Information and Processes Underlying Semantic and Episodic Memory Across Tasks, Items, and Individuals

Gregory E. Cox
Syracuse University

Pernille Hemmer
Rutgers University

William R. Aue
Purdue University

Amy H. Criss
Syracuse University

Abstract

The development of memory theory has been constrained by a focus on isolated tasks rather than the processes and information that are common to situations in which memory is engaged. We present results from a study in which 453 participants took part in five different memory tasks: single-item recognition, associative recognition, cued recall, free recall, and lexical decision. Using hierarchical Bayesian techniques, we jointly analyzed the correlations between tasks within individuals—reflecting the degree to which tasks rely on shared cognitive processes—and within items—reflecting the degree to which tasks rely on the same information conveyed by the item. Among other things, we find that (a) the processes involved in lexical access and episodic memory are largely separate and rely on different kinds of information; (b) access to lexical memory is driven primarily by perceptual aspects of a word; (c) all episodic memory tasks rely to an extent on a set of shared processes which make use of semantic features to encode both single words and associations between words; (d) recall involves additional processes likely related to contextual cuing and response production. These results provide a large-scale picture of memory across different tasks which can serve to drive the development of comprehensive theories of memory.

Keywords: Memory; individual differences; item analysis; Bayesian statistics; principal components analysis.

Introduction

The goal of research into human memory is to understand the information that is contained in memory as well as the processes used to encode, store, retrieve, and make use

of that information. These various aspects of memory are only visible when filtered through particular tasks, that is, specific sets of stimuli, decisions, and actions. Many commentators (e.g., Hintzman, 2011) have recently noted, however, that a focus on individual tasks has impeded progress toward a more comprehensive picture of human memory. While we believe that such focus is critical to the larger research program by which scientific understanding of memory is developed, we believe it is equally crucial to periodically take a step back to appreciate how the fine-grained images provided by particular tasks can be overlaid to form a cohesive mosaic of human memory. For example, the literature studying lexical retrieval is mostly separate from that studying episodic retrieval, making it hard to know what is the relationship between retrieval from semantic and episodic memory (cf. Hintzman & Curran, 1997; Nelson & Shiffrin, 2013). Similarly, a broad understanding of memory is crucial for understanding what diagnostic tasks actually indicate regarding memory deficits, such as those that result from age, injury, or disease (e.g., Siedlecki, 2007; Healey & Kahana, 2016). The purpose of the present work is to examine how performance on different memory tasks is correlated with respect to both the processes engaged by individual participants and the information conveyed by particular items.

In practice, it is difficult to disentangle the roles played by the information in memory and the processes acting on that information. Consider, for example, an episodic recognition task in which a participant studies a list of words and is tested by showing her a word and asking her whether or not it was on the list she just studied. If she says yes, is that because 1) she was able to find a match between the word on that trial and one stored in memory? Or was it because 2) that word generally seems familiar and she would have said yes to it regardless? If the answer is 1, we can attribute performance on that trial to the memory processes she engaged in the episodic recognition task. If, however, the answer is 2, performance is due more to the information contained in the item presented on that trial and how it was made manifest by the episodic recognition task. To resolve this ambiguity, one must collect many observations of each individual participant interacting with many different items as well as each item being processed by many different individuals. In this way, we can distinguish between effects that are consistent with respect to each item (irrespective of the participant) from those that are consistent with respect to each individual (irrespective of the item).

A single task, even with many observations of items and individuals, still only represents one filter through which to view memory. To build up a more complete mosaic, we must examine how memory performance is related *between* different memory tasks. For example, if participants who demonstrate good recall ability also demonstrate good recognition

All data, model code, and stimuli are freely available online at `https://osf.io/dd8kp/`.

This work was supported by National Science Foundation grant BCS-0951612.

Portions of this work in development were presented at the Context and Episodic Memory Symposium in 2013 and 2017, at the Society for Mathematical Psychology in 2015, and at the Psychonomic Society in 2016.

Correspondence concerning this article should be addressed to Gregory E. Cox, Department of Psychology, 430 Huntington Hall, Syracuse University, Syracuse, NY 13244-2340. E-mail: `gregcox7@gmail.com`.

ability—that is, if performance on these tasks is *correlated* at the level of individuals—this indicates that these two tasks rely on similar memory processes. If, however, individuals' performance on these tasks is unrelated, we can infer that they rely on different memory processes. We can draw analogous inferences with regard to the information contained in items that is relevant to memory: If an item is associated with good performance in both recognition and recall, that means it carries information that supports the performance of both of these tasks. Conversely, if recognition and recall of an item are unrelated, then these two tasks make use of different aspects of the information conveyed by the item. Of course, a correlation—or lack thereof—between any given pair of tasks could still result from other sources of variability or from domain-general characteristics (intelligence, engagement, motivation, etc.), which is why it is critical to examine not just the pairwise correlations among tasks but the entire pattern of correlations amongst many tasks in order to better identify meaningful correlations and reject spurious ones. The present study is the first to jointly examine the patterns of correlations across multiple memory tasks for both individuals and items.

Individual memory manifest in different tasks

There have been some previous efforts to study how memory is deployed across tasks. Perhaps the closest analogue to the present study is one reported by Underwood, Boruch, and Malmi (1978). In their study, participants engaged in a variety of experimental tasks, including several different types of memory tasks with several different types of verbal materials designed to emphasize particular attributes of those materials (Underwood, 1969). For example, each participant would complete several blocks of free recall, some with lists of random words, some with lists of paired semantic associates, and some with lists consisting of exemplars of a single category. Factor analysis of outcome measures from each of these tasks lead Underwood et al. (1978) to conclude that five factors best explained their results; these five factors generally represented different task types (paired associates, free recall, memory span, recognition, and verbal discrimination) and did not discriminate between the attributes of the items used in each task. As the original authors argued, the fact that their study was sensitive to task type and insensitive to the information present in each task resulted from the fact that task and information were confounded within an individual: Participants likely engaged in different strategies when, for example, all study words were semantically related from when they were random. Our methods mitigate against this possibility by informing participants of the task only *after* each study phase.

In contrast, certain aging studies have identified factors corresponding to the information contained within a memory task, but without any corresponding task differences. Siedlecki (2007) conducted recognition, cued recall, and free recall tasks on a population with a large age range using verbal, figural (line drawing), and spatial (grid location) stimuli. She found that, when all tasks and information types were analyzed simultaneously, three factors emerged corresponding to the three types of to-be-remembered information. For each information type, correlations in performance among the three tasks were quite high; only in the verbal domain was there evidence for separation between tasks, specifically for a distinction between free recall and cued recall/recognition. These results indicate that the same memory processes are engaged across all types of stimulus materials—hence all tasks are strongly correlated—with the principal differences owing to the kind of information that

is stored in memory.

Studies of the relationship between memory and intelligence lend support to the idea that all memory tasks rely, to an extent, on a single set of core processes. Unsworth (2010) found that a model in which factors corresponding to working memory, recognition, and recall were all related by a single higher-order memory factor provided the best account of his data and that memory was more strongly related to fluid intelligence than crystallized intelligence. To the extent that memory tasks relied on crystallized intelligence, it is likely due to the fact that they employ verbal materials, and vocabulary knowledge is a key component of crystallized intelligence.

Overall, previous studies investigating the nature of correlations between memory tasks have found 1) that while different memory tasks can often be distinguished, they tend to rely on a shared set of underlying processes; and 2) that different kinds of information can yield differences in memory performance, but only when those difference are extreme (words versus line drawings) or when they are not confounded with task. Although most subsequent modeling developments have focused on individual memory tasks, the global memory models of the 1980's, such as SAM (Search of Associative Memory; Raaijmakers & Shiffrin, 1981; Gillund & Shiffrin, 1984) or the Matrix model (Humphreys, Bain, & Pike, 1989), began from the assumption that all memory tasks rely on the same information structures encoded in memory. Such structures include associations between items and other items and between items (or sets of items) and the contexts in which they occur. Tasks were assumed to differ only to the extent that they required different retrieval cues, for example, recall tasks involved primarily context cues whereas recognition tasks required both context and item cues. These models, however, made no specific commitments to how different kinds of item information (e.g., modality or semantic relations) were encoded in memory or how they affected retrieval cuing, but it seems reasonable to assume that when the task requirements are known beforehand (as in the study of Underwood et al., 1978), it becomes impossible to distinguish between encoding and retrieval effects. As described below, our study is designed to explicitly separate variability due to the use of different cognitive processes from that due to the information carried by different memory items, thereby helping to resolve a number of puzzling contradictions in how different items are treated in different situations.

Item information relevant to memory

Much research has focused on the effect that different aspects of an item may have on subsequent memory for that item, so much that we cannot possibly review that entire literature here. We can nonetheless touch on some important normative characteristics of verbal stimuli that have been found to have robust effects on human memory, and which we report for our stimuli below. First among these is word frequency: In lexical decision tasks, high frequency words are correctly identified more often and faster than low frequency words (Scarborough, Cortese, & Scarborough, 1977); in recognition memory, low-frequency words are correctly recognized and correctly rejected more often than high-frequency words (Glanzer & Adams, 1985); while in free recall, high-frequency words are correctly recalled at a higher rate than low-frequency words in pure lists but not in mixed lists, where low-frequency words are typically recalled more correctly (Gregg, 1976). This discrepancy in the effect of frequency across tasks hints that different tasks may rely on different aspects

of an item, even if they share a certain degree of processes. It may also indicate artifacts of how these frequency categories were chosen, since word frequency has been found to have a non-monotonic relationship to both recognition (Hemmer & Criss, 2013) and recall (Lohnas & Kahana, 2013) performance when it is treated as a continuous, rather than categorical variable.

Complicating this picture, however, is the fact that frequency—like many normative characteristics of words—is highly correlated with many other aspects of a word which have been found to have effects on memory. Low frequency words tend to be more orthographically distinctive than high frequency words; in addition to effects of overall word frequency, lower letter-position frequency yields superior recognition performance (higher hit rates, lower false alarm rates) than words with high letter-position frequency (Malmberg, Steyvers, Stevens, & Shiffrin, 2002). High frequency words also tend to be used in more semantic contexts than low frequency words; high context variability, in addition to frequency, leads to lower recognition accuracy (Steyvers & Malmberg, 2003), and it has been argued that context variability accounts for most effects often attributed to word frequency (Adelman, Brown, & Quesada, 2006). Indeed, pure frequency alone—as measured by number of laboratory exposures—appears to lead to a bias effect (higher hit rate *and* higher false alarm rate) rather than the mirror effect reported for normative word frequency, arguing that "frequency" may be a misnomer with regard to the underlying construct that affects memory for words (Maddox & Estes, 1997).

All of these effects are, however, subject to what is known as the "language-as-fixed-effect" fallacy (H. H. Clark, 1973). This fallacy comes about when a small sample of words is separated into, for example, high and low word-frequency or high and low letter-frequency, and effects due to this classification of the sample are generalized to all of language without respecting the fact that the studied words are not necessarily a random sample from the entire lexicon (what counts as "high" or "low" is always relative to the particular sample). Clark suggested a way to circumvent this fallacy, namely by examining the effects of single items without resorting to grouping them into potentially arbitrary or error-prone categories like high- or low-frequency. Freeman, Heathcote, Chalmers, and Hockley (2010) found that applying this kind of analysis to episodic recognition of words supported the presence of a frequency mirror effect but showed that overall effects of orthographic distinctiveness disappeared when item-specific effects were taken into account. We take a similar approach in our analyses, first estimating effects at the level of single items and then correlating these item effects with their normative lexical characteristics, going beyond their foundational work to examine item effects across multiple tasks and considering a wider array of item properties.

Overview of the present study

Of particular note is that previous studies of the relationships between processes and information across memory tasks have tended to examine only one of these dimensions at a time: they assess either the correlations between individual performance (memory processes) or the correlations between item performance (information in memory), but not both. As we noted above, this makes it difficult to know why two tasks may be related. The present work jointly analyses correlations among individuals and among items in the

same dataset, allowing us to properly attribute correlations to the information conveyed by items or the processes used by individuals in a data-driven manner.

By using the same set of items throughout the experiment, we obtain multiple observations of performance for each item in each task. We therefore estimate performance at the level of individual items, rather than relying on externally-defined normative characteristics (e.g., frequency, concreteness, etc.). Although we will subsequently relate the estimated memory qualities of each item to these normative word properties, estimating these qualities is done entirely in a data-driven, bottom-up manner. A drawback of this approach is that we are left with somewhat greater uncertainty at the level of individual items, because estimation is not constrained by normative word properties. This is balanced against the advantage that such properties are themselves subject to estimation error (e.g., from subjective ratings like concreteness) and reflect a variety of decisions that can affect the construct validity of the measure (e.g., which corpus to use for obtaining frequency counts). By eschewing the explicit use of these measures in our estimation procedure, we avoid these problems and ensure that our conclusions are based more on the data at hand than on the vagaries of these normative measures.

Finally, most prior studies of the relationship between different memory tasks at the level of individuals—such as that by Underwood et al. (1978), discussed above—have used different procedures for both the study and test phases of each task. As a result, it is not possible to attribute correlations between tasks to either encoding or retrieval processes or both. For example, Underwood et al. (1978) studied free recall of abstract and concrete words using two different tasks with different compositions of study lists. Any differences or similarities between these tasks could, therefore, be attributed to study strategies adopted by participants in each task, to memory search strategies at retrieval, or to the information contained in either concrete or abstract words that yields different retrieval ability or associative encoding for each word type. In contrast, although we use different test phases for each task, the study phase is identical and all tasks draw their stimuli from the same pool of items. As a result, we can be confident that any differences between tasks at the level of individuals reflect the operation of different retrieval processes, rather than task-specific study strategies or the particulars of the items used in that task.

Method

Participants

Four hundred and sixty-two undergraduate students at Syracuse University participated in exchange for course credit after providing informed consent in accord with local Institutional Review Board policy (IRB #13-003). Seventy-two participants did not finish the full experiment due to time constraints or computer malfunction but are included in the set of analyzed data since the Bayesian hierarchical techniques we employ do not require a balanced design (although there is greater residual uncertainty regarding the parameters of a participant who contributes less data). We did, however, exclude nine participants who always gave the same response (i.e., always “YES” or “NO”) in one or more of the three binary choice tasks (LD, SR, or AR). The following analyses are, therefore, based on data from 453 individual participants.

Materials

We extracted a set of 924 words from the Touchstone Applied Science Associates (TASA) corpus (Landauer, Foltz, & Laham, 1998). These words were not chosen to be within any particular range on the values we are about to describe, but instead to be a diverse sample of words that would span a wide range of potential measures, thereby maximizing the information we could infer about item variability. For the lexical decision task, pseudoword foils were created using the Wuggy pseudoword generator (Keuleers & Brysbaert, 2010). Each word in our stimulus set served as a "base" that was used to generate three similar pseudoword candidates. An independent rater selected one of the three words to maximize "wordiness". The result was that each word in the set had a corresponding similar pseudoword. For example, the word "ACCIDENT" yielded the pseudoword "ADVIGENT"[1]. As mentioned in the Author Note, the complete set of stimulus words, their pseudoword counterparts, and relevant normative statistics are provided online via the Open Science Framework at `https://osf.io/dd8kp/`.

Word frequency. The frequency with which a word is encountered has typically been measured by counting the number of times the word appears within a large representative sample of text (a "corpus"). The standard frequency measure used in memory research are the counts computed by Kuçera and Francis (1967, "KF"). However, the corpus on which these counts are based is now rather outdated, so we also include the frequency counts measured on the more recent HAL corpus (Burgess & Livesay, 1998).

Orthography. An obvious property of a word's orthography is its length, a simplistic measure of a word's visual complexity. We can go beyond this by considering how unusual a word's spelling is relative to other words in the lexicon: The OLD20 measure (Yarkoni, Balota, & Yap, 2008) reports the average orthographic Levenshtein distance (OLD) between a word and its 20 closest neighbors (in terms of OLD). OLD is the minimum number of letter additions, deletions, and substitutions needed to transform one word into another, and therefore measures how differently two words are spelled. The average of the 20 lowest distances from a word thus reflects how regular (low OLD20) or distinctive (high OLD20) its spelling is. OLD20 is correlated with word length, since longer words have more letters that need to be changed in order to transform them into other words.

Phonology. Just as with the orthography of a word, we can measure the auditory length and relative unusualness of a word's pronunciation using the number of syllables and "Phonological Levenshtein Distance". Just like the OLD20 measure, PLD20 reports the average Levenshtein distance between the syllabic transcription of a word and its twenty closest neighbors. Phonological Levenshtein distance (PLD) is the minimum number of additions, deletions, and substitutions of syllables needed to transform the pronunciation of one word into that of another. A word with high PLD20 is one with an relatively unusual pronunciation (it would take a lot of edits to change it into another word's pronunciation) while a word with low PLD20 has a relatively common pronunciation.

Semantic content. Words also vary greatly in their semantic content, which can be measured in several ways. As with orthography and phonology, we tried to find measures

[1] In some cases, this procedure resulted in a pseudoword that, while not in most dictionaries, could be reasonably be considered a word in common usage (like "SLICKED" or "DISSING"). As a result, our analyses below focus on the case in which a word is presented in its proper word form.

Table 1

Descriptive statistics of words in the stimulus set, with illustrative examples of the extremes of each scale. "Num. missing" reports the number of words, out of the 924 total, for which that measure was unavailable. "KF" = Francis and Kucera corpus frequency; "HAL" = HAL corpus frequency; "OLD20" = Orthographic Levenshtein Distance-20; "PLD20" = Phonological Levenshtein Distance-20; "SND" = Semantic Neighborhood Density. See the main text for details on each measure.

