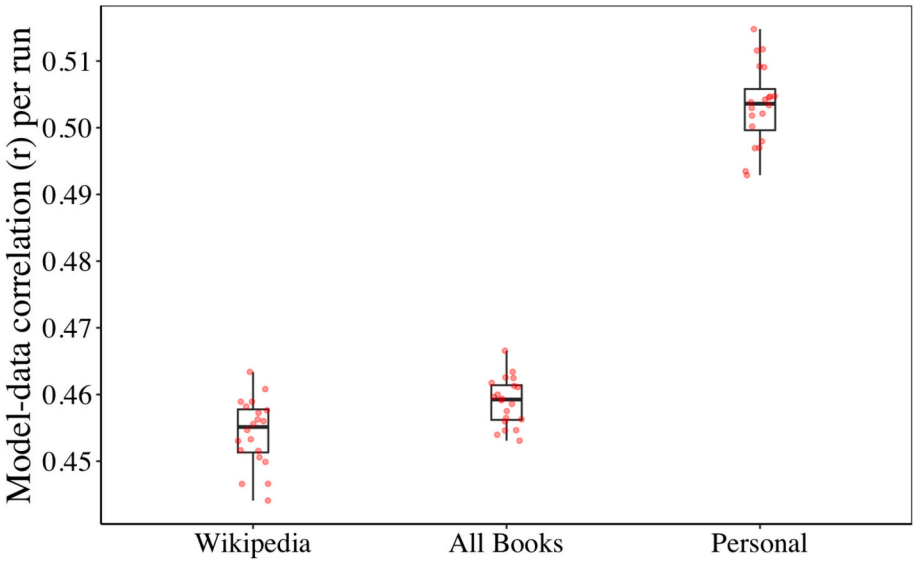

Fig. 3. *Distribution of Model–Data Correlations Across 20 Simulation Runs for Each Representational Condition.*
Note. Each point represents the Pearson correlation between observed and model-predicted proportions across all 3456 cue–target × outcome combinations for a single simulation run (20 runs per representational condition). Boxplots show the median and interquartile range; whiskers extend to 1.5 times the interquartile range. Points are horizontally jittered to reduce overplotting. The personalized representation produced a higher model–data correlation than both generic representations in every one of the 20 runs, and the three distributions did not overlap. The eCFM was run with identical parameter values across all three conditions ($L = 0.37$, $\tau = 3$, $T = 0.434$, $s = 0.01$); only the lexical representation differed.

history of individual differences in participants' language experiences. If we were able to know our participants' corpora rather than inferring and approximating them based on the reading questionnaire, the present proof of concept has the potential to yield additional and potentially substantial gains.

The results of Demonstration 2 establish that personalized embeddings are not merely statistically predictive of memory behaviour but are also functionally compatible with and beneficial when embedded within a formal mechanistic account of memory. The case is not theoretically different from the one that Nosofsky (1985, 1986) made when he used participants' confusion matrices for simple and well-defined perceptual stimulus domains where he showed that individualized MDS solutions of representational spaces combined with a theory for retrieval and decision that operates on those spaces support high-precision model fits to individual data, as well as differential predictions across participants within that domain. Here, however, the representational domain is large (80,838 words), ill-defined, and derived indirectly by mapping individualized (albeit coarse) inferred corpora from which individual semantic representations follow. Nevertheless, despite the difficulties inherent to an analysis of representation from large language sets compared to small perceptual domains, the eCFM applied with identical parameters across representational conditions produced better cue–target-level predictions when its lexicon was tailored to the individual than when it was populated with a generic corpus. That the improvement was obtained without any change to the model's parameters, only its representational input varied, is itself informative: the gain is attributable to the representation rather than to added flexibility in fitting; a result that distinguishes personalization by measured experiential input from post hoc personalization by free parameters. This finding carries an important theoretical implication: improving current computational models of memory may lie not only in some refinement of the retrieval architecture from which memory performance is predicted but also in the representational spaces on which those architectures operate.

General discussion

Across two demonstrations, the present study provides converging evidence that personalized semantic representations improve the prediction of individual memory behaviour relative to generic, population-level alternatives. Demonstration 1 established the result at the level of word pairs using a model-free Bayesian analysis: cosine similarity computed within each participant's personalized semantic space reliably predicted correct recall and omissions, whereas similarity computed within generic spaces did not meet the criterion. Demonstration 2 showed that the gain is not merely descriptive: substituting personalized embeddings into the eCFM improved cue–target-level predictions without altering the model's retrieval architecture or its parameter values. Together, these findings support the claim that experiential variation in language experience as reflected in semantic structure is not a source of noise to be averaged over in computational accounts of memory. Rather, it is a meaningful and tractable influence on remembering that can and should be incorporated into formal theories. The more precisely pre-experimental knowledge is modelled, the more precisely experimental performance can be understood.

What the present findings show, and what they do not

It is important to be clear about what our findings do and do not demonstrate. They do not show that we have solved the problem of individual differences in memory. They do not show that genre-weighted corpora capture the totality or even the majority of the variation that distinguishes one person's semantic representations from another's. They do not show that the eCFM is a complete or final theory of memory; it is only one of several good candidate computational frameworks (Healey et al., 2019; Howard & Kahana, 2002; Logan, 2021; Polyn et al., 2009; Raaijmakers & Shiffrin, 1981). They do not show that the model predicts fine-grained within-measure differences in item difficulty, a standard that no current computational model of memory consistently meets. What the cue–target-level advantage reflects is the model's ability to predict the distribution of outcome types across pairs, which pairs yield relatively more correct responses versus omissions or intrusions, and the level of analysis at which a process model operating over semantic similarity would be expected to show sensitivity to variation in representational structure.

What the present findings do show is constrained but, in our view,

theoretically important. Even a coarse, ecologically grounded index of differential linguistic experience produces measurable improvements in the prediction of memory behaviour, both descriptively and within a formal process model. This constitutes a useful proof of concept that we treat as a first critical step toward a larger research programme in which the representational structure of pre-experimental knowledge is treated as a problem to solve, a problem as important as the structure and function of the memorial mechanisms that operate on that pre-experimental knowledge.

The conservative nature of the present design strengthens rather than weakens this conclusion. The personalized representations were constructed from heuristic, self-reported reading preferences, a coarse proxy for actual reading experience. Participants reported the number of books they read per year across eight genres, and these reports were used to weight genre-specific corpora. This procedure cannot capture the full richness of an individual's linguistic history: it ignores differences in reading depth, vocabulary breadth within genres, reading outside the listed categories, the accumulated effects of non-literary language exposure, and even linguistic variation within the genres (Acheson et al., 2008; Lowder & Gordon, 2017; Mar & Rain, 2015; Vermeiren et al., 2023). It also ignores variation in linguistic experience that shapes semantic structure but is not captured by book reading, including educational history, occupational vocabulary, internet reading, movie watching, everyday conversation, first language phonology, age and cohort effects, and frequency of use across registers (Johns, 2021; Teles et al., 2025; Wulff et al., 2019). However, from our perspective, the fact that reliable predictive gains were obtained in the face of these limitations strengthens rather than weakens the theoretical conclusion. If a rough genre-level approximation of individual experience is sufficient to improve prediction, then more precise characterizations of individual linguistic history should be expected to produce even larger gains. In that sense, knowing a person's language corpus becomes a new problem to solve in the already long history of memory modelling that has tended to focus on mechanisms and intra-experimental features rather than extra-experimental experience and knowledge.