	KF	HAL	Length	OLD20	Num. syllables
Mean	47.93	23046.49	6.72	2.23	2.05
SD	28.54	39447.54	2.16	0.75	0.95
Missing	0	0	0	0	0
Highest	TAX (197)	ARTICLE (861595)	REPRESENTATIVES (15)	NEIGHBORHOOD (5.45)	RESPONSIBILITY (6)
	SECRETARY (191)	FILE (311710)	ADMINISTRATION (14)	OPPORTUNITIES (5.3)	ADMINISTRATION (5)
	SPIRIT (182)	POST (275951)	CHARACTERISTIC (14)	RELATIONSHIPS (5.25)	ASSOCIATED (5)
	COSTS (176)	MAIL (260346)	RESPONSIBILITY (14)	CHARACTERISTIC (5.15)	CHARACTERISTIC (5)
	COMMITTEE (168)	COMPUTER (253358)	TRANSPORTATION (14)	CIRCUMSTANCES (5.15)	CIVILIZATION (5)
Lowest	OCEANS (3)	FERTILE (1394)	CUP (3)	BEAT (1)	BANKS (1)
	MOM (3)	SUPPER (1170)	CAP (3)	BAY (1)	BAND (1)
	BLACKS (3)	DIOXIDE (1022)	BUS (3)	BARE (1)	BAG (1)
	DIOXIDE (2)	COLONISTS (950)	BAY (3)	BAND (1)	AUNT (1)
	COLONISTS (1)	VALLEYS (783)	BAG (3)	BAG (1)	ARMED (1)

	PLD20	Concreteness	Semantic diversity	Num. senses	SND
Mean	2.11	3.32	1.76	7.63	0.76
SD	0.87	1.03	0.26	6.95	0.05
Missing	0	17	5	0	0
Highest	MANUFACTURING (6.35)	APPLE (5)	PREVIOUS (2.28)	BREAKING (60)	TOM (0.95)
	CIRCUMSTANCES (6.25)	BOOTS (5)	FULLY (2.27)	RUNS (57)	VEGETABLES (0.91)
	REPRESENTATIVES (6)	CUP (5)	BRINGING (2.26)	CUTTING (54)	SIGHED (0.91)
	RELATIONSHIPS (5.85)	EAR (5)	MAINLY (2.23)	PLAYS (52)	NODDED (0.89)
	CHARACTERISTIC (5.45)	FINGER (5)	REMAINING (2.22)	HOLDS (45)	PAYMENT (0.89)
Lowest	BEAT (1)	CONCEPT (1.41)	SAIL (1.07)	APARTMENT (1)	CHAIN (0.63)
	BAY (1)	RESPONSIBILITY (1.4)	HYDROGEN (1.03)	ADVICE (1)	PROPERTIES (0.63)
	BARE (1)	POSSIBILITY (1.33)	ATOM (0.94)	ACHIEVE (1)	DEGREES (0.62)
	BAND (1)	LUCK (1.33)	DIOXIDE (0.92)	TOWARDS (0)	EXPRESS (0.61)
	BAG (1)	BELIEF (1.19)	ELECTRONS (0.59)	OUGHT (0)	HOUSEHOLD (0.6)

that reflected both the semantic content inherent to a word (whatever concepts or meanings the word refers to) as well as measures that reflect a word's relative semantic distinctiveness within the lexicon. In terms of a word's inherent content, one measure is its concreteness ("Concr"), the degree to which the word refers to an entity that can be seen or interacted with, reflecting the kinds of experiences one typically has with the word's referent. We use the mean concreteness ratings reported by Brysbaert, Warriner, and Kuperman (2014), which takes a value between 1 (low concreteness) and 5 (high concreteness).

Another measure of a word's inherent semantic content is its "semantic diversity": A word with high semantic diversity is one that has many possible meanings or interpretations, whereas a word with low semantic diversity is one that has only a few possible senses in which it can be used. We quantify this using two measures: First is "NSense", the number of "senses" listed for the word in WordNet (Miller, 1995), analogous to the number of definitions in the dictionary. Second, we also use a more naturalistic measure, "SemD" (Hoffman, Lambon Ralph, & Rogers, 2013), a corpus-based measure that reports the average dissimilarity between documents in which a word appears. It thus represents the diversity of the situations in which a word is used: words with low semantic diversity appear only in

specialized contexts while those with high diversity appear in a wide variety of settings.

Finally, as a measure of a word's semantic content relative to the other words in the lexicon, we computed each word's Semantic Neighborhood Density (SND). We computed SND using a semantic space derived from the HAL model (Hyperspace Analogue to Language; Lund & Burgess, 1996), as described in Günther, Dudschig, and Kaup (2015). A word's SND is defined as the average of its top ten similarity values between that word and all other words in the corpus (similarity between two words is measured by the cosine of the angle between each word's vector representation in the semantic space; Buchanan, Westbury, & Burgess, 2001). Words with high SND tend to have semantically similar neighbors, suggesting that many other words have similar meanings, whereas words with low SND tend to be the only word with their particular meaning. SND is thus analogous to OLD20 and PLD20 in that it measures the relative distinctiveness of a word's semantic (as opposed to orthographic or phonological) content.

Design and Procedure

The experiment included five different tasks: single item recognition, associative recognition, cued recall, free recall and lexical decision. Each task was repeated three times over the course of the experiment for a total of 15 blocks, with 20 test trials per block. The first five blocks consisted of the first presentation of each of the five tasks (randomly ordered for each participant). For the remaining 10 blocks the five task types were presented twice in random order. The task was post-cued, therefore participants could not adopt a study strategy based on the anticipated test type. The items in each block were randomly sampled from the pool of 924 words without replacement for each participant, such that no items repeated between blocks for a given participant.

All blocks (except lexical decision) began with a study phase where participants viewed 20 word pairs presented side by side, one pair at a time. Each pair remained on the screen for 2 seconds and was immediately followed by asking participants to "Please rate the degree of association between the two items you just saw" on a scale from 1–9 where 1 is "not at all associated" and 9 is "highly associated". The word pair was not visible on the screen during the rating. Responses were self-paced by clicking on boxes numbered 1–9 on the screen.

Each study phase was immediately followed by a distractor task. This was a simple math task where participants continuously added a series of 15 random digits drawn with replacement from the range 1–9. Digits were presented at a rate of 3 seconds per digit, for a total presentation time of 45 seconds. After all digits appeared participants typed in their response and received accuracy feedback.

Following the distractor task, participants were presented with one of the following memory tasks. Responses in all tasks were self-paced. Each study/test block was followed by the option to take a self-paced break. The experiment lasted approximately one hour.

Single item recognition. For the target stimuli, ten study items were selected at random from the study list. The ten items could be from either the right or the left presentation position, but not from both the left and the right presentation position for the same study trial. In other words only one of the words in the study word pair could be selected. These ten old items were combined with 10 foils and presented in random order in the center of the screen. Participants are asked to "indicate if the item you see on the screen

was on the list you just studied (YES) or not on the list (NO)". Participants responded by clicking on boxes presented on the computer screen.

Associative recognition. For the test stimuli, ten word pairs were selected at random from the study list. The remaining ten words pairs were scrambled such that none of the pairs remained intact. The scrambled pairs could be rearranged both between earlier and later study positions as well as between right and left presentation positions. The ten intact word pairs were combined with the ten rearranged word pairs and presented in random order in the center of the screen with one word appearing above the other rather than side by side. Participants were asked to "indicate if the PAIR of words you see on the screen was studied as a PAIR on the list you just studied (YES) or were not a pair (NO)". Participants responded by clicking on boxes presented on the computer screen.

Cued recall. For the test stimuli, twenty study items were selected from the study list, one from each pair. Ten of the twenty items were from the right study presentation position and ten were from the left study presentation position, randomly chosen. In this way all twenty study pairs are tested, but half the cue words were from the right presentation position and half were from the left presentation position. The twenty cue words were presented in random order to the left of a box on the computer screen where participants were asked to enter the corresponding word in the pair. Participants are asked to respond by "typing the OTHER WORD in the pair. For example if you studied BRICK BRACK and you now see BRICK your response should be BRACK. If you cannot recall the word click DON'T REMEMBER". It was emphasized to participants that spelling did not matter; rather they should focus on providing as many responses as possible.

Free recall. No words were presented at test, rather participants were asked to "try to recall as many words from the study list as you possibly can. When you cannot recall any more words click on the FINISHED button". Participants were required to attempt to provide responses for a minimum of 90 seconds. A timer appeared on the screen and the finished button could not be clicked until 90 seconds had passed. It was emphasized that spelling did not matter; rather they should focus on providing as many responses as possible.

Lexical decision. This task was not preceded by a study block. For the test stimuli, ten words drawn from the complete word set were combined with 10 pseudo-words and presented in random order in the center of the screen. Participants were simply presented with a word and asked to "indicate if the item you see is a word (YES) or not at word (NO). Respond as QUICKLY as possible". Participants responded "word" by clicking the left mouse button and "non-word" by clicking the right mouse button. Response time was measured from the onset of the word to the click of the mouse button.

Free Response Normalization

Responses for cued and free recall tasks were first corrected for spelling and automatically coded for accuracy via an automated direct word comparison. Any incorrect responses were then corrected for plurals and conjugations and hand coded following a lenient criterion such that if the word could be interpreted as being the same as a target word it was coded as correct. For example, a response of "fiexd" was accepted as the target "fixed" (switched letters), "contricute" was accepted for "contribute" (wrong letter), "aranged" was accepted for "arranged" (missing letter).

Analysis

Our interest is in how performance in each of these tasks is correlated, both among items and among individuals. To study this, we make use of a joint measurement model that characterizes the outcomes of each trial in the experiment in terms of a set of parameters related to the individual on that trial as well as a set of parameters related to the item(s) on that trial[2]. The values of these parameters are merely quantitative measurements of the contributions of items and individuals to performance in each task. The correlations among these two sets of parameters thus reflect how the contributions of items and individuals vary between and within tasks. We wish to emphasize that our modeling is aimed at extracting these quantities (hence the term "measurement model"), not at providing a description of the psychological processes involved in each task (which we would term a "psychological" or "cognitive" model).

We estimate these quantities within a hierarchical Bayesian model. Bayesian methods have two crucial advantages for our purpose (Vandekerckhove, 2014): First, the resulting estimates reflect the different sources and degrees of uncertainty between items and individuals (and between different items and different individuals). Second, the resulting estimates reflect the fact that items and individuals *jointly* contribute to responses across trials, such that information about the items informs the estimates of the individuals and vice versa.

Task models

Our tasks comprise two broad categories: binary-choice tasks (lexical decision and single-item and associative recognition) and recall tasks (cued and free recall). Because of the differences between these types of tasks, different measurement models are required for each. Below, we describe the structure and parameters of these models and how they are related to observed task performance with both items and individuals. We emphasize that the models described below are used as *measurement* tools that act to transform the data into parameters which are more readily interpreted.

Binary choice tasks. For binary choice tasks, performance on each trial is characterized by a final response ("yes" or "no") and response time (RT). A common and robust framework for jointly modeling choice and response time are random-walk/diffusion models (e.g., Busemeyer & Townsend, 1993; Edwards, 1965; Link, 1975; Link & Heath, 1975; Ratcliff, 1978; Stone, 1960). Such models assume that participants continue to draw samples from an evidence-generating process (in this case, the process is memory retrieval) until they have accumulated enough to commit to a decision. Such models are widely used throughout psychology, including in our three binary choice tasks, lexical decision (Ratcliff, Gomez, & McKoon, 2004), single-item recognition (Ratcliff, 1978; Starns, Ratcliff, & McKoon, 2012; Starns & Ratcliff, 2014), and associative recognition (Ratcliff, Thapar, & McKoon, 2011; Voskuilen & Ratcliff, 2016)[3]. Although other frameworks exist to jointly characterize accuracy and response time, notably accumulator models (Vickers, 1970; Usher & McClelland, 2001; Brown & Heathcote, 2008), in practice when used as measurement models, these

[2]This model structure is identical to that used in item response theory.

[3]It is likely that, in associative recognition, multiple accumulation processes are actually involved in making a decision (Cox & Criss, 2017; Cox & Shiffrin, 2017) although a single accumulation process has been found to provide a satisfactory account of the data produced in this task.

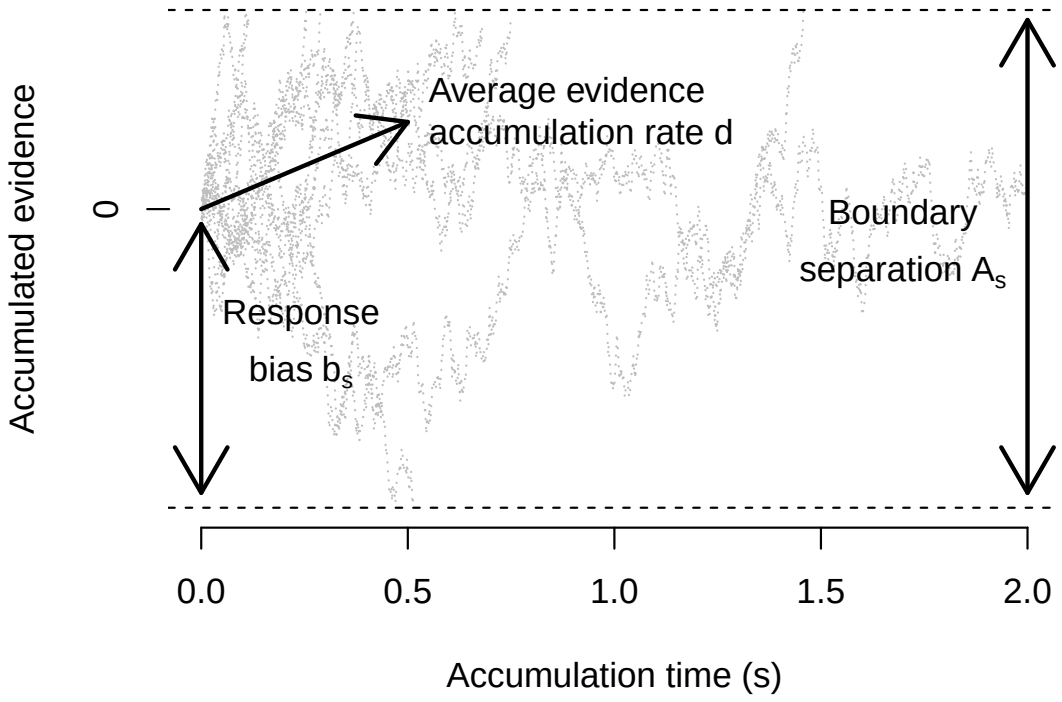


Figure 1. Depiction of a diffusion model of evidence accumulation, illustrating the roles of the boundary separation parameter A_s, response bias parameter b_s, and average rate of evidence accumulation d. Gray dotted lines indicate potential evidence trajectories that eventually lead to one of the two response boundaries (dashed lines), at which time participants commit to a "yes" (upper boundary) or "no" (lower boundary) decision. Total response time is the sum of the time spent accumulating evidence until a boundary is hit plus a residual time R_s for initiating the accumulation process and executing the response.

frameworks lead to identical conclusions regarding cognitive constructs of interest, particularly with regard to the strength of the evidence used to make a decision, which is our primary interest in this analysis (Donkin, Brown, Heathcote, & Wagenmakers, 2011; Rae, Heathcote, Donkin, Averell, & Brown, 2014).[4]

A diffusion model for binary choice, as depicted in Figure 1, assumes that each sample of evidence comes from a Gaussian distribution with mean d and unit variance (the variance amounts to a scale factor that can be fixed for present purposes). Over time, the summed evidence will "drift" with direction and magnitude proportional to d. To the extent that

[4]For analysis purposes, we use only a simple Wiener diffusion model rather than the more elaborate diffusion model popularized by Ratcliff and colleagues. Our model, therefore, does not include trial-by-trial variability in either starting point or residual time, although it does incorporate trial-by-trial variability in drift rates by virtue of including item effects (which, by definition, vary between trials). Including such additional variability may be more psychologically plausible, but is not necessary when using the model as a measurement tool.

$d > 0$, evidence drifts upward and favors a "yes" response; conversely, to the extent that $d < 0$ evidence will drift downward and favor a "no" response. Participants set upper and lower response boundaries reflecting the amount of accumulated evidence needed to commit to a "yes" or "no" response, respectively. These boundaries can be characterized by their degree of separation A_s and bias b_s, such that the upper boundary is $(1 - b_s) \times A_s$ and the lower boundary is $-b_s \times A_s$. As boundary separation A_s increases, more evidence is required to commit to either decision, resulting in longer response times but greater accuracy, owing to the additional evidence accumulated. When $b_s = \frac{1}{2}$, the boundaries are symmetrical meaning that "yes" and "no" responses require an equal amount of accumulated evidence. If $b_s > \frac{1}{2}$, the upper boundary is closer to the start than the lower boundary, meaning less evidence is needed to commit to a "yes" versus a "no" response and we say that there is a "response bias" toward "yes" (and conversely if $b_s < \frac{1}{2}$). The response for a particular trial is given by which of the two boundaries is hit first and the RT is the time taken to hit that boundary, plus a residual time R_s reflecting time needed to detect and begin processing the test item(s) and execute the motor response. The likelihood of making either a "yes" or a "no" response at time t can then be expressed (Feller, 1968; Navarro & Fuss, 2009; Ratcliff, 1978):

$$\begin{aligned}
\Pr(\text{“yes”}, t | A_s, b_s, R_s, d) = \frac{\pi}{A_s^2} \exp\left(dA_s(1-b_s) - \frac{d^2(t-R_s)}{2}\right) \\
\times \sum_{k=1}^{\infty} k \exp\left(-\frac{k^2\pi^2(t-R_s)}{2A_s^2}\right) \sin(k\pi(1-b_s)) \\
\Pr(\text{“no”}, t | A_s, b_s, R_s, d) = \frac{\pi}{A_s^2} \exp\left(-dA_s b_s - \frac{d^2(t-R_s)}{2}\right) \\
\times \sum_{k=1}^{\infty} k \exp\left(-\frac{k^2\pi^2(t-R_s)}{2A_s^2}\right) \sin(k\pi b_s)
\end{aligned}$$

Although these expressions seem formidable, they can be computed to high precision with relative ease using modern numerical methods (Navarro & Fuss, 2009; Tuerlincx, 2004; Wabersich & Vandekerckhove, 2014).

We decompose the evidence accumulation rate d into a sum of two components: β, which reflects the tendency for evidence to drift upwards regardless of the kind of test trial; and δ, the difference in the rate of evidence accumulation between target trials (those for which a positive response is correct) and foil trials (those for which a negative response is correct), such that δ is conceptually similar to the d' measure from signal detection. Thus, for each participant s, five parameters describe their decision process for each binary choice task: Boundary separation A_s, boundary bias b_s, residual time R_s, average evidence drift β_s, and evidence accuracy δ_s. Each item i is associated with two parameters for each binary choice task: an average evidence drift β_i and accuracy δ_i. As we now describe in detail for each task, the participant and the item(s) jointly influence the mean rate of evidence accumulation d for each trial, while response boundaries and residual time are properties of the participants only.

Single-item recognition. We begin by describing the task model for single-item recognition (SR), which will also help to explicate how choice and RT are jointly modeled

across all binary choice tasks. On any one trial, participant s is presented with probe item i. Probe item i can either be a word that was studied (a "target") or one that was not (a "foil"). We use the indicator variable $\mathbb{T}^{SR}$ to represent this: $\mathbb{T}^{SR} = \frac{1}{2}$ if the probe is a target and $\mathbb{T}^{SR} = -\frac{1}{2}$ if it is a foil (the choice of $\frac{1}{2}$ as the magnitude of the indicator is arbitrary, but keeps the drift and accuracy parameters on the same scale).

As mentioned above, participant s is associated with a level of boundary separation A_s^{SR}, boundary bias b_s^{SR}, and residual time R_s^{SR} for the SR task, and these are not affected by the item presented on the current trial. The mean rate of evidence accumulation d is a function of *both* the participant s and the test item i:

$$d = \overbrace{\beta_i^{SR} + \beta_s^{SR}}^{\text{drift}} + \mathbb{T}^{SR} \underbrace{\left(\delta_i^{SR} + \delta_s^{SR}\right)}_{\text{accuracy}},$$

where β_i^{SR} is the drift associated with item i, reflecting the recognition evidence provided by item i regardless of its study status; β_s^{SR} is the drift associated with participant s, reflecting the tendency for participant s to accumulate positive or negative evidence regardless of the test item (note that this is separate from boundary bias b_s^{SR}); δ_i^{SR} is the degree to which item i yields different recognition evidence when it is a target versus a foil; and δ_s^{SR} is the overall ability of participant s to distinguish between targets and foils in SR.

Lexical decision. The task model for lexical decision (LD) is a close analog to that for SR. The critical difference is that, whereas $\mathbb{T}^{SR}$ indicated whether the probe had been studied or not, the corresponding indicator variable for LD, $\mathbb{T}^{LD}$, indicates whether the probe is presented in its usual word form ($\mathbb{T}^{LD} = \frac{1}{2}$) or is instead presented in its distorted pseudoword form ($\mathbb{T}^{LD} = -\frac{1}{2}$). The drift parameters now reflect, for an item, the degree to which it tends to seem like a word (β_i^{LD}), and for a participant, the degree to which they accumulate positive lexical evidence regardless of the probe item (β_s^{LD}). The accuracy parameters describe, for an item, how easily its word and pseudoword forms can be told apart (δ_i^{LD}), and, for a participant, how well they can generally discriminate between words and pseudowords (δ_s^{LD}). The resulting average rate of evidence accumulation on a particular trial is, again, a joint function of the item on that trial (i) and the participant engaging in that trial (s):

$$d = \overbrace{\beta_i^{LD} + \beta_s^{LD}}^{\text{drift}} + \mathbb{T}^{LD} \underbrace{\left(\delta_i^{LD} + \delta_s^{LD}\right)}_{\text{accuracy}}.$$

Finally, recall that each participant is associated with a boundary separation (A_s^{LD}), boundary bias (b_s^{LD}), and residual time (R_s^{LD}) for the LD task.

Associative recognition. In associative recognition (AR), there are two probe items (i and j) instead of just one. In addition, the indicator $\mathbb{T}^{AR}$ now refers to whether the pair is intact ($\mathbb{T}^{AR} = \frac{1}{2}$) or rearranged ($\mathbb{T}^{AR} = -\frac{1}{2}$). In our model, each item contributes independently to drift—the degree to which they appear to be an intact pair—and to accuracy—the ability to tell whether the pair (i, j) is intact or rearranged. As in LD and SR, then, we can express the rate of evidence accumulation on a trial as a function of the

items presented (i and j) and the participant (s):

$$d = \overbrace{\beta_i^{AR} + \beta_j^{AR} + \beta_s^{AR}}^{\text{drift}} + \mathbb{T}^{AR} \underbrace{\left(\delta_i^{AR} + \delta_j^{AR} + \delta_s^{AR}\right)}_{\text{accuracy}}.$$

And, again, each participant has a boundary separation (A_s^{AR}), boundary bias (b_s^{AR}), and residual time (R_s^{AR}) for the AR task.

Recall tasks. The outcome on any one trial of a recall task can be classified as one of the following: a correct recall, an erroneous recall (an intrusion), or a failure to make any response. Unlike in binary choice tasks, we do not attempt to model response time in recall. This is chiefly because it is unclear how RT should be measured in recall, given that participants were allowed unlimited time to retype their responses before committing to a final response (in the few studies that have focused on RT in recall tasks, a different procedure is often used in which participants first produce a clear signal that they are ready to respond, followed by a more protracted period in which they produce the response; see, e.g., Nobel & Shiffrin, 2001).

Our measurement model for recall tasks breaks down each recall observation into two components: a "bias", ρ, to make a response rather than give up; and the accuracy, α of a response conditional on having made one (i.e., a correct recall versus an intrusion). These components, which are influenced by both items and participants, enter into a multinomial logistic link function[5] to yield a likelihood for each possible outcome:

$$\begin{aligned}
\Pr(\text{Correct recall}) &= \frac{\exp(\rho+\alpha)}{\exp(\rho+\alpha)+\exp(\rho-\alpha)+\exp(-\rho)} \\
\Pr(\text{Intrusion}) &= \frac{\exp(\rho-\alpha)}{\exp(\rho+\alpha)+\exp(\rho-\alpha)+\exp(-\rho)} \\
\Pr(\text{No response}) &= \frac{\exp(-\rho)}{\exp(\rho+\alpha)+\exp(\rho-\alpha)+\exp(-\rho)}.
\end{aligned}$$

We now describe how ρ and α are computed for each trial of each recall task.

Cued recall. A single trial of cued recall (CR) involves a cue item (i) and a target item (j). Because each of these items plays a distinct role in cued recall, we estimate separate parameters for each. CR bias ρ and accuracy α are thus functions of the cue and target items (cue i and tar[get] j) and the participant s:

$$\begin{aligned}
\rho &= \beta_i^{\text{CR Cue}} + \beta_j^{\text{CR Tar}} + \beta_s^{CR} \\
\alpha &= \delta_i^{\text{CR Cue}} + \delta_j^{\text{CR Tar}} + \delta_s^{CR},
\end{aligned}$$

where β_s^{CR} reflects the overall tendency for participant s to produce responses in CR; each item β reflects the degree to which that item tends to elicit a response in CR; each item δ reflects the degree to which items in each role yield accurate CR responses, rather than intrusions; and δ_s^{CR} is the tendency for participant s to produce correct CR responses rather than intrusions.

[5] Otherwise known as a softmax or Luce choice rule.

Free recall. The task model for free recall (FR) is essentially a simpler version of that for cued recall, with bias ρ and accuracy α being functions of the participant s and the (potential) response item i:

$$\rho = \beta_i^{FR} + \beta_s^{FR}$$
$$\alpha = \delta_i^{FR} + \delta_s^{FR}.$$

Note that, just as we allow different accuracy parameters δ between cued and free recall, we also allow for different β parameters to reflect overall response tendencies in cued and free recall. Unlike in all the other tasks, FR has no explicit trial structure: participants proceed at their own pace and stop at their leisure. Therefore, we create a set of "ghost" trials in FR: For each list, there are 40 studied items. Each response a participant makes, whether correct or an intrusion, is considered a "trial", with the item i being the word they produced on that trial. For each list item a participant fails to recall, we create a "trial" in which we imagine that the participant attempted to recall that item but failed. This way, even though we directly observe only actual FR responses, our model can still make use of the information about the items that were studied but not recalled. Finally, if the participant gives a response that is not among the 924 items in the stimulus pool, we set $\beta_i = \delta_i = 0$ on that trial.

Bayesian estimation

The task models described above assign a likelihood to the observed outcome (choice and RT for binary choice tasks and response in recall tasks) of each trial for each task. These likelihoods depend on the 19 parameters for each participant, summarized in Table 2, and the 12 parameters for each item, summarized in Table 3. We estimate all of these parameters simultaneously using Bayesian techniques (for an overview, see Kruschke, 2015) implemented in Stan (Carpenter et al., 2017). The model structure is depicted in Figure 2, where we used intentionally broad priors given that the large amount of data would outweigh even moderately informed prior distributions. For estimation purposes, constrained parameters (boundary separation, response bias, and residual time parameters) were first estimated on the real line and then transformed to their constrained scales before using them in the likelihood computations. For boundary separation parameters, an exponential transformation was used to go from the real line to strictly positive values; for response bias parameters, a logistic transformation ($1/\left[1 + \exp\left(-x\right)\right]$) was used to go from the real line to values between 0 and 1; for residual times, which are constrained to be between 0 and the minimum RT for a participant in each task, a scaled logistic transformation was used ($\min \mathrm{RT}_{t,s}/\left[1 + \exp\left(-x\right)\right]$, where $\min \mathrm{RT}_{t,s}$ is the minimum observed RT for participant s in task t).

We obtained 10,000 samples from the joint posterior distribution by using Stan (Carpenter et al., 2017) to run 10 parallel Monte Carlo Markov chains for 1000 iterations each, following 1000 adaptation steps each. The $\hat{R}$ statistic (Gelman & Rubin, 1992) measures the degree to which these chains have converged to a stationary representation of the posterior distribution; the closer $\hat{R}$ is to 1, the stronger the degree of convergence. Across all 453×19 participant parameters, the median $\hat{R}$ is 1.000 and the 97.5% quantile is 1.005; across all 924×12 item parameters, the median $\hat{R}$ is 1.000 and the 97.5% quantile

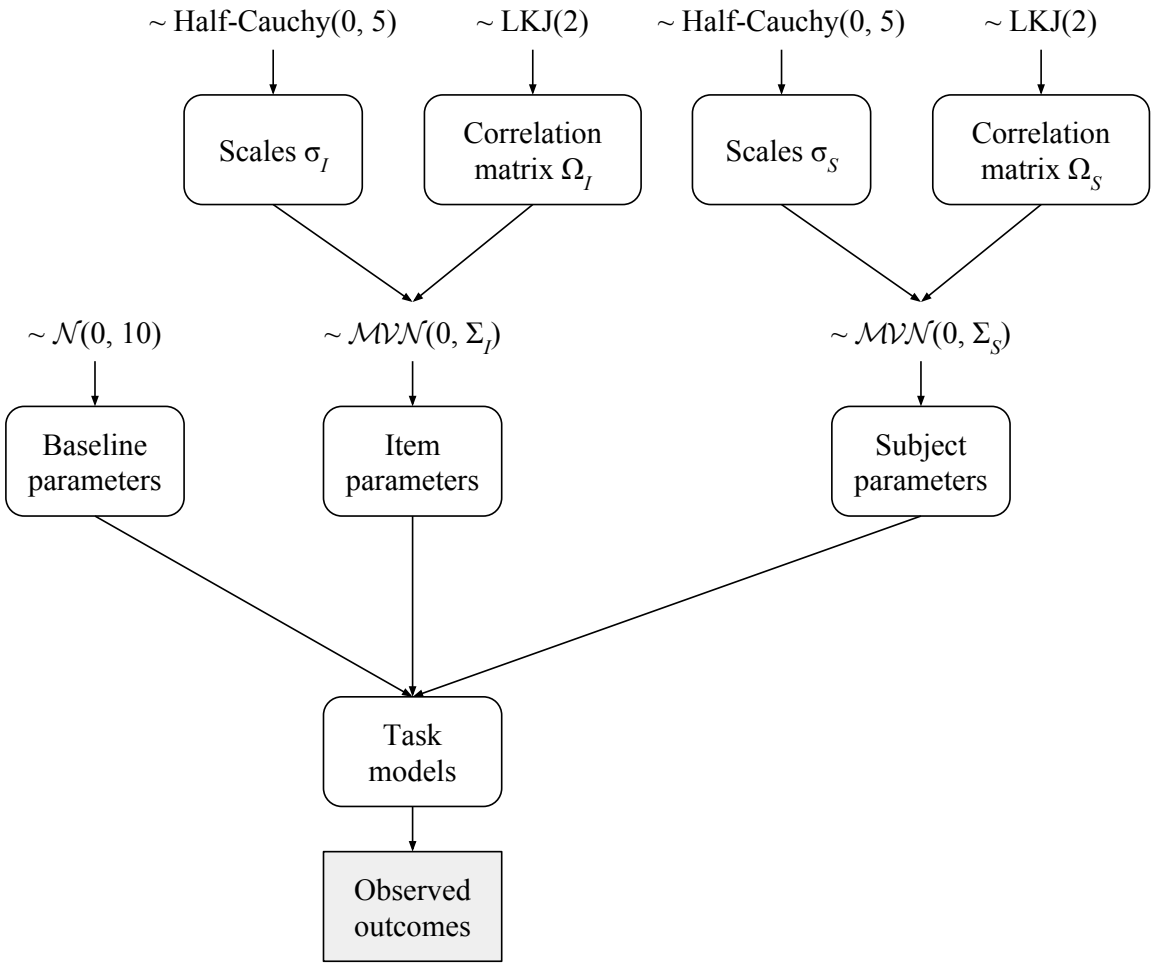


Figure 2. Schematic depiction of the model structure used to estimate parameters for items and subjects across tasks. Individual participant parameters were each constrained to have a mean (across participants) of zero, as were each item parameter, such that the overall means are separately estimated as the "baseline" parameters. The half-Cauchy prior distribution over scales is parameterized in terms of center and scale parameters (here set to a weakly informative 0 and 5, respectively) and is restricted to be strictly positive. The "LKJ" prior distribution over correlation matrices is defined by Lewandowski, Kurowicka, and Joe (2009). Covariance matrices Σ are obtained from the vector of scales σ and correlation matrix Ω via $\Sigma = \text{Diag}(\sigma)\Omega\,\text{Diag}(\sigma)$ where $\text{Diag}(\sigma)$ indicates a diagonal matrix with the vector of scales along its main diagonal. Task models are described in the main text and are used to compute the likelihood of the observed outcome on each trial, conditioned on the baseline, item, and individual participant parameters.

Table 2
Summary of parameters describing individual participants.

Parameter	Description
A_s^{LD}	Boundary separation in lexical decision.
R_s^{LD}	Lexical decision residual time.
b_s^{LD}	Response bias in lexical decision.
β_s^{LD}	Participant's average evidence drift rate in lexical decision.
δ_s^{LD}	Participant's ability to distinguish between words and non-words.
A_s^{SR}	Boundary separation in single-item recognition.
R_s^{SR}	Single-item recognition residual time.
b_s^{SR}	Response bias in single-item recognition.
β_s^{SR}	Participant's average evidence drift rate in single-item recognition.
δ_s^{SR}	Participant's ability to distinguish between studied and unstudied items.
A_s^{AR}	Boundary separation in associative recognition.
R_s^{AR}	Associative recognition residual time.
b_s^{AR}	Response bias in associative recognition.
β_s^{AR}	Participant's average evidence drift rate in associative recognition.
δ_s^{AR}	Participant's ability to distinguish between intact and rearranged pairs.
β_s^{CR}	Average tendency for a participant to give a response in cued recall.
δ_s^{CR}	Rate at which a participant gives correct rather than incorrect responses in cued recall.
β_s^{FR}	Average tendency for a participant to give a response in free recall.
δ_s^{FR}	Rate at which a participant gives correct rather than incorrect responses in free recall.

is 1.001. There is thus strong evidence for convergence. As a measure of how accurately each of these parameters is estimated, we can compute a signal-to-noise ratio, namely, the absolute value of the posterior mean divided by the Monte Carlo standard error (Kass, Carlin, Gelman, & Neal, 1998); greater values indicate greater accuracy, and generally values above 3 indicate a strong signal. Across all 453×19 participant parameters, the median signal-to-noise ratio is 160 and the 2.5% quantile is 6.0; across all 924×12 item parameters, the median signal-to-noise ratio is 93 and the 2.5% quantile is 4.6. There is, then, strong evidence that the joint posterior distribution across all item and participant parameters is accurately represented.

Inferring correlational structure

Each sample from the posterior distribution yields 12 sampled parameters for each item and 19 sampled parameters for each individual. We then compute the Pearson correlation matrices for each of these two sets of parameters for each sample[6]. The resulting

[6]The posterior distribution over the hyper-parameters of the item and individual parameters could also be used for this purpose and, in practice, gives essentially identical results, but is somewhat less stable since it is not as closely tied to the data.

Table 3
Summary of parameters describing items.

Parameter	Description
β_i^{LD}	Item's average evidence drift rate in lexical decision.
δ_i^{LD}	Difference in lexical evidence provided by an item when it is a word versus when it is distorted into a pseudoword.
$\beta_i^{LD} + \delta_i^{LD}/2$	Evidence drift rate when an item is presented in its normal non-distorted form in lexical decision (note that this is not a free parameter, but a function of the previous two parameters).
β_i^{SR}	Item's average evidence drift rate in single-item recognition.
δ_i^{SR}	Difference in memory evidence provided by an item when it was studied versus when it was not.
β_i^{AR}	Item's contribution to the average evidence drift rate in associative recognition.
δ_i^{AR}	Difference in memory evidence provided by an item when it is part of an intact pair versus part of a rearranged pair.
$\beta_i^{\text{CR Cue}}$	Tendency for an item to elicit a response in cued recall when it is given as a cue.
$\beta_j^{\text{CR Tar}}$	Tendency for an item to elicit a response in cued recall when it is the target.
$\delta_i^{\text{CR Cue}}$	Tendency for an item to elicit a correct versus an incorrect response in cued recall when it is given as a cue.
$\delta_j^{\text{CR Tar}}$	Tendency for an item to elicit a correct versus an incorrect response in cued recall when it is the target.
β_i^{FR}	Overall tendency for an item to be produced as a response in free recall.
δ_i^{FR}	Overall tendency for an item to be produced correctly versus incorrectly in free recall.

posterior distribution over correlation matrices for item parameters and individual parameters contains all the information needed to determine how parameters are related among items and individuals, respectively. Unfortunately, it is difficult to directly interpret even a single 12×12 or 19×19 correlation matrix if there are complex relationships among the entities involved, let alone 10,000 such matrices. We therefore augment analysis of the correlation matrices by obtaining the principal components—that is, the eigenvalues and eigenvectors—of each sample of each correlation matrix and examine the resulting distribution of parameter loadings on each component. The principal components represent the latent dimensions along which items and individuals may vary (for an overview of principal components analysis, see Jolliffe, 2002). This representation does not lose any information—it is merely a decomposition of the original correlation matrix[7]. To the extent that the cor-

[7]Although an additive factor analysis model could also be used, doing so would require choosing the number of dimensions/factors ahead of time, whereas PCA allows this decision to be driven purely by the data.

relation matrix contains meaningful structure, these transformed dimensions—the principal components—convey information about the extent to which different groups of parameters covary along the same dimensions. Moreover, because these dimensions are, by definition, orthogonal to one another, we gain insight about how different groups of parameters can vary *independently* of one another, thus reflecting different sources of variability.