What the personalized representations capture

A natural question about these results is whether the advantage of personalized representations reduces to familiarity: whether a personalized space is simply a record of which words a person has encountered often. If true, the effects we report might be interpreted differently as the influence of raw word exposure rather than of semantic structure. The question is worth addressing directly, because the answer clarifies our contribution.

Item-level properties cannot produce the effects we report. Every participant studied the same 144 targets, and cue words were constrained at selection for length, morphological overlap, and comparable frequency across the eight genre corporas. More importantly, the three similarity measures we compared differ in a way that is diagnostic. Similarity computed in the generic all-books and Wikipedia spaces is a property of a cue–target pair alone: it takes the same value for every participant who studies that pair. Similarity computed in a participant's personalized space is a property of a pair and a person: it varies across participants who study the identical pair conditional on their reported reading habits. Any variable that is constant across participants for a given item (e.g., word frequency, concreteness, length, or similarity in a generic space) cannot in principle account for variance of this second kind. The improvement observed when personalized similarity replaces generic similarity is therefore evidence for person-specific semantic structure, and not for any property of the words themselves. Demonstration 2 makes the same point in a different way. The model's parameters were identical across the three conditions and only its lexicon was exchanged. Thus, the improvement in cue–target-level prediction cannot be attributed to differences in the flexibility with which the models had opportunity to accommodate the data.

The construction of the representations reinforces this reading. The global negative and distribution of associations transformations applied to all three spaces are designed to remove base-rate occurrence structure (Johns et al., 2019). The GN transformation subtracts the co-occurrence that would be expected given how often words appear, and the row-wise standardization in the DOA transformation corrects for the tendency of high-frequency words to show elevated co-occurrence values simply because they occur in many sentences. What survives these transformations is distinctive co-occurrence: the degree to which two words appear together more than their individual rates of occurrence would predict. Raw exposure is, in a formal sense, what the transformations discard. And if aggregate exposure were nevertheless the operative variable, the generic spaces should have performed at least as well as the personalized ones. The all-books corpus spans all eight genres and 1.73 billion words; it encodes the frequency structure of the language these participants read at least as faithfully as any individual's weighted sample of it. What the personalized corpora add is not more exposure information but differently weighted contexts.

We would therefore resist framing familiarity and distributional structure as competing explanations. They are better understood as different levels of description of the same underlying variable. Reading within a genre does two things at once: it increases how often a reader encounters particular words, and it repeatedly places those words in the company of other related words. The first is what a frequency count measures, in a single number per word. The second is what a distributional space represents, as a geometry over the whole lexicon. Our hypothesis throughout has been that the second is the operative one for cued recall, because cued recall concerns relations between items rather than items in isolation, and because the retrieval mechanism of the eCFM is explicitly a similarity computation over stored traces. A frequency account can say which words are easy to remember; it cannot say which cue will retrieve which target for which reader. That is the level at which the present effects appear. This position is continuous with a broader line of work showing that frequency is an imperfect summary of linguistic experience and remembering, and that measures preserving the contexts in which words occur (i.e., their contextual diversity) account for lexical behaviour better than counts of occurrence alone (Adelman et al., 2006; Johns, 2021; Johns et al., 2019; Johns & Jones, 2010; Jones et al., 2012). On that view, personalized semantic structure is not a proxy for familiarity; familiarity is a compressed summary of the experiential record that a distributional space represents more completely.

What remains unresolved should be stated plainly. Because reading history shapes exposure and co-occurrence structure jointly, the present design cannot fully orthogonalize them: a reader of fantasy encounters *dragon* more often and also encounters it near *sorcery* more often than a nonreader of that genre. What the design does establish is that the person-specific component of the effect cannot be attributed to properties of the items, and what the model establishes is that relational structure is sufficient to improve prediction when combined with a retrieval mechanism that operates on similarity. Separating exposure from structure more sharply would require materials in which the two are deliberately decorrelated within an individual's corpus, and so, nevertheless, we regard that as a valuable direction for future work.

Theoretical implications for model evaluation

The present findings have implications for how computational models of memory should be evaluated. The standard practice in the field is to report aggregate fit statistics, comparing model predictions to group-level averages, attending to condition-specific but overlooking item-specific differences within experimental data sets. Demonstration 2 shows that this practice can obscure representational nuances that are visible only at the cue–target level (an observation that Dienes, 1992, argued and that Jamieson & Mewhort, 2010, took up as a critical step for

model development in the more constrained context of artificial grammar learning). Because the decision threshold T was optimised jointly across all three representational conditions to match aggregate proportions, the three versions of the eCFM tested in this paper begin from the same aggregate baseline by construction; the advantage of the personalized representation emerged at the cue–target level and would have been invisible to an analysis that stopped at group means.

This suggests that models claiming to account for the role of semantic knowledge in memory should potentially be evaluated against behaviour at the level of individual cue–target pairs, rather than only against condition means or serial position functions. The distinction we intend concerns the units entering the evaluation, not the use of summary statistics: correlations and explained variance may remain the natural way to express model–data correspondence, but what they are computed over matters. In the present case they are computed across 3456 cue–target × outcome combinations, and the advantage of the personalized representation emerged at that level while being statistically undectable invisible in the aggregate proportions, which were matched across conditions by construction. We should be clear that this criterion is one our own model meets only partly. As reported above, correspondence within individual outcome measures was near zero for all three representations, so the model does not predict fine-grained differences in item difficulty within a measure; what it predicts is the distribution of outcome types across pairs. Item-level prediction is better understood as a standard the field is moving toward than one any current model, including ours, has met (Chang et al., 2025; Guitard et al., 2025a, 2025b, Guitard et al., 2026; Johns, 2026; Johns et al., 2012; Reid & Jamieson, 2023; Reid et al., 2026). Conversely, models that lack a principled model for representation, including those that rely on random vectors, may produce adequate aggregate fits by a mechanism that is theoretically uninformative about the contribution of semantic structure to individual memory outcomes (Johns & Jones, 2010).

Limitations

Several limitations of the present work should be acknowledged. First, our memory paradigm was a simple cued recall task with a constrained word pool drawn from the experimental materials. Whether the advantage in using personalized-embeddings generalises to serial recall, free recall, recognition memory, paired-associate learning over longer retention intervals, autobiographical recall, or naturalistic memory remains an empirical question. In particular, free recall depends heavily on inter-item semantic associations that govern clustering and contiguity (Howard & Kahana, 2002; Polyn et al., 2009), and these may or may not benefit from personalization as cleanly as we observed with cued recall. Second, the eCFM is one of several plausible process architectures within which personalized representations could be tested. Retrieved-context models (Howard & Kahana, 2002; Lohnas et al., 2015; Polyn et al., 2009), other instance-based models (MINERVA 2; Hintzman, 1988; SAM; Raaijmakers & Shiffrin, 1981), and global-matching recognition models (Osth & Dennis, 2024) make different commitments about how stored knowledge is accessed at retrieval, and the magnitude of the personalization advantage may depend on those commitments (Chang et al., 2025).

Thirdly, the summation method we used to encode cues and targets in episodic traces could be further considered. Summation follows from the instance-theoretic view of the trace described above, but it does not capture properties that other formal treatments of association were designed to capture. Convolution-based schemes, in which the association is a unique representation from and between two items have been developed as explicit statements on the special nature of associative encoding and retrieval (Eich, 1982; Jones & Mewhort, 2007; Murdock, 1982). Cox and Criss (2020) have shown that a dynamic model in which associative features are built from pairwise conjunctions of co-active item features represents a direct theory on the nature of association that accounts for the observed relationship between item similarity and associative strength. In contrast, the eCFM assumes that episodic memory traces record cue-target cooccurrence and that association emerges at retrieval: presenting the cue retrieves the trace in which it rests and carries along the target as a corollary. Whether the advantage of personalized representations reported here obtains under convolution-based or co-active associative encoding is an open question and one worth pursuing. We note, however, that our process model and model parameters were held constant across all three representational conditions and so the differences we report that show an advantage personalized word representations are not attributable to that choice; the limitation bounds the generality of the present implementation rather than the interpretation of the representational contrast.