Results

Prior to analysis, we excluded trials in the binary choice tasks (lexical decision, single-item recognition, and associative recognition) in which responding was either exceptionally short or exceptionally long, since these were unlikely to reflect the processes of interest. Specifically, responses shorter than 200 ms were excluded because they could not have been produced by processing the stimuli on a given trial (this resulted in exclusion of 445 trials in lexical decision, 107 trials in single-item recognition, and 108 trials in associative recognition). Responses greater than 10 seconds were also excluded because they are likely to be contaminated by processes other than those involved in the task (this resulted in exclusion of 34 trials in lexical decision, 26 trials in single-item recognition, and 60 trials in associative recognition). The following analyses are thus based on a total of 164,001 trials (26,093 in lexical decision, 26,253 in single-item recognition, 26,171 in associative recognition, 26,495 in cued recall, and 9884 in free recall) produced by 453 participants.

As a very coarse summary of the data, we present average response probabilities in each task in Figure 3 and average median correct RT for binary-choice tasks in Figure 4. We note for the moment the apparently high intrusion rate in FR. As shown in Figure 5, although overall rate of responding in FR diminishes over the experiment, the relative proportions of different types of intrusions and correct responses stays relatively consistent after the first block. Comparing the responses in FR when it is the first block versus any subsequent block, it is clear that although the relative proportion of extra-list intrusions remains constant across blocks, intrusions of words from prior lists effectively "consume" responses that, in the first block, would have been correct; it thus appears that many intrusions in FR can be attributed to confusions between the current and previous lists. Moreover, such prior-list intrusions come more often from recent lists rather than more temporally distant lists (Zaromb et al., 2006). Intrusions in FR can be contrasted with CR, where although overall responding once again tends to decrease over the course of the experiment, prior-list and even within-list intrusions are comparatively rare (Figure 6). These differing intrusion patterns will turn out to be important in the Discussion. We now turn to the modeling results for greater insight into the structure of the data.

Posterior predictive

To ensure that our model provides an accurate description of the data across tasks, we simulated a complete set of responses for each sample from the posterior and used these to produce a distribution of predicted performance across individual participants (Figure 7) and, separately, over individual items (Figure 8). The first thing to note is that the predictions are centered at the observed values (i.e., about the diagonal of each plot), justifying the model as providing a legitimate description of the data. The second thing to note is the spread of the predictive distribution around each diagonal, which reflects the

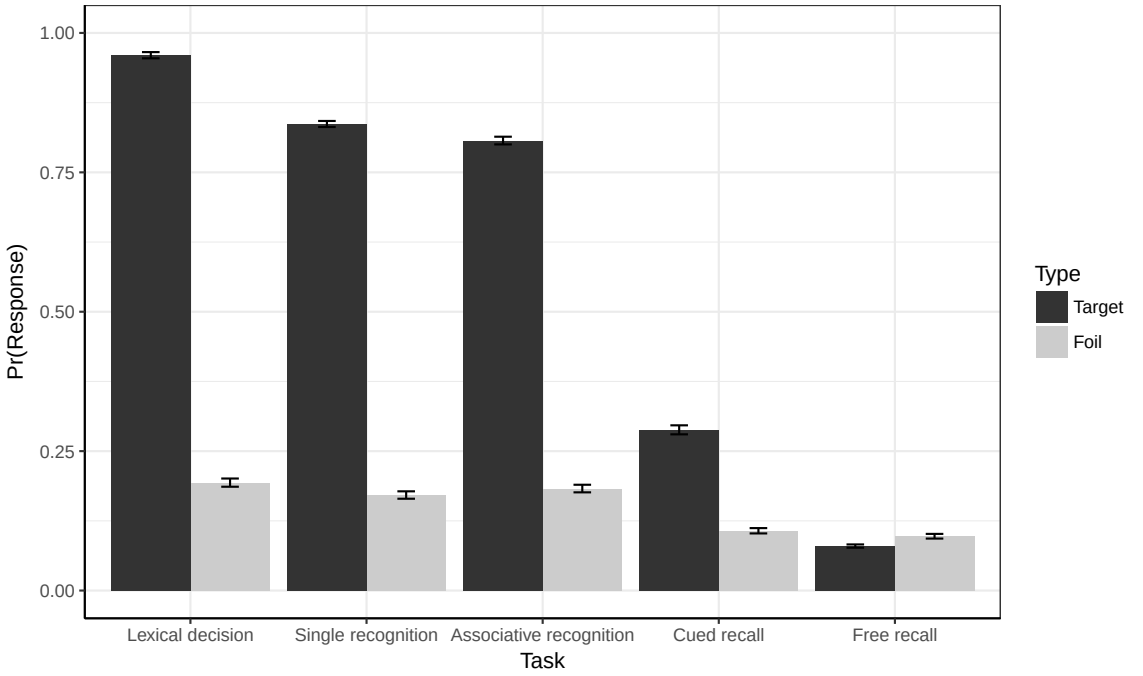


Figure 3. Observed mean response probability in each task. For binary choice tasks (lexical decision, single recognition, and associative recognition), response probability is the probability of giving a positive response ("YES") to the given test. "Targets" and "foils" are defined as words and pseudowords in lexical decision; studied and unstudied words in single recognition; and intact and rearranged pairs in associative recognition, respectively. For cued recall, response probability is the probability of producing a response that is either the target item (correct recall) or a nontarget/foil item (intrusion). For free recall, response probability is the proportion out of 40 possible responses that are correct studied (target) items or nontarget/foil items (intrusions). Error bars denote within-subjects standard errors around the mean.

level of uncertainty inherent in the data. This spread is generally smaller for individuals than items, suggesting that individuals, marginalized over items, yield more consistent performance than items, marginalized over individuals.

Comparison of binary-choice task parameters

There are clear differences among the three binary choice tasks (LD, SR, and AR) in terms of both accuracy and RT. By examining the posterior distribution of mean parameters, we can figure out how to attribute these differences to the cognitive processes these parameters represent. As shown in Figure 9, participants are, indeed, more accurate in LD than in SR and AR in the sense that, on average, they accumulate stronger evidence discriminating between words and nonwords than they do discriminating studied from unstudied items or pairs. Both evidence drift and response bias are higher in LD as well, consistent with the high rate at which words are endorsed as well as the high false alarm rate to nonwords. Finally, the increased RT in AR relative to the other tasks can be

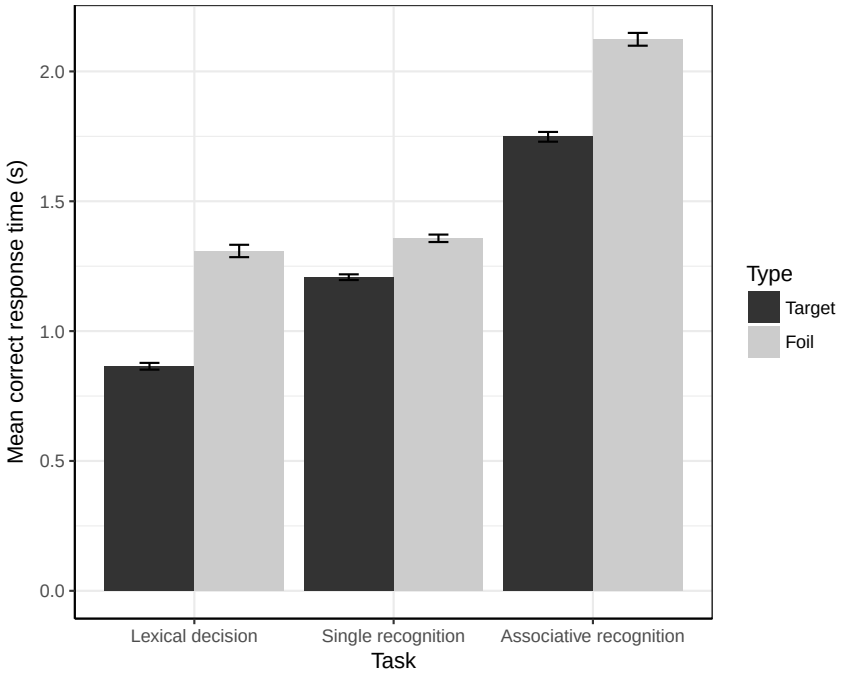
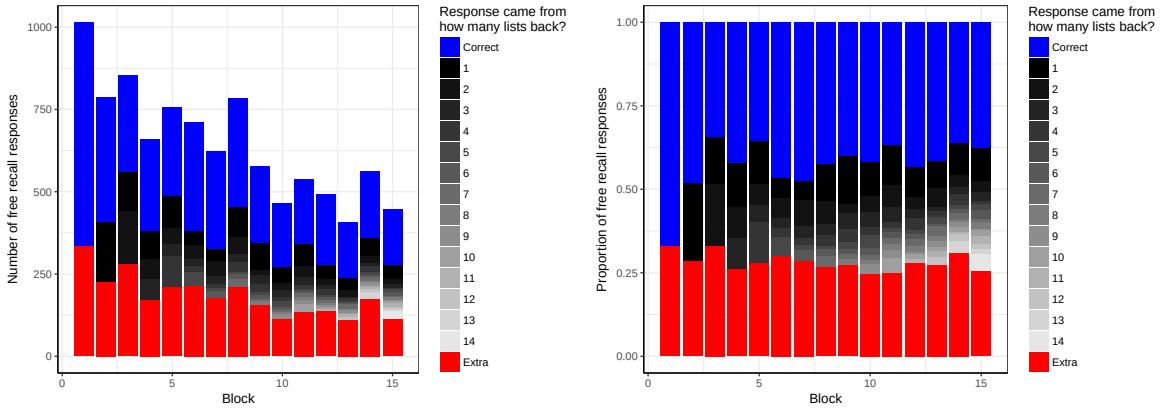
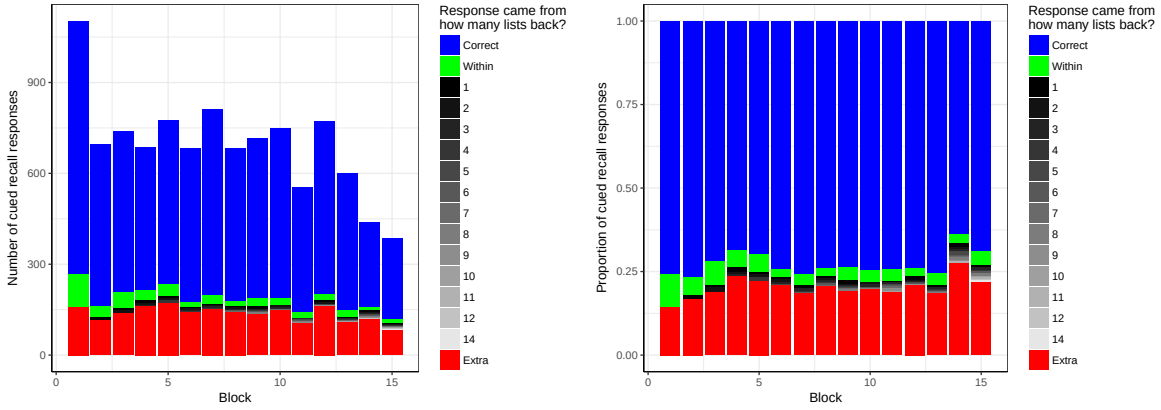


Figure 4. Observed mean correct response time in binary choice tasks. "Targets" and "foils" are defined as words and pseudowords in lexical decision; studied and unstudied words in single recognition; and intact and rearranged pairs in associative recognition, respectively. Error bars denote within-subjects standard errors around the mean.



(a) Total number of responses. (b) Relative proportion of responses.

Figure 5. Pattern of responses in free recall as a function of block within the experiment. "Extra" indicates an extra-list intrusion.



(a) Total number of responses. (b) Relative proportion of responses.

Figure 6. Pattern of responses in cued recall as a function of block within the experiment. "Extra" indicates an extra-list intrusion while "within" indicates a within-list intrusion.

attributed both to an increase in the amount of evidence needed to commit to a decision (boundary separation) as well as an increase in residual time. A simple reason why residual time may be higher in AR relative to SR is that participants must process two words in AR rather than just one, although response-signal studies suggest that the cost of reading two vs. one word is relatively minor compared to the cost of requiring associative information rather than only item information (Gronlund & Ratcliff, 1989). Other than the difference in the number of stimuli in AR, response demands (pressing a mouse button) are equivalent across the binary-choice tasks, suggesting that many of the differences in residual time can be attributed to additional processes engaged while accessing the relevant evidence in memory (see Cox & Shiffrin, 2017, for additional discussion on differences in residual times and additional processing in AR relative to SR).

Correlations at the individual level

The posterior distribution over the matrix of correlations between all individual participant parameters is shown in Figure 10. A visual inspection reveals mostly positive correlations. Accuracy-related parameters (δ's) in all tasks are positively correlated, with accuracies in single-item and associative recognition representing the strongest pairwise correlation, with overall recall levels (β_s^{CR} and β_s^{FR}) also being correlated between tasks and with accuracy in all episodic tasks. Accuracy parameters in the four episodic tasks are negatively correlated with overall evidence drift in single-item recognition (β_s^{SR}), indicating that a tendency to accrue positive evidence in SR (irrespective of the test item) is negatively associated with episodic accuracy overall. There are also positive correlations among the boundary separation (A's), response bias (b's), and residual time (R's) parameters across binary-choice tasks, suggesting that these represent relatively stable characteristics of individuals. In general, residual time and boundary separation are positively correlated with accuracy, perhaps indicating that participants who are more deliberate and/or focus more time on task tend to be more accurate. However, this is only true for SR and AR; in LD, accuracy and boundary separation are negatively correlated, reflecting a kind of

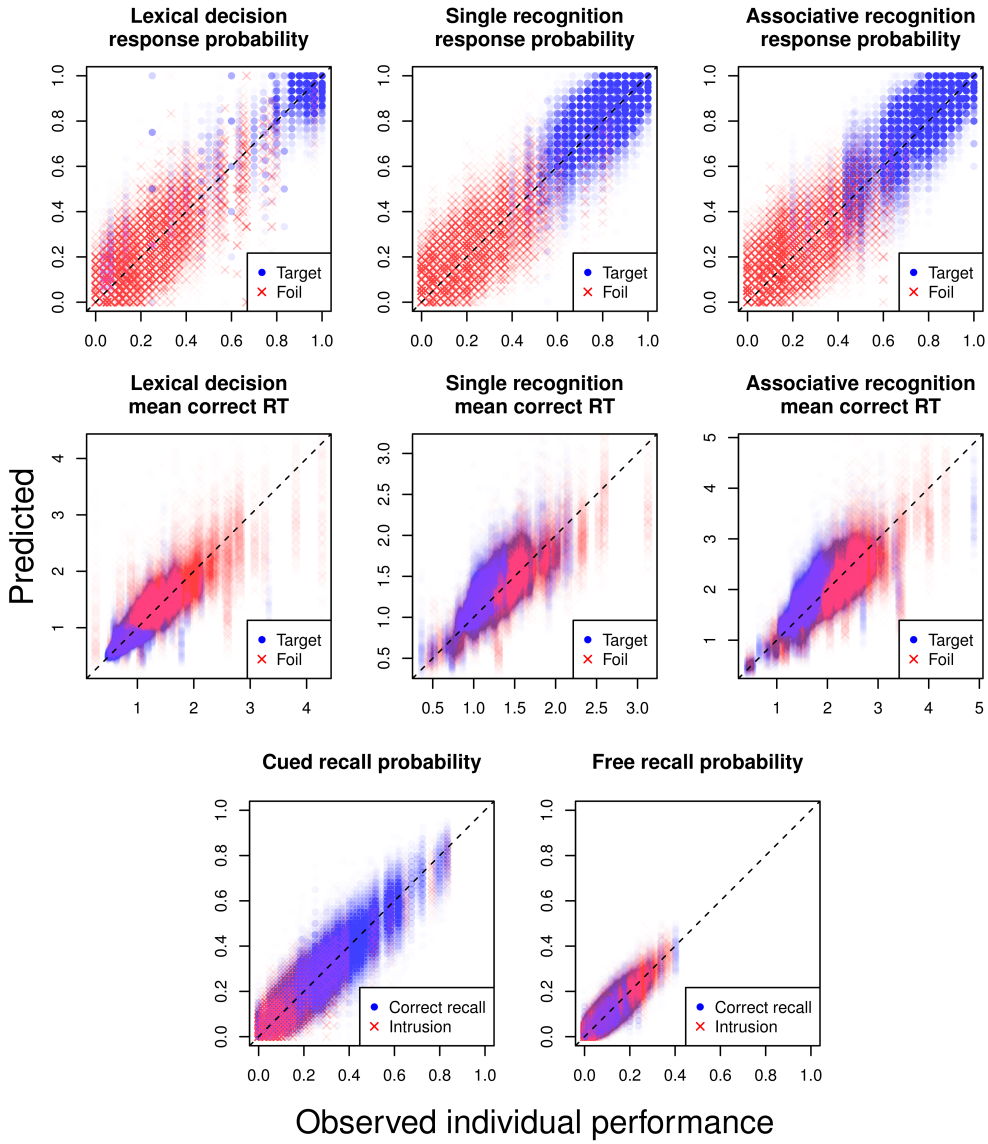


Figure 7. Posterior predictive distributions for performance of individual participants. Each point represents a sample of single participant.

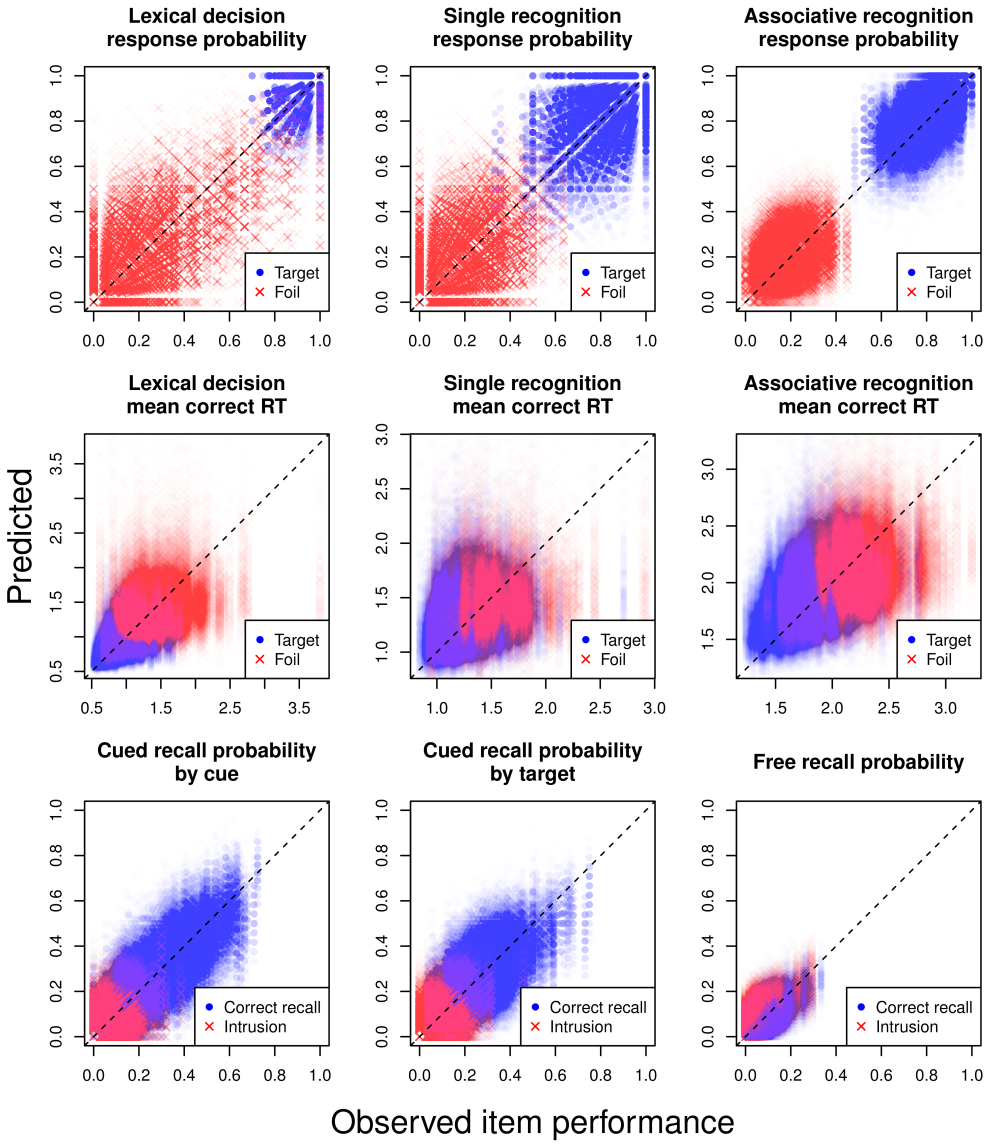


Figure 8. Posterior predictive distributions for performance at the item level. Each point represents a sample of single item.

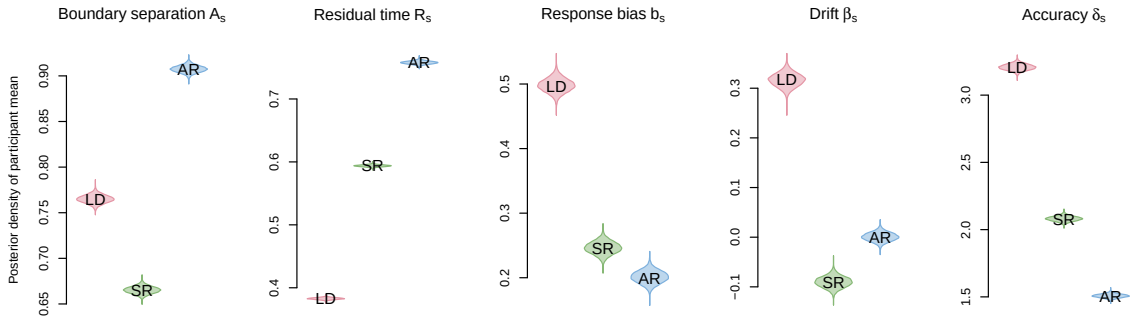


Figure 9. Posterior densities of the mean of participant parameters for the three binary choice tasks, lexical decision (LD), single-item recognition (SR), and associative recognition (AR). Boundary separation parameters are on a log-scale, residual time parameters are in seconds, bias parameters are on a logistic scale, and evidence parameters (β and δ) are measured in units of evidence per second, as described in the main text.

reverse-speed-accuracy trade-off (participants who find the task especially easy don't bother spending as much time accumulating evidence). Finally, we note the negative correlation between overall level of free recall (β_s^{FR}) and free recall accuracy (δ_s^{FR}), meaning that participants who recall many words also tend to produce many intrusions whereas those who recall fewer words do so more selectively.

Our subsequent analyses focus on the individual parameters that directly relate to the memory evidence used by participants in each task, namely, the β and δ parameters reflecting, respectively, response tendencies and response accuracy in each task (see Appendix B for additional analyses of the complete set of participant parameters). In binary choice tasks, β and δ parameters are related to the rate at which participants accumulate memory evidence and can be contrasted with boundary separation, boundary bias, and residual time, which act to transform the underlying memory evidence into an observed response at a particular time.

The many pairwise correlations between individual evidence parameters, shown in Figure 11 extracted from the full matrix of correlation distributions, could well manifest from a smaller set of underlying factors which can be revealed by examining the principal components of each sample of the correlation matrix. As described above, we obtained the eigenvalues and eigenvectors of each sample of the correlation matrix among evidence-related individual parameters. Because eigenvectors are only defined up to a change in sign, we adjusted the sign of each sample of each eigenvector to maximize its dot product with a common basis vector, thereby identifying the resulting components. As shown in Figure 12, there are four eigenvalues that are credibly greater than or equal to one, indicating that the pattern of correlations among individual parameters can be satisfactorily accounted for by four principal components which collectively account for a median of 70 percent (with a 95% credible interval of 69%–72%) of the total correlation among individuals (Larsen & Warne, 2010). To aid interpretation, we performed an orthogonal rotation on each sample of the loading matrix of these top four principal components according to the infomax criterion, which tends to emphasize simplicity of the resulting pattern of loadings (that is, each parameter will tend to load on only one dimension) while discouraging them from

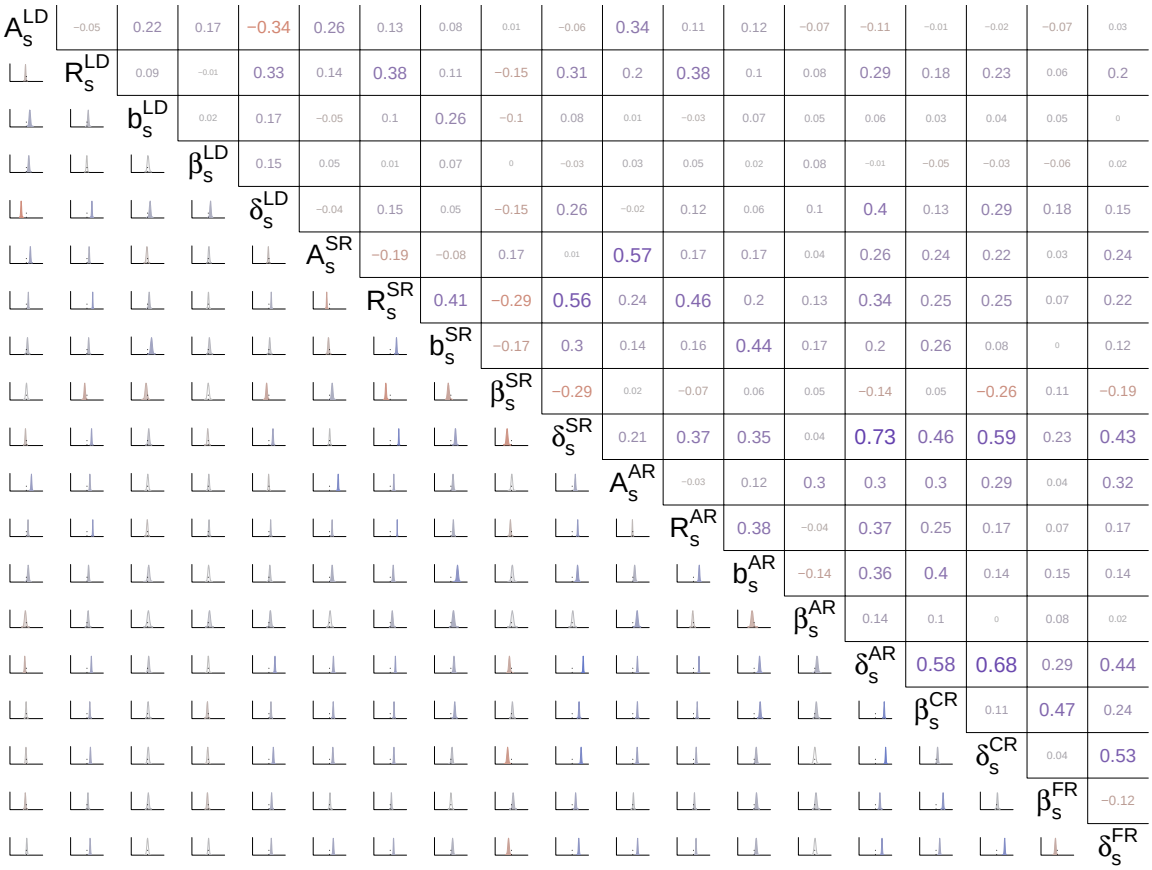


Figure 10. Posterior distributions over correlations between individual participant parameters. Parameter names are given along the diagonal (see Table 2). The lower diagonal depicts the marginal posterior density of each pairwise correlation while the upper diagonal gives the posterior mode of each pairwise correlation. For visualization purposes, colors range between red (negative correlations) and blue (positive correlations) depending on the magnitude of the median correlation and the degree to which the densities in the lower right diagonal are filled reflects the width of the widest highest density interval that excludes zero (smaller for distributions that assign zero a high probability).

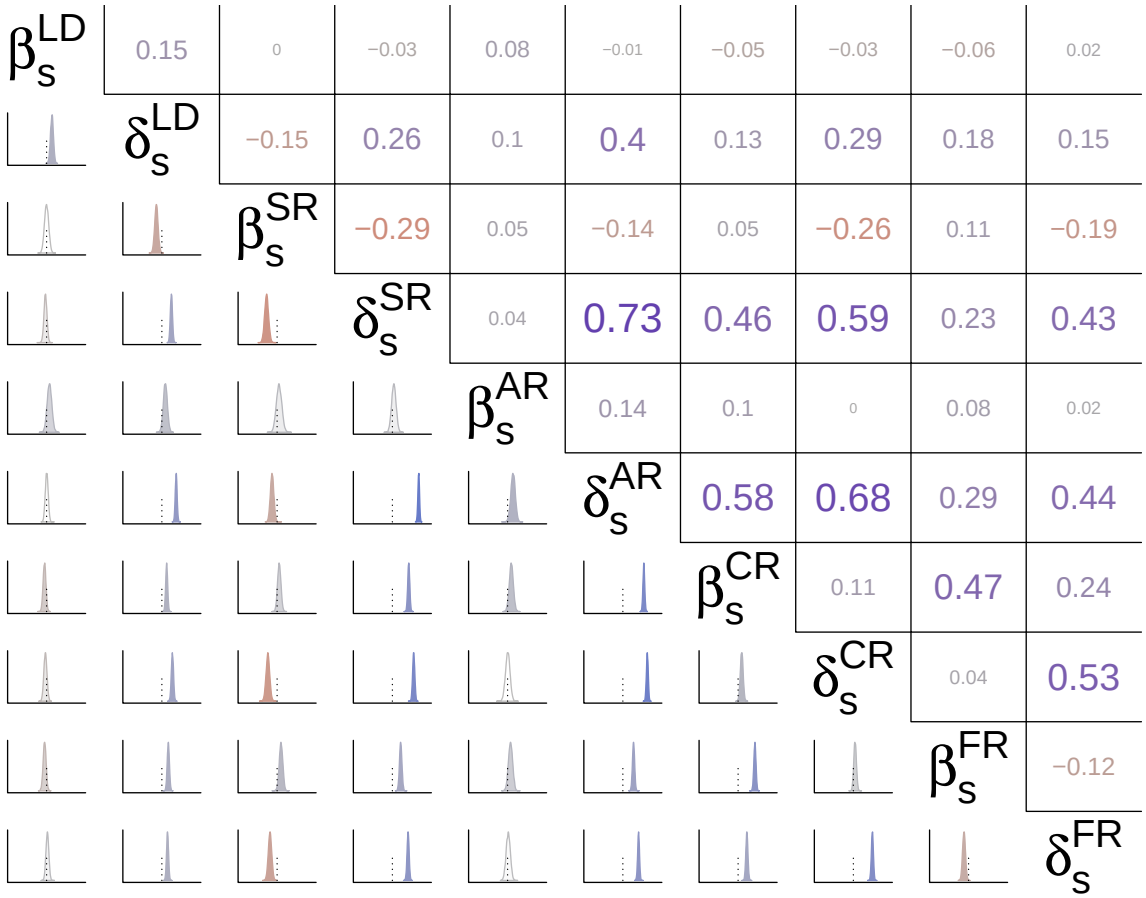


Figure 11. Posterior distributions over correlations between individual participant parameters related to memory evidence (for the full matrix of correlation distributions, see Figure 10). Parameter names are given along the diagonal (see Table 2). The lower diagonal depicts the marginal posterior density of each pairwise correlation while the upper diagonal gives the posterior mode of each pairwise correlation. For visualization purposes, colors range between red (negative correlations) and blue (positive correlations) depending on the magnitude of the median correlation and the degree to which the densities in the lower right diagonal are filled reflects the width of the widest highest density interval that excludes zero (smaller for distributions that assign zero a high probability).

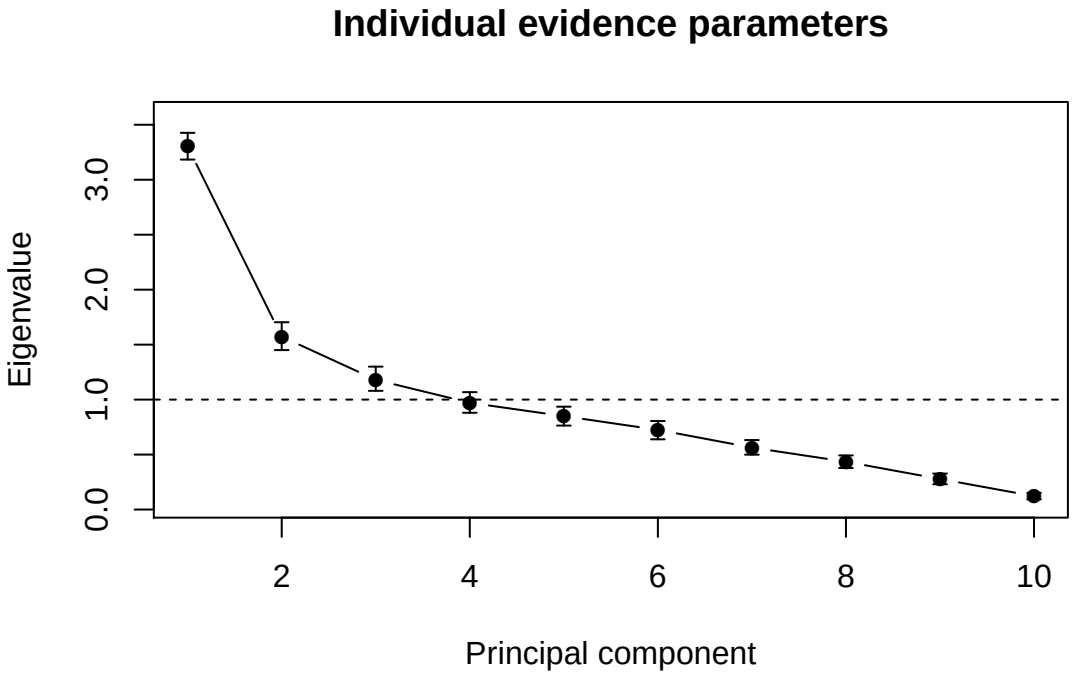


Figure 12. Posterior distribution over eigenvalues of the matrix of correlations between only evidence-related (β and δ) individual participant parameters. Bars depict 95% credible regions and points depict posterior means.

collapsing onto a single component (McKeon, 1968; Browne, 2001). Details about this rotation procedure are given in Appendix A, but we note here that a single rotation was applied across all samples, such that rotation yielded a linear transformation, rather than a distortion, of the principal components. We refer to the resulting rotated components as "factors" (although they are not strictly identical to the results of an additive factor analysis); these factors are a set of basis vectors that represent the most important latent dimensions of the data.

The posterior distributions over the loadings of each parameter on these four factors are shown in Figure 13. Because each factor is orthogonal to the others, it is possible for a participant to vary along one factor without altering the level of any of the others. We now describe how each factor can be interpreted in a way that makes it easier to understand the pattern of correlations among participant parameters. While the factors themselves are a product of the data, the labels we apply to them involve a degree of subjectivity, and others may prefer different labels.

Individual Evidence Factor 1: Episodic accuracy. This first factor corresponds to accuracy across all episodic memory tasks (SR, AR, CR, and FR), which is slightly negatively correlated with drift in SR (β_s^{SR}). That lexical accuracy does *not* load on this factor suggests that it does not reflect general ability, but is specific to tasks that involve episodic memory. Thus, a participant who scores highly on this factor is good at

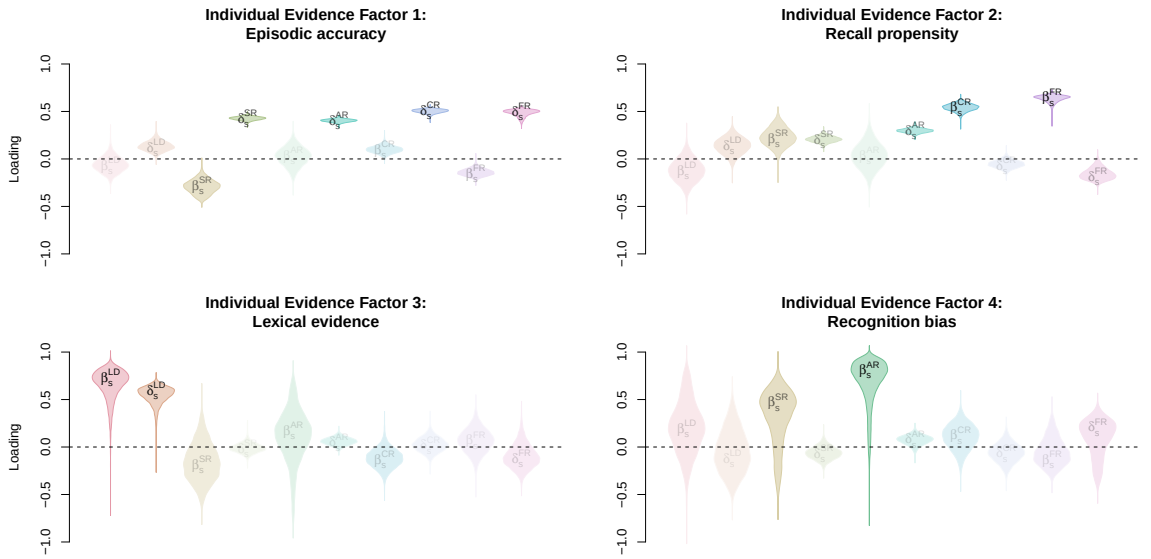


Figure 13. Posterior distributions of the loadings of evidence-related individual participant parameters (see Table 2) on factors formed by orthogonal rotation of their top four principal components. As described in the main text, each factor was assigned a label on the basis of which parameters loaded most strongly on that factor. The saturation of each distribution and label visually indicate the magnitude and uncertainty of each loading.

discriminating between studied and unstudied items and between intact and rearranged pairs and, when they produce a response in a recall task, the response tends to be correct rather than an intrusion.