Finally, our sample, although large by memory-research standards, was drawn from English-speaking adults in two countries; whether the same pattern holds across languages, age groups, and cultural contexts requires separate investigation. Lastly, our characterization of personalization operates at the level of broad genre exposure; finer-grained personalization, perhaps based on individual language production logs, sociolinguistic profiles, or behavioural measures of semantic structure (Aeschbach et al., 2025; Johns, 2023, 2024), may reveal additional dimensions of representational variation that genre weighting cannot capture. We believe records of people's daily language interactions might yield similar effects to the ones reported here based on reading habits, but might have the advantage of amplifying social factors that influence people's different understandings of word meanings.

Future directions

These limitations point to several directions for future work. Methodologically, the field would benefit from validated, multi-dimensional instruments for indexing individual linguistic experience, going beyond self-reported reading frequency to incorporate vocabulary breadth, register exposure, and cumulative reading depth. For example, at first cut, if Reddit users could be identified and brought into the lab to participate in experiments, there might be an opportunity to derive representations for individuals based on their Reddit reading and writing histories to predict differences in performance at more objective and fine tuned scales (Johns, 2021, 2023, 2024, 2025). Computationally, the same logic applied here to BEAGLE-derived representations could be applied to contextualised embeddings from large language models, allowing tests of whether contextual semantic structure also benefits from individualization. Theoretically, the present framework can be extended to ask not only whether personalization improves prediction but whether it explains qualitative phenomena that population-level models struggle with, including individual differences in false recall susceptibility, in retrieval-induced forgetting, and in cohort-level changes in memory across the lifespan (Teles et al., 2025; Wulff et al., 2019).

At a broader level, the present findings reinforce a long-standing view of the relationship between cognitive architecture and the experiential environment. That the eCFM improved its account of individual performance without any change to its encoding, retrieval, or decision assumptions, assumptions inherited largely intact from MINERVA 2, emphasizes the point: the mechanisms articulated four decades ago remain adequate, and what changed was the knowledge over which they operate. This is continuous with Simon's (1956) argument and with Nosofsky's (1985, 1986) demonstration that individualized representations combined with a shared process model yield high-precision accounts of individual remembering behaviour. What we would add is not a revision of that principle but an observation about modelling practice: the principle has been more readily accepted in the abstract than implemented at the level of individual experiential histories, in part because the corpora and instruments needed to do so have only recently become available. A genuinely individual-centred theory of memory will need to specify both what the system does and what knowledge it does it over, and the latter cannot be assumed to be uniform across individuals.

A caution is warranted about the limiting case of this programme. Personalization pursued without constraint risks producing models that predict individual behaviour with great precision while forfeiting any general account of it. We regard the goal as the conjunction rather than the trade-off: an ideal theory should account for the whole and for each of its parts, such that the aggregate is recovered by the sum of the individual accounts whilst each individual account remains separately predictive. Aggregate performance is then not a separate target requiring its own model but a consequence of the individual ones, and the two levels of description are reconciled rather than traded against one another.

Two features of the present approach serve that objective. First, personalization enters through measured input rather than through free parameters: the eCFM was run with a single fixed parameter set applied identically across all representational conditions, so the personalized models had no additional latitude with which to accommodate the data. This distinguishes personalization by experiential history from personalization by parameterization fitting: the same distinction that allowed Nosofsky's (1985, 1986) individualized representations to increase precision without reducing the generality of the underlying theory. Second, the personalized spaces are not idiosyncratic constructions but point within a common, low-dimensional family: each participant's representation is a reweighting of the same corpora by the same learning algorithm, differing only by the proportion with which each subcorpus was weighted. Individuals differ in their coordinates within a shared representational structure, not in kind, and the shared process architecture operates identically over all of them. The general account resides in the family of representations and the mechanisms defined over it. From our perspective, modelling the aggregate as arising from the individualized representational spaces is more theoretically sound than modelling the individualized representational spaces as special parameterized cases of the aggregate.

The criterion that keeps this programme disciplined is generalization. A personalized representation earns its theoretical keep when it predicts held-out items, tasks, or occasions for that individual, and when the aggregate of such predictions reproduces group-level performance, not when it merely improves fit to the data from which it was derived (Navarro, 2019; Pitt et al., 2002; Roberts & Pashler, 2000). The objective is not maximal personalization but the level of representational granularity at which individual prediction improves while the general account is preserved. Thus, a future test of our account would involve testing if the account developed in cued recall generalises to other remembering behaviours as well.

A further direction concerns the items themselves. Because each genre-specific space is constructed by the same algorithm from the same pipeline, the semantic neighbourhood of a given word can be compared directly across genres: the five words most closely related to a target in the fantasy space need not be the five most closely related to it in the crime fiction space, and the degree of that divergence is directly computable. On the present account, words whose neighbourhoods shift most across genres are precisely those for which individual reading history should matter most, and selecting such items in advance would permit a considerably more sensitive test than the broad sampling used here where items were constrained only to be diagnostic within a single genre.[4] More generally, comparing a word's neighbourhood across spaces offers a way to characterize how individual experience shapes the representation of each word rather than of the lexicon as a whole. We intend to pursue this in future work.

The deeper question motivating this work is one that no single empirical study can settle: how does the human mind become the mind it is? Memory, perhaps more than any other cognitive faculty, is the place where the structure of experience and the structure of cognition meet. Every word an individual knows is connected, however indirectly, to every other word that individual has previously encountered, and to the indirect patterns of co-occurrence that shape the geometry of meaning over a lifetime. Computational models of memory have made remarkable progress in characterizing the processes by which traces are formed and retrieved, but they have done so largely by abstracting away from the variability of the individual experiential record. The present study suggests that this abstraction, while pragmatically convenient, comes at a theoretical cost. Closing the gap between process and representation, between the universal mechanisms of memory and the particular history of the individual remembering, will require additional, sustained, and programmatic effort: better instruments for measuring experience, better corpora that reflect individual exposure, better representational models that can ingest individualized input, and better process models that can use that input to generate item-level predictions. We have taken one step in that direction, but many more will be needed.

Conclusion

The title of this article was chosen to be slightly tongue in cheek. We claim only that what a person reads contributes in a measurable and theoretically tractable way to what that person remembers. The field of memory research has made extraordinary progress over the past half century, and the shoulders on which we stand have taught us a great deal about the mechanisms of encoding, storage, and retrieval. We do not argue that this work is wrong. We argue that it is incomplete in a specific and remediable way. A large portion of the experience that shapes human cognition, the books people read, the language environments they inhabit, the semantic geometries they build up over a lifetime, have often been set aside as a source of noise rather than treated as a source of signal. The present demonstrations are a first, also incomplete step toward correcting that oversight. Our operationalization of individual experience is coarse, our paradigm is constrained, and our model is one of many plausible accounts of the processes underlying cued recall. We offer this work not as a final solution but as a proof of concept: the infrastructure for connecting what people read to what they remember now exists, and it produces detectable, theoretically interpretable results even at rough granularity. We hope the present work gives the field motivation to invest in that programme.