Individual Evidence Factor 2: Recall propensity. This factor chiefly accounts for the tendency to produce responses in either cued or free recall (β_s^{CR} and β_s^{FR}, respectively), separately from the accuracy of those responses. Accuracy in both single-item and associative recognition also load slightly on this factor, such that high recognition accuracy is associated not only with correct responding in cued and free recall (factor 1, above) but to a lesser extent with overall response rate as well. It is critical to note that, while a participant who scores highly on this factor may produce many responses in cued and free recall, those responses are not guaranteed to be correct, since CR and FR accuracy do not load on this factor.

Individual Evidence Factor 3: Lexical evidence. This factor reflects the positive correlation between overall drift and accuracy in lexical decision. That these are positively correlated (i.e., load in the same direction) implies that participants who score highly on this factor are particularly good at detecting words (for which the evidence accumulation rate is proportional to the *sum* of drift and accuracy), rather than at rejecting pseudowords.

Individual Evidence Factor 4: Recognition bias. The tendency to accumulate positive evidence in single-item recognition is associated with a tendency to accumulate positive evidence in associative recognition, but not with the same tendency in lexical decision. As with factor 1 (above), the distinction between episodic and lexical tasks implies that this factor is related to a tendency to accumulate positive evidence for episodic decisions

Table 4
Examples of the words with the highest and lowest median scores on the four item factors, the loadings of which are shown in Figure 16.

	Item Factor 1: Supports episodic memory	Item Factor 2: Often recalled given a cue	Item Factor 3: Word-like	Item Factor 4: Biased to recognize
Highest	MOM	SUGGEST	DAD	COMMUNITIES
	DAD	FINISH	GOAL	EXPERIMENT
	SEX	STRONGER	DANCE	SECRETARY
	CHINESE	BIGGER	FOOTBALL	MACHINERY
	GRANDMOTHER	EDUCATIONAL	FUNNY	MECHANICAL
Lowest	FORMING	STOCK	CONTINENT	SUGGEST
	RECOGNIZE	HAVEN	CHOSE	ADDING
	GRANTED	NEIGHBORHOOD	OUGHT	FEWER
	INTENDED	WHEAT	RODE	NODDED
	SLIGHT	TOM	ILLUSTRATION	LEG

only, rather than being a general characteristic of binary choice tasks.

Correlations at the item level

Before we analyze the correlations among item parameters, we remind the reader that the pseudoword foils in our LD task were each generated by distorting one of the 924 stimulus words. As such, we parameterized the items in LD in terms of an average evidence drift β_i^{LD} associated with each word, regardless of whether it was distorted or not, and an evidence accuracy δ_i^{LD} reflecting the difference in evidence accumulation between the distorted and non-distorted forms of the word. Because the other tasks only used the original non-distorted words, however, our analysis below will only consider the LD evidence accumulation rate of the non-distorted word, which is given by $\beta_i^{LD} + \delta_i^{LD}/2$. The analysis including the pseudoword drift rates is presented in Appendix C and differs from the following only in that it is harder to interpret correlations between LD drift rates and a word's normative characteristics. This difficulty arises because many of these characteristics are subtly altered when a word is distorted into a pseudoword (e.g., its orthographic regularity may differ) or they are not applicable to pseudowords (e.g., pseudowords are not clearly associated with any particular semantic content). By focusing on only undistorted words, the following analysis avoids this complication.

The posterior distributions over the pairwise correlations between item parameters are shown in Figure 14. Once again, visual inspection hints at the structure inherent in these correlations: There are many strong positive correlations, including between accuracy parameters (δ's) in all episodic memory tasks, as well as overall recall rates (β_i^{FR}, $\beta_i^{\text{CR Cue}}$, and $\beta_i^{\text{CR Tar}}$) and bias in AR (β_i^{AR}). As with individuals, there is a negative correlation between response bias in SR (β_i^{SR}) and episodic accuracy, and correlations between lexical and episodic task parameters are minimal. As with the parameter correlations for individuals, we now employ principal components analysis to better understand the structure of the correlations between item parameters.

As shown in Figure 15, the first four principal components are associated with eigenvalues that are credibly greater than or equal to one, so we focus on these for our analysis

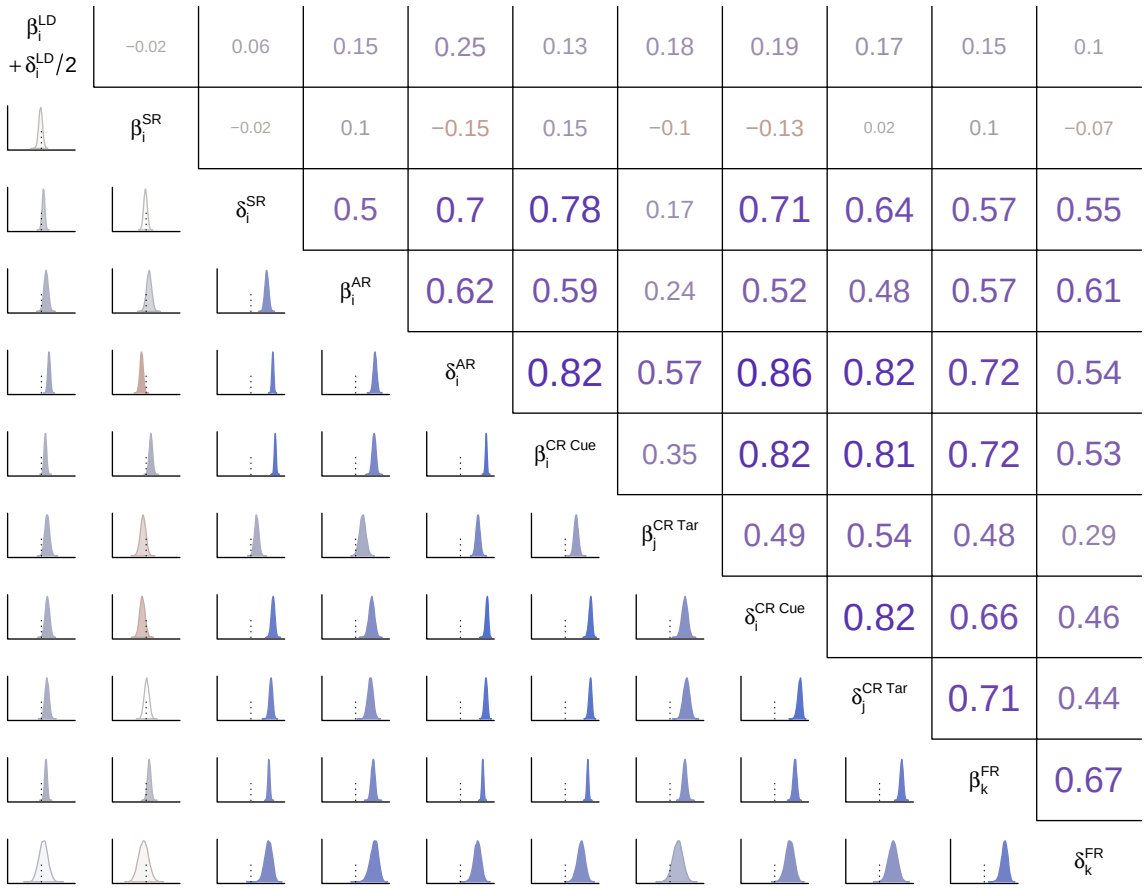


Figure 14. Posterior distributions over correlations between item parameters. Parameter names are given along the diagonal (see Table 3). The lower diagonal depicts the marginal posterior density of each pairwise correlation while the upper diagonal gives the posterior mode of each pairwise correlation. For visualization purposes, colors range between red (negative correlations) and blue (positive correlations) depending on the magnitude of the median correlation and the degree to which the densities in the lower right diagonal are filled reflects the width of the widest highest density interval that excludes zero (smaller for distributions that assign zero a high probability).

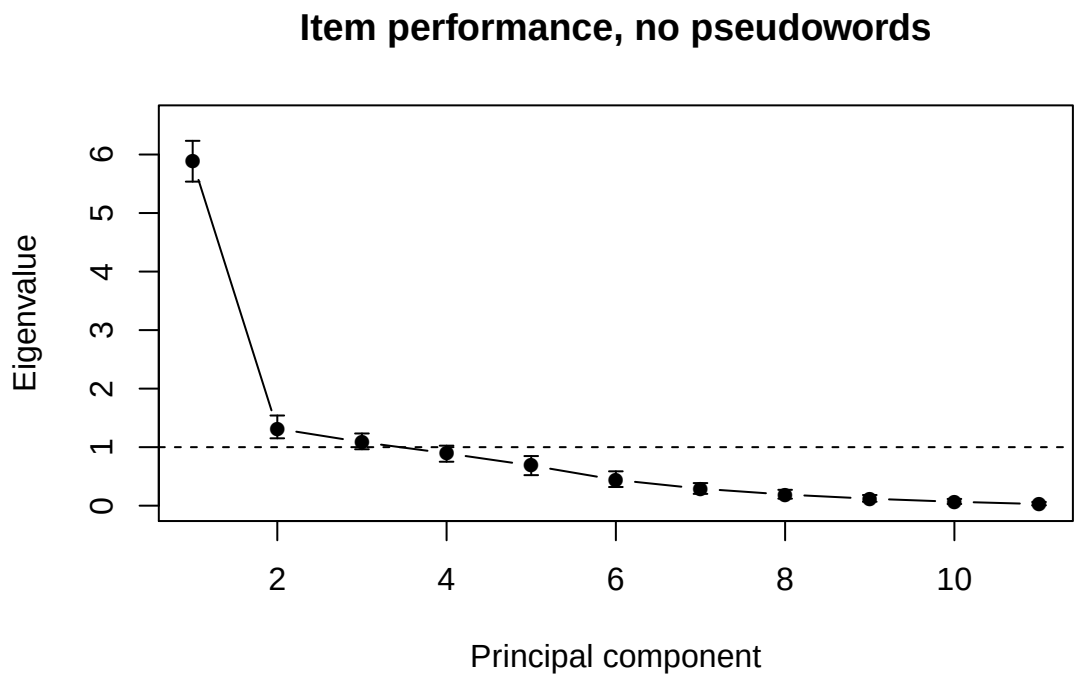


Figure 15. Posterior distribution over eigenvalues of the correlations between item parameters. Bars depict 95% credible regions and points depict posterior means.

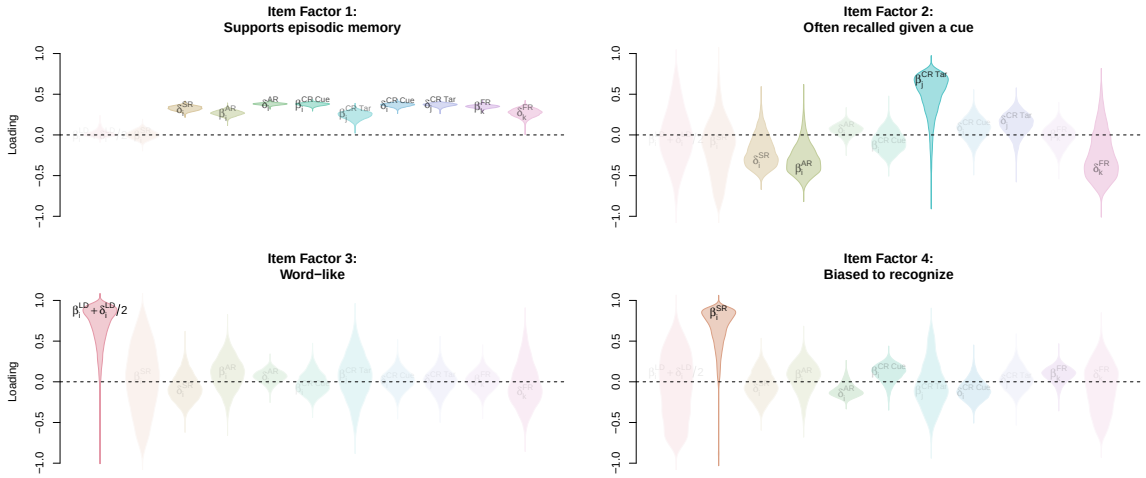


Figure 16. Posterior distributions of the loadings of each item parameter (see Table 3) on factors formed by orthogonal rotation of the top four principal components. As described in the main text, each factor was assigned a label on the basis of which parameters loaded most strongly on that factor. The saturation of each distribution and label visually indicate the magnitude and uncertainty of each loading.

Table 5
Kendall's τ rank correlations between normative characteristics of the item stimuli.

	HAL	Length	OLD20	Num. syll.	PLD20	Concr	SemD	Num. sense	SND
KF	0.36	0.11	0.12	0.15	0.12	-0.16	0.17	0.02	-0.05
HAL	—	-0.03	-0.01	0.02	0.01	-0.12	0.17	0.11	-0.06
Length	—	—	0.73	0.75	0.71	-0.25	0.15	-0.22	-0.05
OLD20	—	—	—	0.69	0.75	-0.26	0.12	-0.20	-0.05
Num. syll.	—	—	—	—	0.73	-0.30	0.18	-0.29	-0.05
PLD20	—	—	—	—	—	-0.26	0.13	-0.27	-0.05
Concr	—	—	—	—	—	—	-0.41	0.10	0.06
SemD	—	—	—	—	—	—	—	0.12	-0.18
Num. sense	—	—	—	—	—	—	—	—	-0.16

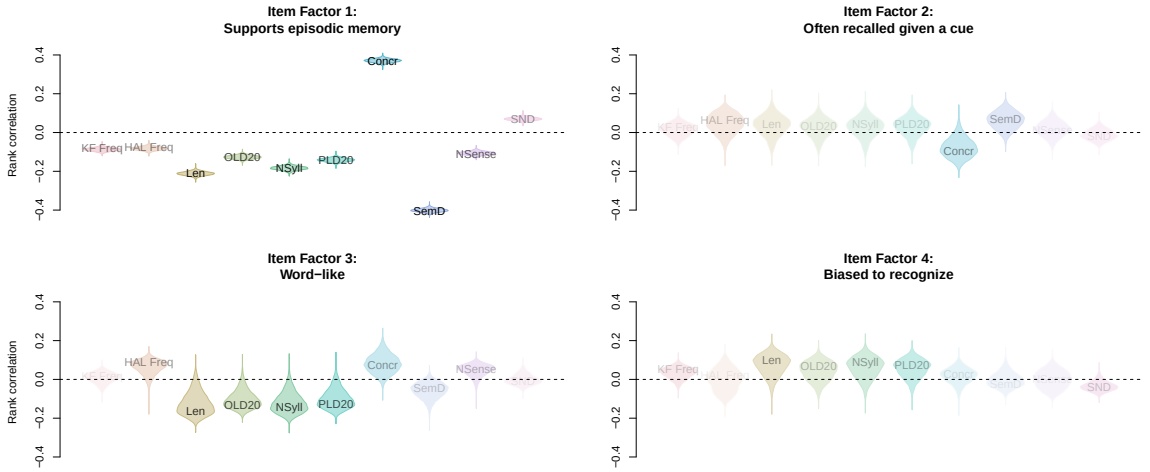


Figure 17. Posterior distributions over Kendall's τ rank correlations between each item's score on each factor (as depicted in Figure 16) and its normative lexical characteristics (as described in Table 1). For visualization purposes, the lightness of each distribution reflects how strongly it deviates from zero.

(these four components account for a median of 83% of variability, with a 95% credible interval of 81%–86%). As above, we rotated the top four principal components according to the "infomax" criterion to obtain interpretable factors that describe the underlying dimensions along which items can independently vary. The posterior distributions of loadings on each factor are shown in Figure 16 and are readily interpreted. To aid intuition, we provide examples of words that have the lowest and greatest median scores on each of these factors in Table 4.

In addition, for each sample of item factor scores, we computed Kendall's τ rank correlations[8] with the normative properties of the items given in Table 1, yielding posterior distributions over rank correlations between factor scores and normative characteristics, shown in Figure 17. Although each of these characteristics are themselves correlated with one another (see Table 5), they can help interpret the nature of the item information that

[8]By taking the rank correlation, we acknowledge the different distributional characteristics of each property as well as the fact that their relationships with memory performance may be nonlinear.

underlies these different factors. Once again, we emphasize that while the factors themselves arise from correlations in the data, it is possible to assign different labels to these factors than those we have provided.

Item Factor 1: Supports episodic memory. This factor accounts for accuracy across all episodic tasks, as well as drift in associative recognition (β_i^{AR}), the propensity for either a cue or target to elicit a response in cued recall ($\beta_i^{\text{CR Cue}}$ and $\beta_i^{\text{CR Tar}}$), and the tendency for an item to be produced in free recall (β_i^{FR}). The fact that this factor accounts for both drift and accuracy in AR suggests that it reflects an item's ability to enhance recognition of intact pairs rather than to reject rearranged pairs. Coupled with the fact that this factor is also related to accuracy and overall responding in cued recall, this suggests that items that yield good recognition performance (δ_i^{SR}) also yield the encoding of stronger associations (which manifests in increased recognition of intact pairs and correct responding in cued recall). Word frequency and orthographic/phonological complexity are somewhat negatively correlated with this factor, however the strongest correlations are with a word's semantic qualities, namely, concreteness and semantic diversity. A word tends to score highly on this factor if it refers to a specific concrete entity (high concreteness) and/or occurs only in specific discourse contexts (low semantic diversity). Concreteness and semantic diversity are themselves negatively correlated, however; words that refer to specific entities tend to be used in specific settings (and therefore also have fewer senses).

Item Factor 2: Often recalled given a cue. This factor accounts for the tendency of an item to elicit a response when it is the target in cued recall, while other item parameters either have weak or uncertain (broadly distributed) loadings on this factor. Coupled with the fact that this factor does not demonstrate any strong correlations with normative lexical properties, it is likely that this factor reflects idiosyncratic properties of a word that allow it to seem "target-like" in cued recall.

Item Factor 3: Word-like. This factor accounts primarily for the degree to which a word yields positive lexical evidence (given by $\beta_i^{LD}+\delta_i^{LD}/2$, as described above). As would seem logical, this property is negatively associated with measures of word length (number of letters/syllables) and complexity (OLD20 and PLD20), such that shorter and/or less complex words are more readily identified as such. This factor is only weakly, albeit positively, correlated with normative word frequency (at least using the HAL measure), suggesting that, at least over the range of frequencies represented in our stimulus pool, overall prior occurrence of a word plays less of a role in LD than its orthographic/phonological properties. Correlations with semantic measures are also weak, again emphasizing the importance of low-level features for word identification.

Item Factor 4: Biased to recognize. Finally, this factor accounts for the tendency for a word to elicit a positive recognition response, irrespective of whether it had been studied or not (β_i^{SR}). As with factor 2, reflecting a bias to produce responses in CR, a word's score on this factor is not strongly correlated with any of its normative properties, implying that this factor pertains more to idiosyncratic properties of words and how participants engaged with them, rather than anything that systematically holds across the lexicon.

Comparison of individual and item factors

In many ways, the patterns of correlations we found between item and individual parameters break down along similar lines, at least for the parameters that are directly comparable—namely, the evidence-related β and δ parameters. When focusing on only these evidence-related parameters, four factors were found to describe their correlational structure for both items and individuals. In broad terms, these factors related to 1) episodic memory tasks, 2) recall rate, 3) lexical decision, and 4) recognition bias. As described in Appendix D, this general breakdown of task parameters into factors remains consistent for both individuals and items, even as the number of principal components used to construct these factors differs, indicating that these are dimensions that most clearly define the structure of the correlations within and between tasks. While the sets of factors describing items and individuals are similar in many ways, they differ in other important aspects:

Episodic memory. For both individuals and items, the accuracy of episodic memory across all tasks was related. While only accuracy-related parameters tended to load together for individuals, for items both episodic accuracy and response bias parameters (β's) loaded together. This suggests that the bias-related parameters for individuals can be interpreted separately from their episodic memory ability, with ability arising from encoding and retrieval processes and bias arising from decision-related processes. For items, as noted above, the coupling of accuracy and bias parameters indicates that this factor represents an item's ability to support correct memory (the sum of β and δ) rather than its resistance to false memory.

Recall rate. For individuals, the rate at which they produce responses in both cued and free recall loaded together, and were slightly correlated with accuracy in single-item and associative recognition. For items, only the rate at which a target item elicits responses loaded on a separate factor, with uncertain relation to other item parameters or normative characteristics. Thus, while the first two factors for both items and individuals delineate the relationships between different aspects of episodic memory performance, they do so along different lines. The memory processes engaged by individuals can be described in terms of overall across-task accuracy (individual factor 1) and rate of responding in recall tasks specifically (individual factor 2), while the information provided by items can be described by across-task accuracy and response rate together (item factor 1) with a special role for target response rate in cued recall (item factor 2). This points to a difference between how cue and target items affect cued recall performance—the quality of a cue is more closely related to its other mnemonic properties (e.g., it's ability to be correctly recognized or freely recalled) while the quality of a target can be influenced by other factors.

Lexical decision. Parameters related to lexical evidence loaded on separate factors for both individuals and items, emphasizing that not only do episodic and lexical memory engage different processes (at the individual level), they rely on different kinds of information (at the item level).

Bias. For items, a bias to be recognized as having been studied loaded on its own factor while, for individuals, a tendency to accumulate positive recognition evidence for single words was related to the tendency to do the same thing for word pairs. As noted above, however, the apparent idiosyncratic nature of an item's recognition bias—which was not correlated with any of a word's normative properties—means that it is difficult to identify

any regularities in the information carried by an item that tends to yield positive recognition responses. Instead, there appears to be more structure in how participants accumulate recognition evidence, where similar processes appear to be involved in recognizing words and word pairs.

Discussion

We presented results from a large-scale study investigating how individual performance was correlated between different memory tasks and how the information contained within different items supports performance in each of these tasks. We analyzed these data using a hierarchical Bayesian model to simultaneously estimate parameters reflecting individual performance and item contributions for each task. We interpreted the resulting correlations among item and individual parameters with the help of their principal components, identifying latent dimensions that describe how groups of parameters between and within tasks covary. We identified four latent dimensions that characterized the variation among participant performance as well as four similar dimensions reflecting how item information contributed to performance in each task. The four evidence-related dimensions for items and individuals related to episodic memory accuracy, rate of recall, lexical memory, and bias. We further investigated correlations between item dimensions and normative characteristics of words, finding that concreteness and semantic specificity support episodic memory for words while lexical memory largely depends on the orthographic/phonological properties of a word. Below, we summarize our key results and explicate how they fit with and can help to unify existing memory theory, followed by a discussion of future analyses and implications for individual differences.

Theoretical implications

The major findings from our analyses can be summarized as follows:

1. Lexical access depends primarily on orthographic/phonological information. Words that are shorter and/or more orthographically/phonologically typical are easier to identify as words.

2. Episodic memory depends primarily on semantic information. Memory for single items and associations as well as the ease with which an item can be recalled are all enhanced when a word refers to a specific concrete entity and/or is used only in specific discourse contexts.

3. Among individual participants, accuracies across all episodic memory tasks are strongly correlated with one another but only weakly correlated with accuracy in lexical decision. This suggests that while all episodic memory tasks rely to an extent on a shared set of processes, these are largely separate from those involved in lexical access.

4. For both items and individuals, response rate in recall tasks involved a separate factor from other episodic tasks, suggesting that recall involves additional processes and information beyond those involved in other episodic memory tasks.

As we describe below, many of these results can be related to the structure of global memory models (e.g., Murdock, 1982; Gillund & Shiffrin, 1984; Hintzman, 1988; Humphreys et al., 1989; S. E. Clark & Gronlund, 1996), particularly in the claim that different memory tasks involve similar memory structures that are simply accessed in different ways depending

on the task. Further, the distinction we identified between lexical and episodic memory, popularized by Tulving (1985), is closely related to modal models of memory (Atkinson & Shiffrin, 1968) and to the reactivation theory presented by Bower (1996), in that while semantic and episodic memory may be separable, they are mutually dependent. Indeed, the differences we found in residual time needed to access the relevant memory evidence (Figure 9) are consistent with the kind of staged processing proposed by these models. In the following, we describe how our results can help to further develop global theories of memory that extend beyond individual tasks and help explicate the relationship between lexical, semantic, and episodic memory.

Orthography and lexical access. Orthographic and phonological information play a central role in many models of lexical access (e.g., McClelland & Rumelhart, 1981; Jacobs & Grainger, 1994; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001). According to these models, when a string of letters is presented, its orthographic features (letters and letter combinations) are used to activate traces in lexical memory corresponding to individual words or senses (sometimes called "logogens"). These lexical entries are activated to the extent that their orthographic features are similar to those of the presented string, such that words with more regular spellings will tend to be activated by more word-like letter strings, owing to their overlapping orthographic features. Semantic aspects of a word may also play a role in lexical access, with multiple senses or rich semantics providing a boost in overall lexical activation (Buchanan et al., 2001; Pexman, Hargreaves, Siakaluk, Bodner, & Pope, 2008; Yap, Tan, Pexman, & Hargreaves, 2011), we found only weak correlations between lexicality and semantics. The overall degree of lexical activation engendered by a letter string is used to make a lexical decision (e.g., Wagenmakers et al., 2004).

We note that features of the current context must also play a role in lexical decision, owing to the ubiquitous long-term priming effects that have been found in this task (Scarborough et al., 1977; Logan, 1988; Schooler, Shiffrin, & Raaijmakers, 2001). Since individuals in our experiment did not encounter the same word in more than one block, we are not in a position to assess the degree to which context may play a role in lexical decision. It is possible, for instance, that the boost granted to a word by virtue of having multiple senses may be due to context, with these words having a higher baseline level of activity by virtue of being used in multiple ways. It has also been found that repetitions of words that retain specific features (e.g., speaker or prosody) yield improved lexical access, suggesting that episodic (or at least context-specific) memory can interact with lexical access (Goldinger, 1998).

The role of semantic information. Lexical memory associates the phonological and orthographic features of a word with its semantic features. Semantic features must, of course, be learned over time via linguistic training and experience. Many theories of semantics postulate that these semantic features reflect not only the features of the entities to which a word refers (e.g., "four leggedness" and "furry" for dogs Collins & Quillian, 1969; Rosch & Mervis, 1975; E. E. Smith, Shoben, & Rips, 1974) but their relationships to other words and the situations in which they are used. The notion that semantics are derived from patterns of use—reflected in the oft-cited aphorism that "you shall know a word by the company it keeps" (Firth, 1957)—is ubiquitous in computational linguistics and information retrieval, serving as the basis for systems like Latent Semantic Analysis (Landauer & Dumais, 1997) and Topic modeling (Griffiths, Steyvers, & Tenenbaum, 2007), both of

which have been shown to account for human semantic judgments. Other psychologically-motivated contextual models of semantics (e.g., Lund & Burgess, 1996; Jones & Mewhort, 2007) also embody the idea that what we call semantics can be interpreted as a word's aggregate pattern of use. While most of these models are based purely on linguistic co-occurrence data, this turns out to be strongly correlated with perceptual experience (Riordan & Jones, 2010), consistent with the strong negative correlation between concreteness (perceptual content) and semantic diversity (pattern of use) in our study (see Table 5). Indeed, even very abstract concepts may have a basis in concrete perception (Barsalou, 1999). Further, developmental studies have found that even very young infants are sensitive to the aggregate patterns of co-occurrence between words and between words and their environment (L. B. Smith & Yu, 2008; L. B. Smith, Suanda, & Yu, 2014), suggesting that such "statistical" information is fundamental to word-learning.

Our results suggest that both single words and word pairs are encoded in terms of the semantic features associated with the words involved, such that more distinctive semantic features yield stronger memory at both encoding and retrieval. Here, "distinctive" semantic features means that a word refers to a specific concrete entity, and is thereby associated with perceptual features of that entity (Paivio, 1969), and/or it is used only in specific discourse contexts (low semantic diversity) and is therefore associated with a narrow set of patterns of use (Adelman et al., 2006). These strong memories manifest at encoding by virtue of allowing for better recognition of intact pairs and better recall of an associated word. Strength of semantic features must also play a role at retrieval, in that it is associated with discriminability in single-item recognition (i.e., not just recognition of studied words, but the ability to *reject* an unstudied but semantically distinctive word). Feature distinctiveness and the resulting differentiation between event memories is a core component of likelihood-based models of episodic memory (McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997) and has been found to be critical for explaining a variety of memory phenomena across a variety of stimuli (not just words; Cox & Shiffrin, 2017; Kılıç, Criss, Malmberg, & Shiffrin, 2017; Nosofsky & Zaki, 2003; Steyvers & Malmberg, 2003).

Relation between memory for items and associations. We found that accuracy across all episodic memory tasks were mutually correlated, however single-item and associative recognition were correlated particularly strongly, with accuracy on each task loading on the same factors for both individuals and items. This implies that these two tasks, despite ostensibly asking different questions ("was this item studied?" vs. "given that these items were studied, were they studied at the same time?") actually rely on similar memory processes and similar information stored in memory. This is in contrast to many dual-process theories which assume that associative recognition can only be carried out using a secondary recall-like process that is independent of the processes involved in recognition of single items (e.g., Jacoby, 1991; Yonelinas, 1997).

Our results do, however, accord with a growing body of evidence for a close similarity between recognition of single items and of associations. This similarity could arise either because items and associations are encoded using similar information or because the processes used to retrieve item and associative information are related, or both. We are not in a position to adjudicate this question here, although we note that other work from our laboratory supports the idea that item and associative information are *both* stored *and* retrieved using similar processes (Cox & Criss, 2017), and that memory for associative information

involves elaborating or interrelating information about items (Cox & Shiffrin, 2017; see also McGee, 1980; Dosher, 1984; Dosher & Rosedale, 1989, 1991, 1997). Consistent with the idea that item and associative information are *encoded* in a similar manner, we found that the same item properties, like concreteness and semantic specificity, that make a single word easy to recognize (and to reject when unstudied) also make it easy to recognize a *pair* in which that word appears. Finally, the fact that an item's ability to support accurate cued recall (as either a cue or a target) loads on the same factor as its ability to support accurate single-item and associative recognition lends credence to the idea that these correlations arise from how well an item supports the encoding of an association.

Additional processes in recall. Recall parameters loaded on additional factors for both items and individuals, suggesting the involvement not only of additional processes in recall tasks but additional information as well, beyond that relevant to other episodic memory tasks. Two aspects of recall distinguish it from recognition, which may account for the additional processes and information involved: first is the increased importance of context; second is the need to produce an item without support from an explicit cue.

Context is more important in recall—particularly free recall—than in recognition in part because of the structure of our experiment. Context is, by definition, crucial for any episodic memory task—at least those that use familiar stimuli, like words—since participants must be able to distinguish occurrences of an item during the experiment from those prior to the experiment. The distinction between lexical decision and single-item recognition makes this clear: in lexical decision, participants should give a positive response if they have encountered the test item *at all*, whereas in single-item recognition, they should only give a positive response if they have encountered the test item *in the preceding study context*. However, because no participant would encounter the same item in more than one study/test block, the contextual information needed for recognition tasks (SR and AR) would not need to distinguish between different blocks, only between the experimental context and context outside the experiment. While a variety of obvious features distinguish the experimental context from other experience (e.g., the unusual tasks, the unusual setting, etc.), few such features distinguish between blocks *within* our experiment (cf. Mensink & Raaijmakers, 1988; Klein, Shiffrin, & Criss, 2007), as required for recall and as made evident in the pattern of prior-list intrusions in FR. Indeed, to the extent that the current context were used as a retrieval cue, one would expect a greater mnemonic advantage for recent experience *within* a list as well, consistent with the presence of recency effects in CR and FR and the absence of such in either SR or AR (see Figure 18; note also the primacy effect that is exclusive to FR). It may be that prominent items—those that are more easily recognized and/or recalled—can overcome any mismatch between the current list context and that of prior lists, such that the ability to produce *only* items from the most recent study list entails additional processes on the part of participants. These processes may involve forming associations between the specific study context and a sufficiently distinctive array of semantic features (Raaijmakers, 2003). In addition, a semantically distinctive item may also be more effective in driving the formation of the temporal context associated with the list (Howard & Kahana, 2002; Polyn, Norman, & Kahana, 2009).

Recall also entails producing a response "from scratch", rather than choosing between a set of given alternatives (like "yes" or "no"). For that reason, one might expect either that overall rate of recall would be related to verbal/lexical ability or that more typical

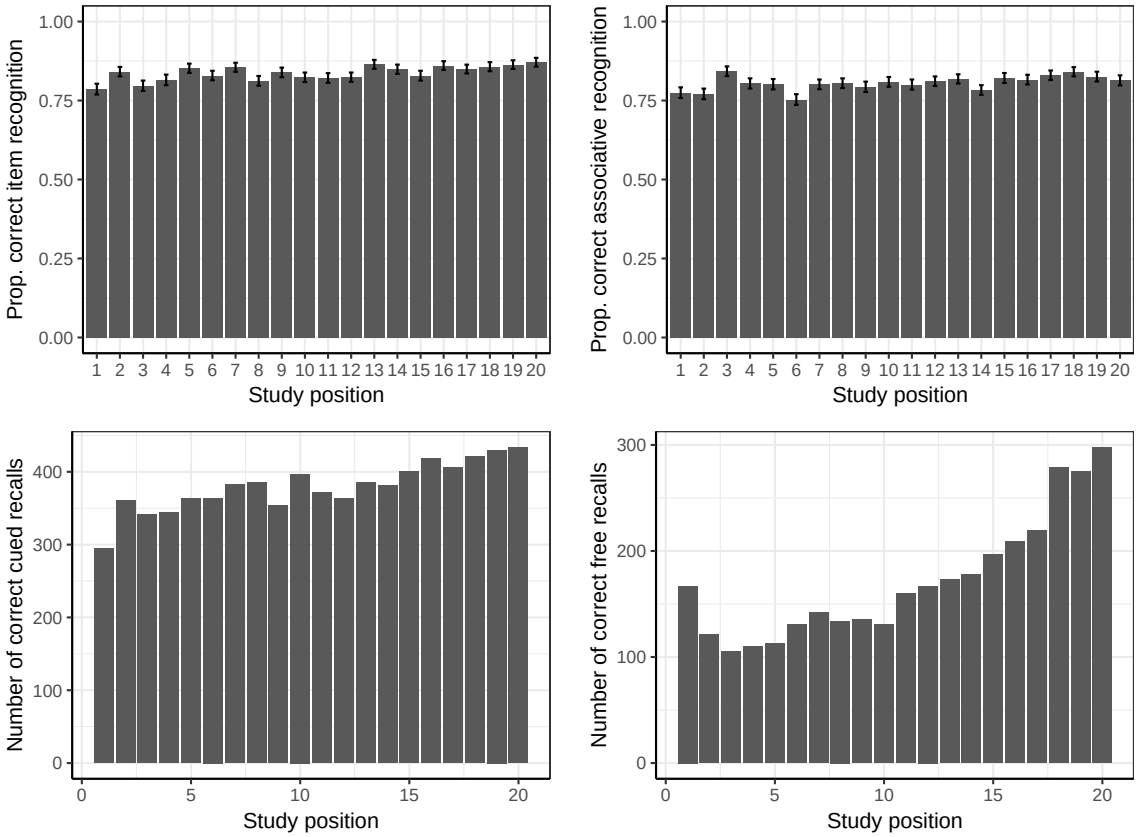


Figure 18. Top panels depict the probability of correct recognition as a function of serial position within the study list of either the item (for SR, top left) or pair (for AR, top right); error bars in these panels denote ±1 within-subject standard error of the mean. Bottom panels depict the number of observed correct responses in cued recall (bottom left) and free recall (bottom right) as a function of serial position within the study list. Note that, because words were studied in pairs, study position refers to the position at which the *pair* containing the word was studied.

words would be produced more often in recall. While there are positive correlations between CR and FR parameters and lexical memory (δ^{LD}) for both items and individuals, they are not strong enough to emerge as separate factors in our analyses. However, the degree to which a word supports good episodic memory was negatively correlated with orthographic/phonological complexity, albeit not as strongly as with its semantic characteristics. This is consistent with a role for lexical memory at encoding, as noted above, but perhaps also at retrieval: If, as we have found, episodic memory is encoded largely in terms of semantic features, then it is these features, not orthographic/phonological ones, that must be used to decide what word to produce. In essence, producing a recall response reverses the process of episodic storage: During storage, orthography is used to access lexical memory, from which semantic features are extracted and used to encode the item. During recall, stored semantic features are used to access lexical memory, from which orthography is extracted and used to produce the item. Various extant models involve something like

this production process (e.g., J. A. Anderson, 1973; Metcalfe Eich, 1982; Hintzman, 1984; Plate, 2003), which is sometimes referred to as "recovery" of an item, in contrast to the "search" process that isolates a memory trace (e.g., Gillund & Shiffrin, 1984; Raaijmakers & Shiffrin, 1981). We note, however, that once participants have produced a response, it is possible for them to use that response to cue subsequent retrieval, as postulated by many models of recall (Raaijmakers & Shiffrin, 1981; Howard & Kahana, 2002; Polyn et al., 2009). The extent to which this "auto-cuing" plays a role in our free recall data is currently under investigation, and we note that recent work has identified this ability as an important indicator of individual differences (Healey, Crutchley, & Kahana, 2014; Healey & Kahana, 2016; Kılıç, Criss, & Howard, 2013).