Cognitive science has made extraordinary progress by studying the mind in abstraction from the particulars of individual experience. That abstraction has been productive, but it has a cost. The mind that encodes, stores, and retrieves is always a mind shaped by a particular history of encounters with language and the world. The present work suggests imperfectly but, we think, compellingly, that we cannot fully understand human memory without taking that history (i.e., knowledge) into account. Experience is not noise; t is part of the signal.

Open practices statements

The demonstrations reported in this article were theoretically motivated but were not preregistered. The stimuli are described in the text and reproduced in full in Appendix A. All materials, including stimuli, analysis scripts, and simulation code are available on the Open Science Framework page associated with this article at OSF. Because of their large size, the vector files could not be deposited on OSF but are available from the first or last author upon request.

CRediT authorship contribution statement

Dominic Guitard: Writing – original draft, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Randall K. Jamieson:** Writing – review & editing, Validation, Software, Methodology, Conceptualization. **Jean Saint-Aubin:** Writing – review & editing, Methodology, Funding acquisition, Conceptualization. **Brendan T. Johns:** Writing – review &

[4] We thank an anonymous reviewer for this suggestion.

editing, Writing – original draft, Software, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada through a Discovery Grant awarded to Randall K. Jamieson (RGPIN-2025-06861), Jean Saint-Aubin (RGPIN-2023-05943), and Brendan T. Johns (RGPIN-2020-04727).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Target Words and Genre-Specific Cue Words

Table A1 presents the complete set of 144 cue-target pairs used in the experiment. Each row lists one target word (column 1, bold) and the eight genre-specific cue words assigned to it, one per genre. All 1296 words across the table are unique: no word appears more than once. Target words are listed in alphabetical order.

Table A1
Target Words and Genre-Specific Cue Words for All 144 Stimulus Pairs.

Target	Mystery	Fantasy	Horror	Literary Fiction	Romance	Science Fiction	Thriller	Non-fiction
ahold	gyro	pigtail	venison	ballots	coffees	kennels	mormon	glee
airlocks	griefs	falcons	renegade	alleys	coolies	entryway	widget	solidity
amulets	sutures	rebound	decanter	dearth	grist	livery	slang	codon
apaches	mohair	gammon	crypto	warder	deities	cabby	rapier	recce
archduke	drapes	jerseys	smoothie	robbery	canons	spinster	lanyard	kaisers
artistry	truffles	postmark	lullaby	hackers	asterisk	inflow	flukes	damsel
axle	gala	purses	bums	prune	chintz	ayah	servo	pawn
bankroll	juncture	parades	rowers	snippets	drifts	sirloin	locust	shooter
baptists	herpes	quilts	inquest	milady	cinder	thorium	turnout	browns
bast	creeps	peep	beater	epics	mucus	fixer	lilacs	pronto
bastions	surfeit	totems	wrestler	egoism	donkeys	cypher	epistle	bordello
berserk	fable	slayer	cabinets	giraffes	zigzags	boars	exporter	fatigues
bibles	argent	ticks	ironwork	entirety	seminole	deans	limbo	firearm
bios	snails	typhus	palais	minors	karate	koran	earful	logon
blobs	poodles	chine	magna	fishery	sleeper	rodents	wallop	shack
blowup	armpits	gals	motorist	nodules	brazier	choirs	choline	stirrups
bodywork	gnashing	capering	tankard	subways	manpower	monorail	grandmas	adjunct
bookmark	farthing	workup	lyceum	shipload	teamwork	jaybird	widower	toolkit
brutes	guises	wildcats	beamer	inductor	hexes	stews	eyeglass	essences
buildup	papas	washroom	shorty	bison	firmness	epsilon	icebergs	froth
butane	wrappers	gusts	coliseum	gags	haddock	fags	chagrin	vapours
buyout	prowess	imager	holiness	lawgiver	takers	zebras	baronet	breaches
byline	breakout	cocker	reels	redo	township	villain	rood	vitals
byword	vertebra	haircuts	shreds	puce	triads	heron	silt	hilarity
cannery	flickers	stash	pitter	snapper	commune	meatball	apogee	windmill
casks	betimes	louse	koala	rectum	idiom	jogger	brisket	taverns
chits	biopsy	esprit	pecks	morph	gulag	encoder	beards	shovel
choir	dung	scythes	pullout	sampler	yucca	alpaca	critter	stucco
cilantro	omission	sailboat	madrigal	sconce	cadets	seaboard	feasts	ruffians
clout	wads	handset	scraper	rants	nary	pima	opus	porno
cobbler	censors	hypoxia	tyrants	scrubber	butts	camphor	cookery	liters
cordon	evasions	taffy	pest	webbing	pager	skimmers	foursome	fillet
coterie	wildfire	nettle	caftan	callus	bulldogs	octet	quilter	firemen
counsels	goings	jockeys	skyline	kleenex	mistral	agitator	oddness	volition
couplet	decency	hokey	smallpox	leotard	annex	chipmunk	giveaway	migraine
coupon	quips	missis	mousse	grands	bipeds	misuse	nozzles	breakage
craw	grafts	gums	ruckus	zephyr	hist	umlaut	tithe	clamor
creamer	leopards	vestry	aldermen	caving	dutchman	peyote	grandad	barnyard
cronies	vesper	chainsaw	pounder	fritters	landmass	senorita	skeet	leftist
cynics	funnels	guesses	leprosy	leaching	fracture	brats	inkling	clown
dace	chore	lupus	ovary	plugs	hoods	armful	coves	acacia
dents	codeine	hauberk	sliver	albino	antes	blister	recon	septum
devotees	slammer	lemons	hyphen	cossack	basins	oxytocin	slayers	mortuary
dint	fudge	lager	diaper	totem	florin	touts	buds	semen
ditty	fuse	ploys	jugular	legwork	hampers	colds	gulls	cramp
dogwood	boarders	kilter	hygiene	mobsters	tonne	shrew	tryst	spurs
drifters	airboat	streaker	keyhole	revivals	meiosis	eyelash	platters	poplar
drummer	thrush	tonic	snippet	guano	flasks	courant	readout	trilby
dynamism	bullocks	eardrums	cosine	conduits	agility	epithet	bosses	crusades
emporium	intuit	roommate	snipers	brocade	mainmast	divider	waivers	portside

(*continued on next page*)

Table A1 (*continued*)

Target	Mystery	Fantasy	Horror	Literary Fiction	Romance	Science Fiction	Thriller	Non-fiction
eunuchs	blinker	amnesty	bitches	abductor	arsehole	bumpkin	rhesus	vassals
favours	colloquy	finals	brahmin	platelet	chaperon	czechs	hyena	disdain
filament	agonies	portent	unrest	golfers	bringer	parrots	headset	pincer
flange	damask	britches	larks	animus	cobblers	westerns	okra	grease
flints	soles	spire	fete	dockside	rosebuds	scape	pimples	layover
foetus	caldron	unison	cashier	kinks	caliber	tequila	muslin	tepee
footpath	organist	poachers	helpings	almanac	puncture	upslope	wireman	wildcat
frauds	cottons	whist	florists	laziness	cartons	thumper	restorer	itch
freeways	sorrel	winder	tonnes	cotter	madman	bargains	cutaway	locales
garnish	lumps	linnet	panthers	orations	umbrage	pagodas	adherent	drizzle
girder	tonality	tractors	togs	nonstick	oilskin	cabling	uplift	charade
glitch	misstep	romp	midrash	axon	goodbyes	wayfarer	anthrax	compiler
glitters	fervor	cigars	navies	repeater	refunds	berets	ugliness	roubles
gristle	postures	apparel	arbiter	latency	rovers	autocrat	detours	saffron
halves	sashes	hurdles	rascals	crusts	eggplant	gent	seaweed	hitches
hamper	cougars	hoop	pansies	monogram	blunder	kamikaze	centrum	conveyor
handout	fascists	slowdown	kinsmen	sausages	hometown	furlough	toddlers	scamp
handouts	entreaty	seagulls	edifice	solver	robber	weathers	burdens	keynote
harpoon	culprits	twang	tetanus	poling	farts	marksman	wingtip	carrion
hetero	snail	incision	lite	thrasher	firefly	pups	girdle	ligation
hiatus	flats	hugs	collie	apoplexy	innings	faiths	hombres	carnage
highball	pigments	cutters	rampart	bistros	cradles	croupier	berths	dragoon
hobbit	shingles	gizmo	molds	staffer	scum	flamingo	exhale	pall
hobo	divan	napalm	samba	flaws	phial	quip	auras	foodie
hoovers	nitrite	psalms	mover	duvet	chianti	tatami	gents	hansom
humpback	egrets	hotshot	rarity	larynx	takeoff	turncoat	tugboat	paddles
humps	exiles	peaches	codex	gauges	scabs	phage	canary	donkey
hypo	exits	truism	swat	tweaks	humbug	scruff	rook	borax
jerkin	dodgers	chutes	hostelry	touristy	hardcore	tannery	hothead	stockade
jitters	scruples	sailings	tortilla	peephole	dweller	corundum	fiend	graviton
kneecaps	allure	foreplay	prisons	frills	tycoon	solenoid	liquors	kerosene
lather	trolleys	ostler	spotters	bluebird	benzoate	fathoms	prelates	guava
lawman	profiler	mutiny	hawkers	aprons	prithee	gypsum	charisma	anteroom
leavings	termite	noontime	gouges	kidneys	unease	raccoons	galleon	marquise
lepers	climber	earners	terrors	coconuts	minotaur	picker	voucher	squatter
longbow	caption	turnpike	nudity	tutelage	earshot	loophole	cretin	potions
lounges	mixers	moraine	typeset	tassels	environs	roost	hostler	vibes
madhouse	collard	outboard	chaser	frosting	beetles	cabaret	evasion	predawn
makings	cadaver	racquet	jawbone	furor	gelatin	teammate	nudge	grates
mandates	mealtime	haircut	licorice	uterus	earlobe	lingua	cedars	jurists
maniacs	yahoo	modifier	liturgy	footwork	tamarind	comanche	tribunal	topaz
mariner	gunny	adverb	issuance	eldritch	midbrain	ceviche	globule	shithead
mash	geyser	posse	madmen	caff	glands	idols	realty	hanger
millet	echelon	foci	orgy	baldy	peony	offenses	deltas	gossamer
misheard	bracer	wardroom	sockets	seance	girders	gimmick	moorland	clumps
mittens	orbiter	mesons	pitfalls	specs	filings	asbestos	larch	freedmen
neophyte	feeder	ricochet	sitter	inbound	trophies	unknowns	niacin	thorns
newsmen	middy	frond	guffaw	haves	zillion	checker	spawn	bayonet
nightie	fleas	woodpile	steels	mamas	cutout	doberman	ouija	vexation
notary	parquet	salons	cliques	notables	spades	durbar	escapees	audition
outbound	drifter	rancher	granola	aerosol	inaction	rapport	choppers	stalker
pastrami	raptor	capstan	cuckoos	climates	ailments	guilders	pathos	holies
peddlers	hindus	sickbay	moralist	bathtubs	snapshot	vicarage	suntan	jingle
pervert	racket	rickets	doorman	freebies	synch	falafel	turkeys	bicep
pianist	tassel	palaver	seraphim	legit	scoops	paras	flipper	baritone
pianos	bondsman	capsules	imprint	coauthor	londoner	demiurge	bravery	snare
plagues	juggler	dropout	hoofs	raisin	wingman	necrosis	maser	passover
plums	bleach	waltz	tinge	dummy	twister	lags	punches	honeys
podium	ideation	mantles	palmers	pyres	morel	viewport	hooch	tirade
polyglot	incest	retort	allusion	vortices	justices	unshaken	forceps	machete
pontoons	tadpole	debrief	henchman	ringer	satiety	mileage	beluga	fondant
poplars	flatiron	marten	tutorial	whiting	archery	dressers	jesuits	nugget
praises	quads	guppy	prickles	growers	atoll	fridges	bushman	enmity
pulley	grinder	gimp	looters	shin	nepotism	scone	misfit	bearings
pussies	trillion	honours	locale	atrophy	labia	burrito	hangers	breaker
redeemer	scalps	kidder	parsley	basalt	waveform	purveyor	sparring	nearness
regimen	possum	rigours	sorbet	quietude	brandies	loner	quarrels	bedtime
sabbath	forelock	hooker	raincoat	tedium	prisms	harpy	orchards	exodus
sadist	moire	stoat	airspeed	chives	compost	blokes	wayside	hangman
salami	gambit	fucker	berm	damper	clique	whalers	coolie	butchers
scarab	ruse	wicket	sheaf	pacifism	pennies	pong	destruct	grubs
shackles	comedies	pathogen	aqueduct	dungeons	aryans	lingerie	spinner	fugitive
sherbet	sayings	dryness	crawlers	inserts	boxing	fiddler	speck	padding
slabs	zippers	thimble	ardour	bazaars	orators	prong	ringers	nibbles
slings	dupes	lira	adage	rustlers	bursa	whisker	tonnage	heiress
slurs	steins	remorse	culprit	arks	mantra	demo	melee	ladle

(*continued on next page*)

Table A1 (*continued*)

Target	Mystery	Fantasy	Horror	Literary Fiction	Romance	Science Fiction	Thriller	Non-fiction
spawning	sickroom	suture	commode	swimsuit	padlocks	resins	bookie	frolic
spoons	solarium	pigment	kudos	emulsion	shippers	invader	vibrator	hocus
stardust	thickets	teatime	seepage	weirdo	contests	depots	stipend	loathing
stroller	myelin	dictator	bribes	triples	sprayer	duckling	briton	kimono
suckers	watchmen	dentists	drywall	atropine	emissary	thermos	namesake	sneer
tapers	legacies	polka	prank	wenches	foxes	airfare	cyborg	impeller
thistle	surveyor	goulash	fistful	calla	pharaoh	weakling	sluice	buffy
timidity	hatchets	sponsors	prefects	delicacy	clansmen	swifts	sacristy	acquit
tributes	loveseat	mayors	truckers	eyepiece	attics	postings	tremors	epitaph
trier	gaga	tariffs	portals	duffel	loch	smirks	cramps	eunuch
tumour	draughts	termites	numeral	coinage	academia	latte	beauties	heparin
turnkey	ironclad	grower	crayons	chaplain	bazooka	penlight	chisels	shovels
urinal	hedonism	maths	jubilee	emir	leniency	baguette	spiel	sleet
veils	llama	mould	smarts	lint	volleys	fumes	wrecks	snicker
washers	parapet	thiamine	saxons	conjurer	delirium	bangers	drivel	pedal
whores	taoist	charmer	corp	oxen	butters	artisan	fiddles	rioting
willies	bouncers	naivete	ancestry	drape	rifleman	bighorn	icicle	knell
wingspan	manure	drunks	bursar	ghouls	lepton	falconer	hawthorn	pyjamas

Note. Each row presents one target word (bold, leftmost column) and the eight genre-specific cue words associated with it, one per genre column. All targets and cues are English nouns between four and eight letters in length. Cue words were selected from genre-specific distributional semantic spaces (BEAGLE model; Jones & Mewhort, 2007, with Global Negative and Distribution of Associations transformations; Johns et al., 2019) such that each cue was associatively related to its target within its genre (cosine $\geq$0.15) but not in any other genre (cosine $\leq$0.08). No word appears more than once across the entire table. Stimuli are listed in alphabetical order by target word.

Appendix B. Reading Experience Questionnaire

The following pages reproduce the exact wording of the reading-experience questionnaire as it was presented to participants. The questionnaire was administered online prior to the memory task. Participants read an introductory screen explaining that the questions concerned their reading habits, and that their honest responses were important for the purposes of the study. All questions were mandatory; participants could not advance to the next page without providing a response. The questionnaire was divided into two pages. The first page (questions 1 to 4) assessed overall reading frequency, reading duration, and genre preferences. The second page (questions 5 to 12) assessed the approximate number of books read per year for each of the eight target genres.

1. How often do you typically read books?

Please select the option that best describes your reading frequency.

○ Every day.
○ At least 4 times per week.
○ At least 2 times per week.
○ About once per week.
○ Less than once per week.
○ Rarely.

2. When you read, how much time do you typically spend reading?

Please select the option that best describes a typical reading session.

○ About 15 min.
○ About 30 min.
○ About 1 h.
○ About 2 h.
○ More than 2 h.

3. About how many books do you read in a typical year?

Please include books from all genres combined.

○ None.
○ Fewer than 5.
○ Between 5 and 10.
○ Between 10 and 20.
○ More than 20.

4. Please indicate your preference for the following genres of books.

Genre	1 (Least preferred)	2	3	4	5	6	7	8 (Most preferred)
Mystery	○	○	○	○	○	○	○	○
Fantasy	○	○	○	○	○	○	○	○
Horror	○	○	○	○	○	○	○	○
Literature	○	○	○	○	○	○	○	○
Romance	○	○	○	○	○	○	○	○
Science Fiction	○	○	○	○	○	○	○	○
Thriller	○	○	○	○	○	○	○	○
Non-fiction	○	○	○	○	○	○	○	○

5. About how many Mystery books do you read per year?
○ None.
○ Fewer than 2.
○ 2 to 5.
○ 6 to 10.
○ More than 10.
6. About how many Fantasy books do you read per year?
○ None.
○ Fewer than 2.
○ 2 to 5.
○ 6 to 10.
○ More than 10.
7. About how many Horror books do you read per year?
○ None.
○ Fewer than 2.
○ 2 to 5.
○ 6 to 10.
○ More than 10.
8. About how many Literary Fiction books do you read per year?
○ None.
○ Fewer than 2.
○ 2 to 5.
○ 6 to 10.
○ More than 10.
9. About how many Romance books do you read per year?
○ None.
○ Fewer than 2.
○ 2 to 5.
○ 6 to 10.
○ More than 10.
10. About how many Science Fiction books do you read per year?
○ None.
○ Fewer than 2.
○ 2 to 5.
○ 6 to 10.
○ More than 10.
11. About how many Thriller books do you read per year?
○ None.
○ Fewer than 2.
○ 2 to 5.
○ 6 to 10.
○ More than 10.
12. About how many Non-fiction books do you read per year?
○ None.
○ Fewer than 2.
○ 2 to 5.
○ 6 to 10.
○ More than 10.

B.1. The global negative and distribution of associations transformations

To apply the GN transformation, a global negative vector is first constructed that reflects the base-rate occurrence values for each column in the matrix. This vector is calculated by summing the values within each column of the word-by-feature matrix:

$$\mathrm{GN}_j = \sum_{i=1}^{n} \mathrm{M}_{i,j} \tag{5}$$

where GN is the global negative vector, where M is the word-by-feature matrix defined above, j is the column being calculated, and i increments through all n rows in the matrix. The elements of the GN vector depend on the frequency with which each word occurs across sentences. The second step of the transformation is to unit-normalize this vector so that it has a total magnitude of 1. This is accomplished by dividing each element of the vector by the sum of all elements in the vector:

$$\mathrm{GN}_j = \frac{\mathrm{GN}_j}{\sum_{k=1}^{n} \mathrm{GN}_k} \tag{6}$$

where k increments through each index in the GN vector. The resulting transformed matrix contains a balanced mixture of positive and negative

information. Positive values indicate that the association between two words exceeds what would be expected based on their base-rate co-occurrence. In contrast, negative values or values close to zero indicate that the words do not share a distinctive association.

The GN transformation was developed to approximate the negative sampling procedure used in neural embedding models. The DOA transformation achieves a similar balance between positive and negative information through a different mechanism, primarily by emphasizing distinctive co-occurrence patterns through an assessment of the relative magnitude of matrix elements. To accomplish this, the DOA transformation first converts each column of the matrix into z-scores:

$$M_{i,j} = \frac{M_{i,j} - \mu_j}{\sigma_j}, \tag{7}$$

where i represents a row in the matrix, j represents a column, μ_j represents the mean of the column, and σ_j represents the standard deviation of the column. The values in the resulting matrix indicate how many standard deviations a given co-occurrence value deviates from the mean of the column. In this way, the transformation reflects how distinctive a particular word-word co-occurrence is relative to the other values within the same column.

However, the resulting z-scores are biased by the word frequency of a word, where higher frequency words will have high z-scores on average just because they tend to occur in more sentences. Thus, the second step is to transform each row into a z-score, in order to emphasize the important associations within a single word:

$$M_{i,j} = \frac{M_{i,j} - \mu_i}{\sigma_i}, \tag{8}$$

where μ_i is the mean of a word's row, and σ_i is the standard deviation of that row. The result of this transformation is that a word's representations contains normalized feature values.

Data availability

All data and code required to reproduce the findings are publicly available on OSF at https://osf.io/r7jad. Due to file size limitations, the complete set of vector representations is available from the corresponding author, first author, or last author upon reasonable request.

References

Acheson, D. J., Wells, J. B., & MacDonald, M. C. (2008). New and updated tests of print exposure and reading abilities in college students. *Behavior Research Methods, 40*(1), 278–289. https://doi.org/10.3758/BRM.40.1.278

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science, 17*(9), 814–823. https://doi.org/10.1111/j.1467-9280.2006.01787.x

Aeschbach, S., Mata, R., & Wulff, D. U. (2025). Measuring individual semantic networks: A simulation study. *PLoS One, 20*(8), Article e0328712. https://doi.org/10.1371/journal.pone.0328712

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the Dangers of Averaging Across Subjects When Using Multidimensional Scaling or the Similarity-Choice Model. *Psychological Science, 5*(3), 144–151. https://doi.org/10.1111/j.1467-9280.1994.tb00651.x

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

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

Chang, M., Johns, B. T., & Brainerd, C. J. (2025). True and false recognition in MINERVA2: Integrating fuzzy-trace theory and computational memory modelling. *Psychological Review, 132*(4), 857–894. https://doi.org/10.1037/rev0000541

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4*(1), 55–81. https://doi.org/10.1016/0010-0285(73)90004-2

Cowan, N., Bao, C., Bishop-Chrzanowski, B. M., Costa, A. N., Greene, N. R., Guitard, D., … Ünal, Z. E. (2024). The relation between attention and memory. *Annual Review of Psychology, 75*, 183–214. https://doi.org/10.1146/annurev-psych-040723-012736

Cox, G. E. (2026). Similarity as likelihood ratio: Coupling representations from machine learning (and other sources) with cognitive models. *Psychonomic Bulletin & Review, 33*(1), 6. https://doi.org/10.3758/s13423-025-02828-w

Cox, G. E., & Criss, A. H. (2020). Similarity leads to correlated processing: A dynamic model of encoding and recognition of episodic associations. *Psychological Review, 127*(5), 792–828. https://doi.org/10.1037/rev0000195

Cox, G. E., Hemmer, P., Aue, W. R., & Criss, A. H. (2018). Information and processes underlying semantic and episodic memory across tasks, items, and individuals. *Journal of Experimental Psychology: General, 147*(4), 545–590. https://doi.org/10.1037/xge0000407

Dienes, Z. (1992). Connectionist and memory-array models of artificial grammar learning. *Cognitive Science, 16*(1), 41–79. https://doi.org/10.1207/s15516709cog1601_2

Eich, J. M. (1982). A composite holographic associative recall model. *Psychological Review, 89*(6), 627–661. https://doi.org/10.1037/0033-295X.89.6.627

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin, 53*(2), 134–140. https://doi.org/10.1037/h0045156

Gatti, D., & Günther, F. (2026). Zero-shot pseudowords memorability via representational content analysis. *Psychonomic Bulletin & Review, 33*(4), Article 137. https://doi.org/10.3758/s13423-026-02875-x

Goldberg, Y., & Levy, O. (2014). Word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv*. https://doi.org/10.48550/arXiv.1402.3722

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*(2), 211–244. https://doi.org/10.1037/0033-295X.114.2.211

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., … Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology, 81*, 80–97. https://doi.org/10.1016/j.jmp.2017.09.005

Guitard, D., Cowan, N., Saint-Aubin, J., Reid, J. N., & Jamieson, R. K. (2026). Toward a comprehensive account of verbal memory: An embedded computational model across representational domains. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001618

Guitard, D., Gabel, A. J., Saint-Aubin, J., Surprenant, A. M., & Neath, I. (2018). Word length, set size, and lexical factors: Re-examining what causes the word length effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(11), 1824–1844. https://doi.org/10.1037/xlm0000551

Guitard, D., Miller, L. M., Neath, I., & Roodenrys, S. (2019). Does contextual diversity affect serial recall? *Journal of Cognitive Psychology, 31*(4), 379–396. https://doi.org/10.1080/20445911.2019.1626401

Guitard, D., Saint-Aubin, J., Reid, J. N., & Jamieson, R. K. (2025a). An embedded computational framework of memory: Accounting for the influence of semantic information in verbal short-term memory. *Journal of Memory and Language, 140*, Article 104573. https://doi.org/10.1016/j.jml.2024.104573

Guitard, D., Saint-Aubin, J., Reid, J. N., & Jamieson, R. K. (2025b). An embedded computational framework of memory: The critical role of representations in veridical and false recall predictions. *Psychonomic Bulletin & Review, 32*(5), 2004–2058. https://doi.org/10.3758/s13423-025-02669-7

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science, 14*(6), 1006–1033. https://doi.org/10.1177/1745691619861372

Harris, Z. S. (1954). Distributional structure. *Word, 10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Healey, M. K., & Kahana, M. J. (2014). Is memory search governed by universal principles or idiosyncratic strategies? *Journal of Experimental Psychology: General, 143*(2), 575–596. https://doi.org/10.1037/a0033715

Healey, M. K., Long, N. M., & Kahana, M. J. (2019). Contiguity in episodic memory. *Psychonomic Bulletin & Review, 26*(3), 699–720. https://doi.org/10.3758/s13423-018-1537-3

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers, 16*(2), 96–101. https://doi.org/10.3758/BF03202365

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93*(4), 411–428. https://doi.org/10.1037/0033-295X.93.4.411

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95*(4), 528–551. https://doi.org/10.1037/0033-295X.95.4.528

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology, 46*(3), 269–299. https://doi.org/10.1006/jmps.2001.1388

Jamieson, R. K., Avery, J. E., Johns, B. T., & Jones, M. N. (2018). An instance theory of semantic memory. *Computational Brain & Behavior, 1*(2), 119–136. https://doi.org/10.1007/s42113-018-0008-2

Jamieson, R. K., Crump, M. J. C., & Hannah, S. D. (2012). An instance theory of associative learning. *Learning & Behavior, 40*, 61–82. https://doi.org/10.3758/s13420-011-0046-2

Jamieson, R. K., Johns, B. T., Vokey, J. R., & Jones, M. N. (2022). Instance theory as a domain-general framework for cognitive psychology. *Nature Reviews Psychology, 1*(3), 174–183. https://doi.org/10.1038/s44159-022-00025-3

Jamieson, R. K., & Mewhort, D. J. K. (2010). Applying an exemplar model to the artificial-grammar task: String completion and performance on individual items. *Quarterly Journal of Experimental Psychology, 63*, 1014–1039. https://doi.org/10.1080/17470210903267417

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.

Johns, B. (2026). *Collective memory modeling: Utilizing individualized representations to drive processing models*. https://doi.org/10.31234/osf.io/km6jx_v1

Johns, B. T. (2021). Disentangling contextual diversity: Communicative need as a lexical organizer. *Psychological Review, 128*(3), 525–557. https://doi.org/10.1037/rev0000265

Johns, B. T. (2023). Computing word meanings by aggregating individualized distributional models: Wisdom of the crowds in lexical semantic memory. *Cognitive Systems Research, 80*, 90–102. https://doi.org/10.1016/j.cogsys.2023.02.009

Johns, B. T. (2024). Determining the relativity of word meanings through the construction of individualized models of semantic memory. *Cognitive Science, 48*(2), Article e13413. https://doi.org/10.1111/cogs.13413

Johns, B. T. (2025). Large-scale social feedback and language usage: A distributional analysis. *Computational Brain & Behavior, 8*, 503–516. https://doi.org/10.1007/s42113-025-00247-7

Johns, B. T., Dye, M., & Jones, M. N. (2021). Estimating the prevalence and diversity of words in written language. *Quarterly Journal of Experimental Psychology, 73*(6), 841–855. https://doi.org/10.1177/1747021819897560

Johns, B. T., & Dye, M. W. (2019). Gender bias at scale: Evidence from the usage of personal names. *Behavior Research Methods, 51*, 1601–1618. https://doi.org/10.3758/s13428-019-01234-0

Johns, B. T., & Jamieson, R. K. (2018). A large-scale analysis of variance in written language. *Cognitive Science, 42*(4), 1360–1374. https://doi.org/10.1111/cogs.12583

Johns, B. T., & Jamieson, R. K. (2019). The influence of place and time on lexical behavior: A distributional analysis. *Behavior Research Methods, 51*(6), 2438–2453. https://doi.org/10.3758/s13428-019-01289-z

Johns, B. T., Jamieson, R. K., & Jones, M. N. (2023). Scalable cognitive modelling: Putting Simon's (1969) ant back on the beach. *Canadian Journal of Experimental Psychology, 77*(3), 185–201. https://doi.org/10.1037/cep0000306

Johns, B. T., & Jones, M. N. (2010). Evaluating the random representation assumption of lexical semantics in cognitive models. *Psychonomic Bulletin & Review, 17*(5), 662–672. https://doi.org/10.3758/PBR.17.5.662

Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2012). A synchronization account of false recognition. *Cognitive Psychology, 65*(4), 486–518. https://doi.org/10.1016/j.cogpsych.2012.07.002

Johns, B. T., Mewhort, D. J. K., & Jones, M. N. (2019). The role of negative information in distributional semantic learning. *Cognitive Science, 43*(5), Article e12730. https://doi.org/10.1111/cogs.12730

Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology, 66*(2), 115–124. https://doi.org/10.1037/a0026727

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114*(1), 1–37. https://doi.org/10.1037/0033-295X.114.1.1

Kahana, M. J. (2020). Computational models of memory search. *Annual Review of Psychology, 71*, 107–138. https://doi.org/10.1146/annurev-psych-010418-103358

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Logan, G. D. (2021). Serial order in perception, memory, and action. *Psychological Review, 128*(1), 1–44. https://doi.org/10.1037/rev0000253

Lohnas, L. J., Polyn, S. M., & Kahana, M. J. (2015). Expanding the scope of memory search: Modelling intralist and interlist effects in free recall. *Psychological Review, 122*(2), 337–363. https://doi.org/10.1037/a0039036

Long, D. L., & Freed, E. M. (2021). An individual differences examination of the relation between reading processes and comprehension. *Scientific Studies of Reading, 25*(2), 104–122. https://doi.org/10.1080/10888438.2020.1748633

Lowder, M. W., & Gordon, P. C. (2017). Print exposure modulates the effects of repetition priming during sentence reading. *Psychonomic Bulletin & Review, 25*(1), 442–449. https://doi.org/10.3758/s13423-017-1248-1

Mar, R. A., & Rain, M. (2015). Narrative fiction and expository nonfiction differentially predict verbal ability. *Scientific Studies of Reading, 19*(6), 419–433. https://doi.org/10.1080/10888438.2015.1069296

Martí, L., Wu, S., Piantadosi, S. T., & Kidd, C. (2023). Latent diversity in human concepts. *Open Mind, 7*, 79–92. https://doi.org/10.1162/opmi_a_00072

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *arXiv*. https://doi.org/10.48550/arXiv.1310.4546

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review, 89*(6), 609–626. https://doi.org/10.1037/0033-295X.89.6.609

Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition, 18*(3), 251–269. https://doi.org/10.3758/BF03213879

Navarro, D. J. (2019). Between the Devil and the Deep Blue Sea: Tensions Between Scientific Judgement and Statistical Model Selection. *Computational Brain & Behavior, 2*, 28–34. https://doi.org/10.1007/s42113-018-0019-z

Nosofsky, R. M. (1985). Overall similarity and the identification of separable-dimension stimuli: A choice model analysis. *Perception & Psychophysics, 38*(5), 415–432. https://doi.org/10.3758/BF03207172

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115*(1), 39–57. https://doi.org/10.1037/0096-3445.115.1.39

Osth, A. F., & Dennis, S. (2024). Global matching models of recognition memory. In M. J. Kahana, & A. D. Wagner (Eds.), *The Oxford handbook of human memory*. Oxford University Press.

Osth, A. F., Zhang, L., & Williams, S. (2026). A global matching model of choice and response times in the Deese–Roediger–Mcdermott semantic and structural false recognition paradigms. *Psychological Review, 133*(4), 771–819. https://doi.org/10.1037/rev0000596

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). https://doi.org/10.3115/v1/D14-1162

Petilli, M. A., Rodio, F. M., Gatti, D., Marelli, M., & Rinaldi, L. (2026). Multimodal prior knowledge determines false memory formation. *Journal of Experimental Psychology: General, 155*(2), 327–343. https://doi.org/10.1037/xge0001852

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109*(3), 472–491. https://doi.org/10.1037/0033-295X.109.3.472

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review, 116*(1), 129–156. https://doi.org/10.1037/a0014420

R Core Team. (2026). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88*(2), 93–134. https://doi.org/10.1037/0033-295X.88.2.93

Recchia, G., Sahlgren, M., Kanerva, P., & Jones, M. N. (2015). Encoding sequential information in semantic space models: Comparing holographic reduced representation and random permutation. *Computational Intelligence and Neuroscience, 2015*, Article 986574. https://doi.org/10.1155/2015/986574

Reid, J. N., Guitard, D., & Jamieson, R. K. (2026). MINERVA OPS: A computational framework for the representation and recognition of orthographic, phonological, and semantic associates. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-025-02792-5

Reid, J. N., & Jamieson, R. K. (2023). True and false recognition in MINERVA 2: Extension to sentences and metaphors. *Journal of Memory and Language, 129*, Article 104397. https://doi.org/10.1016/j.jml.2022.104397

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*(2), 358–367. https://doi.org/10.1037/0033-295X.107.2.358

Saint-Aubin, J., Yearsley, J. M., Poirier, M., Cyr, V., & Guitard, D. (2021). A model of the production effect over the short-term: The cost of relative distinctiveness. *Journal of Memory and Language, 118*, Article 104219. https://doi.org/10.1016/j.jml.2021.104219

Schad, D. J., Betancourt, M., & Vasishth, S. (2021). Toward a principled Bayesian workflow in cognitive science. *Psychological Methods, 26*(1), 103–126. https://doi.org/10.1037/met0000275

Schwering, S. C., Ghaffari-Nikou, N. M., Zhao, F., Niedenthal, P. M., & MacDonald, M. C. (2021). Exploring the relationship between fiction reading and emotion recognition. *Affective Science, 2*(2), 178–186. https://doi.org/10.1007/s42761-021-00034-0

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—Retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*(2), 145–166. https://doi.org/10.3758/BF03209391

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review, 63*(2), 129–138. https://doi.org/10.1037/h0042769

Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology, 41*, 1–19. https://doi.org/10.1146/annurev.ps.41.020190.000245

Stanovich, K. E., & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly, 24*(4), 402–433. https://doi.org/10.2307/747605

Stoet, G. (2010). PsyToolkit: A software package for programming psychological experiments using Linux. *Behavior Research Methods, 42*(4), 1096–1104. https://doi.org/10.3758/BRM.42.4.1096

Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology, 44*(1), 24–31. https://doi.org/10.1177/0098628316677643

Teles, M., Moore, I., & Kenett, Y. N. (2025). How retrieval processes change with age: Exploring age differences in semantic network and retrieval dynamics. *Canadian Journal of Experimental Psychology, 79*(1), 109–123. https://doi.org/10.1037/cep0000332

Thompson, B., Roberts, S. G., & Lupyan, G. (2020). Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour, 4*, 1029–1038. https://doi.org/10.1038/s41562-020-0924-8

Unsworth, N. (2019). Individual differences in long-term memory. *Psychological Bulletin, 145*(1), 79–139. https://doi.org/10.1037/bul0000176

Unsworth, N., Miller, A. L., & Robison, M. K. (2019). Individual differences in encoding strategies and free recall dynamics. *Quarterly Journal of Experimental Psychology, 72*(11), 2495–2508. https://doi.org/10.1177/1747021819847441

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved R̂ for assessing convergence of MCMC. *Bayesian Analysis, 16*(2), 667–718. https://doi.org/10.1214/20-BA1221

Vermeiren, H., Vandendaele, A., & Brysbaert, M. (2023). Is the author recognition test a useful metric for native and non-native English speakers? An item response theory analysis. *Behavior Research Methods, 55*(3), 1162–1186. https://doi.org/10.3758/s13428-021-01556-y

Wang, X., & Bi, Y. (2021). Idiosyncratic tower of babel: Individual differences in word-meaning representation increase as word abstractness increases. *Psychological Science, 32*(10), 1617–1635. https://doi.org/10.1177/09567976211003877

Williams, E. J. (1959). The comparison of regression variables. *Journal of the Royal Statistical Society: Series B (Methodological), 21*(2), 396–399. https://doi.org/10.1111/j.2517-6161.1959.tb00346.x

Wulff, D. U., De Deyne, S., Jones, M. N., Mata, R., & The Aging Lexicon Consortium. (2019). New perspectives on the aging lexicon. *Trends in Cognitive Sciences, 23*(8), 686–698. https://doi.org/10.1016/j.tics.2019.05.003