Relation between cued recall and other tasks. For individuals, the accuracy of their responses in cued recall was closely related to their accuracy in all other episodic tasks, including those that only require item information (e.g., single-item recognition). Individuals' response rates in cued recall were, however, more closely associated with their response rates in free recall than with those in recognition. Coupled with the observations of a strong recency effect in free recall, weak recency effects in item or associative recognition, and a moderate recency effect in cued recall (Figure 18), cued recall appears to exhibit properties of both recognition and free recall. While accurate cued recall depends on the same memory processes that support recognition, responding in cued recall relies in part on the contextual cuing and response generation processes needed in free recall.

Relation between lexical and episodic memory. Taken together, the weak correlations between lexical and episodic accuracy and the relative importance of semantic versus orthographic/phonological information for lexical and episodic tasks suggests that lexical access and episodic retrieval are largely separate, although as we noted above, they interact in important ways. Indeed, the very fact that semantic information is so important for episodic memory would imply that information stored in lexical memory is crucial for forming episodic memories. That lexical access acts as a "first stage" prior to the encoding of semantic features in episodic memory is consistent with the greater speed at which people can make lexical decisions versus recognition memory decisions (Figures 4 and 9; see also Hintzman & Curran, 1997) and is a key feature of the modal model of memory (Atkinson & Shiffrin, 1968) and the reactivation theory of Bower (1996). Ease of lexical access also aids the formation of associations between words. This reinforces the notion, noted above, that associations are encoded by elaborating or interrelating the semantic features of the items involved—to the extent that such features are easy to access, they can enter the elaboration/interrelation process more readily. In order for the semantic features extracted from lexical memory to be helpful, however, they must be sufficiently distinct (as noted above); highly typical features will tend to be shared by many words, thereby diluting the specificity of any associations encoded based on those features (note the relation between an item's word-likeness and tendency to be produced in cued recall).

Over longer timescales than those involved in the present study, episodic memory must also contribute to lexical memory in that the perceptual and conceptual features stored in lexical memory must be learned from experience: perceptual features of an object referent must be associated with its label, giving rise to the importance of concreteness; and conceptual features of the discourse contexts in which a word is used grow to be associated with a word, giving rise to the importance of semantic diversity/specificity. In general, our

results support a role for separate but interactive lexical and episodic retrieval systems (e.g., Kumaran & McClelland, 2012; Nelson & Shiffrin, 2013).

Limitations and future directions

In the present study, we presented a large-scale correlational analysis with the aim of characterizing the broad outlines of how memory is deployed across various tasks. While we focused on the strongest sets of correlations, since these are most likely to be robust, taking this broad view meant that we were forced to skip over many fine details. However, we do not expect this will be the final word with regard to this dataset or the tasks represented therein. In addition to ongoing work in our laboratories, by making our data and analysis methods publicly available (via the Open Science Framework; `osf.io/hctyg`), we enable and encourage other investigators to conduct their own explorations of the relationships between different memory tasks among items and individuals, including to develop and test more elaborate cognitive models of these tasks than the relatively simplistic measurement models we employed as part of our analyses.

One interesting avenue for future investigation concerns the within- or between-trial dynamics of recall in either cued or uncued settings. Within-trial dynamics in cued recall have been studied more extensively than that for free recall (see, e.g., Diller, Nobel, & Shiffrin, 2001; Nobel & Shiffrin, 2001; Aue, Criss, & Novak, 2017; Sederberg, Howard, & Kahana, 2008; Hopper & Huber, 2016), however unlike with binary choice tasks, there is not even a marginally agreed-upon set of modeling frameworks for addressing recall dynamics. Across-trial dynamics have been found to be especially important in free recall (Howard & Kahana, 2002; Polyn et al., 2009) and test position effects are often found in recognition as well (Roediger & Schmidt, 1980; Schwartz, Howard, Jing, & Kahana, 2005; Criss, Malmberg, & Shiffrin, 2011; Aue, Criss, & Prince, 2015). While including these factors in the present analyses would go far beyond the scope of the present study, deeper understanding of recall dynamics will undoubtedly be a boon to the study of memory in general.

While we studied the relationship between items and the entire lexicon—their orthographic and phonological regularity, their semantic distinctiveness, etc.—we did not study the relations among items within each study list. This is an active and important area of study (Criss & Shiffrin, 2004; Freeman et al., 2010) and another that we and others are pursuing. Unfortunately, estimating item-item interactions in a purely data-driven manner would require orders of magnitude more observations than even the large amount we already had, since one would need at least one observation for each combination of two items. Bringing in outside measures of word-word similarity, on the other hand, is highly model-dependent in that different measures of semantic and/or orthographic similarity/relatedness depend on models of orthographic and semantic representation. The aim of the present study was to see what could be concluded from the data alone, without strong commitment to any particular models, thus we did not include these aspects in our broad-level overview.

Finally, we note some features of our design that preclude certain inferences: First, our design means that we cannot estimate interactions/correlations *between* items and individuals. To measure such interactions would require multiple observations for each combination of item and individual, which would entail either a prohibitive amount of data collection or a much more restricted set of stimuli. While there is some virtue in using a smaller stimulus pool, doing so would limit the ability to generalize beyond a limited set of stimuli and make

inferences about the relationship between normative stimulus characteristics and memory performance, as we were able to accomplish using a larger stimulus pool. Second, while informing participants of the task only *after* study means that the study situation is effectively balanced across tasks, it also means that we cannot study any interactions between study strategy and test task. For example, if participants did not know that associative information was important, would we have found as strong of a correlation between single-item and associative recognition? Perhaps not, although we note that while participants can adjust the degree to which they encode associative information, encoding more associative information does not impair the encoding of item information (Hockley & Cristi, 1996) nor does expectation strictly govern what material is retained (R. C. Anderson & Pichert, 1978), suggesting that any interactions between study and test, at least for the kinds of materials and tasks we examined, are unlikely to distort the picture that emerged from our efforts.

Implications for individual differences

Our analyses imply that certain memory processes may be functionally dissociated from one another at an individual level, such that they may be related to various individual differences such as age or intelligence/capacity (as noted in the introduction). Although episodic memory performance in general declines with age, it does so especially for associative recognition and recall tasks (e.g., Naveh-Benjamin, 2000; Zacks, Hasher, & Li, 2000; Ratcliff et al., 2011). Decrements in recall performance have been attributed to the need for additional resources beyond those needed for recognition (Craik & McDowd, 1987) as well as an ability to use temporal context as an effective retrieval cue (Healey & Kahana, 2016); both are consistent with a separate factor related to responding in recall (individual factor 2). Differences in measures of working memory capacity—not necessarily related to age—have been used to explain differences in the ability to conduct controlled search of episodic memory (Oberauer, 2005) and that such processes are related to "fluid" intelligence (Unsworth & Engle, 2007). Consistent with an interpretation of intelligence/capacity as being related to controlled use of context in retrieval, Healey et al. (2014) reported that a factor pertaining to temporal context predicted IQ better than other aspects of free recall (although recall accuracy remained a good predictor of intelligence as well).

While the selective decrease in recall performance with age and/or working memory is consistent with a functional dissociation between overall episodic memory quality (individual factor 1) and recall rate (individual factor 2), a selective decline in associative recognition over item recognition (Ratcliff et al., 2011) appears to conflict with our results. Further complicating matters, a working memory deficit has been invoked to explain the decline in associative recognition with age (Buchler, Faunce, Light, Gottfredson, & Reder, 2011), just as with recall, but in this case working memory is needed not just to retrieve, but to encode the association as well. It is possible that we did not detect a similar distinction because of the limited age range of our participants, although we note that, unlike in many other studies of memory and aging, our participants could not "tune" their encoding to a specific memory task; thus it is possible that some differences in memory performance with age could result from older adults adopting different study strategies than younger adults.

Concluding remarks

We have presented results of a large-scale correlational analysis examining how different memory tasks engage similar or different processes and rely on similar or different information. We found that the processes and information involved across a variety of episodic memory tasks were largely similar, relying primarily on semantic information, and were generally distinct from the information and processes involved in lexical access, which depended primarily on orthographic/phonological information. Episodic tasks that depended on memory for associations were strongly correlated with those that depended on memory for single items, for both individuals and items, suggesting that such tasks rely on similar processes and that item and associative information are encoded in a similar manner. Recall differed from recognition primarily in terms of response rates, which we hypothesized to be due to the relative importance of contextual information in recall and the need to select a response on the basis of a more impoverished cue. Although a complete portrait of the information and processes underlying human memory requires detailed accounts of each task, our analyses provide a glimpse of its overall shape that can guide future theoretical developments.

References

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*, 814–823.

Anderson, J. A. (1973). A theory for the recognition of items from short memorized lists. *Psychological Review*, *80*(6), 417–438.

Anderson, R. C., & Pichert, J. W. (1978). Recall of previously unrecallable information following a shift in perspective. *Journal of Verbal Learning and Verbal Behavior*, *17*, 1–12.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 2, pp. 89–195). New York: Academic Press.

Aue, W. R., Criss, A. H., & Novak, M. D. (2017). Evaluating mechanisms of proactive facilitation in cued recall. *Journal of Memory and Language*, *94*, 103–118.

Aue, W. R., Criss, A. H., & Prince, M. (2015). Dynamic memory searches: Selective output interference for the memory of facts. *Psychonomic Bulletin and Review*.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*, 577–660.

Bower, G. H. (1996). Reactivating a reactivation theory of implicit memory. *Consciousness and Cognition*, *5*, 27–72.

Brown, S., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*, 153–178.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*(1), 111–150.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, *46*, 904–911.

Buchanan, L., Westbury, C., & Burgess, C. (2001). Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin & Review*, *8*(3), 531–544.

Buchler, N. G., Faunce, P., Light, L. L., Gottfredson, N., & Reder, L. M. (2011). Effects of repetition on associative recognition in young and older adults: Item and associative strengthening. *Psychology and Aging*, *26*(1), 111–126.

Burgess, C., & Livesay, K. (1998). The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kucera and Francis. *Behavior Research Methods, Instruments, & Computers*, *30*(2), 272–277.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic–cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*(3), 432–459.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1).

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335–359.

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review*, *3*(1), 37–60.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*, 240–247.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*(1), 204–256.

Cox, G. E., & Criss, A. H. (2017). Parallel interactive retrieval of item and associative information from event memory. *Cognitive Psychology*, *97*, 31–61.

Cox, G. E., & Shiffrin, R. M. (2017). A dynamic approach to recognition memory. *Psychological Review*, *124*(6), 795–860.

Craik, F. I. M., & McDowd, J. M. (1987). Age differences in recall and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(3), 474–479.

Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory. *Journal of Memory and Language*, *64*, 316–326.

Criss, A. H., & Shiffrin, R. M. (2004). Context noise and item noise jointly determine recognition memory: A comment on Dennis and Humphreys (2001). *Psychological Review*, *111*(3), 800–807.

Diller, D. E., Nobel, P. A., & Shiffrin, R. M. (2001). An ARC-REM model for accuracy and response time in recognition and recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(2), 414–435.

Donkin, C., Brown, S., Heathcote, A., & Wagenmakers, E.-J. (2011). Diffusion versus linear ballistic accumulation: Different models but the same conclusions about psychological processes? *Psychonomic Bulletin & Review*, *55*, 140–151.

Dosher, B. A. (1984). Discriminating preexperimental (semantic) from learned (episodic) associations: A speed-accuracy study. *Cognitive Psychology*, *16*, 519–555.

Dosher, B. A., & Rosedale, G. (1989). Integrated retrieval cues as a mechanism for priming in retrieval from memory. *Journal of Experimental Psychology: General*, *118*(2), 191–211.

Dosher, B. A., & Rosedale, G. (1991). Judgments of semantic and episodic relatedness: Common time-course and failure of segregation. *Journal of Memory and Language*, *30*, 125–160.

Dosher, B. A., & Rosedale, G. (1997). Configural processing in memory retrieval: Multiple cues and ensemble representations. *Cognitive Psychology*, *33*, 209–265.

Edwards, W. (1965). Optimal strategies for seeking information: Models for statistics, choice reaction times, and human information processing. *Journal of Mathematical Psychology*, *2*, 312–329.

Feller, W. (1968). *An introduction to probability theory and its applications: Vol. I* (3rd ed.). New York: Wiley.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930–1955. In *Studies in linguistic analysis* (pp. 1–32). Oxford, England: Philological Society.

Freeman, E., Heathcote, A., Chalmers, K., & Hockley, W. (2010). Item effects in recognition memory for words. *Journal of Memory and Language*, *62*, 1–18.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–511.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*(1), 1–67.

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition*, *13*(1), 8–20.

Goldinger, S. D. (1998). Echoes of echoes? an episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279.

Gregg, V. H. (1976). Word frequency, recognition and recall. In J. Brown (Ed.), *Recall and recognition.* Oxford, England: John Wiley & Sons.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244.

Gronlund, S. D., & Ratcliff, R. (1989). Time course of item and associative information: Implications for global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(5), 846–858.

Günther, F., Dudschig, C., & Kaup, B. (2015). LSAfun—an R package for computations based on Latent Semantic Analysis. *Behavior Research Methods*, *47*, 930–944.

Healey, M. K., Crutchley, P., & Kahana, M. J. (2014). Individual differences in memory search and their relation to intelligence. *Journal of Experimental Psychology: General*, *143*(4), 1553–1569.

Healey, M. K., & Kahana, M. J. (2016). A four-component model of age-related memory change. *Psychological Review*, *123*(1), 23–69.

Hemmer, P., & Criss, A. H. (2013). The shape of things to come: Evaluating word frequency as a continuous variable in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1947–1952.

Hintzman, D. L. (1984). MINERVA2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, *16*(2), 96–101.

Hintzman, D. L. (1988). Judgements of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*(4), 528–551.

Hintzman, D. L. (2011). Research strategy in the study of memory: Fads, fallacies, and the search for the "coordinates of truth". *Perspectives on Psychological Science*, *6*(3), 253–271.

Hintzman, D. L., & Curran, T. (1997). Comparing retrieval dynamics in recognition memory and lexical decision. *Journal of Experimental Psychology: General*, *126*(3), 228–247.

Hockley, W. E., & Cristi, C. (1996). Tests of encoding tradeoffs between item and associative information. *Memory & Cognition*, *24*(2), 202–216.

Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, *45*, 718–730.

Hopper, W. J., & Huber, D. E. (2016). The primary and convergent retrieval model of recall. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th annual conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*, 269–299.

Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review*, *96*(2), 208–233.

Jacobs, A. M., & Grainger, J. (1994). Models of visual word recognition—sampling the state of the art. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(6), 1311–1334.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*, 513–541.

Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). New York, NY: Springer.

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.

Kass, R. E., Carlin, B. P., Gelman, A., & Neal, R. M. (1998). Markov chain Monte Carlo in practice: A roundtable discussion. *The American Statistician*, *52*(2), 93–100.

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*(3), 627–633.

Kılıç, A., Criss, A. H., & Howard, M. W. (2013). A causal contiguity effect that persists across time scales. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(1), 297–303.

Kılıç, A., Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2017). Models that allow us to perceive the world more accurately also allow us to remember past events more accurately via differentiation. *Cognitive Psychology*, *92*, 65–86.

Klein, K. A., Shiffrin, R. M., & Criss, A. H. (2007). Putting context in context. In J. S. Nairne (Ed.), *The foundations of remembering: Essays in honor of Henry L. Roediger III.* New York: Psychology Press.

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). London: Academic Press.

Kuçera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

Kumaran, D., & McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychological Review*, *119*(3), 573–616.

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.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284.

Larsen, R., & Warne, R. T. (2010). Estimating confidence intervals for eigenvalues in exploratory factor analysis. *Behavior Research Methods*, *42*(3), 871–876.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*, 1989–2001.

Link, S. W. (1975). The relative judgment theory of two choice response time. *Journal of Mathematical Psychology*, *12*, 114–135.

Link, S. W., & Heath, R. A. (1975). A sequential theory of psychological discrimination. *Psychometrika*, *40*, 77–105.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*(4), 492–527.

Lohnas, L. J., & Kahana, M. J. (2013). Parametric effects of word frequency in memory for mixed frequency lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1943–1946.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, and Computers*, *28*(2), 203–208.

Maddox, W. T., & Estes, W. K. (1997). Direct and indirect stimulus-frequency effects in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(3), 539–559.

Malmberg, K. J., Steyvers, M., Stevens, J. D., & Shiffrin, R. M. (2002). Feature frequency effects in recognition memory. *Memory & Cognition*, *30*(4), 607–613.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review*, *105*(4), 724–760.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, *88*, 375–405.

McGee, R. (1980). Imagery and recognition memory: The effects of relational organization. *Memory & Cognition*, *8*(5), 394–399.

McKeon, J. J. (1968). Rotation for maximum association between factors and tests. *Unpublished manuscript, Biometric Laboratory, George Washington University*.

Mensink, G., & Raaijmakers, J. G. W. (1988). A model for interference and forgetting. *Psychological Review*, *95*(4), 434–455.

Metcalfe Eich, J. (1982). A composite holographic associative recall model. *Psychological Review*, *89*(6), 627–661.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, *89*(3), 609–626.

Navarro, D. J., & Fuss, I. G. (2009). Fast and accurate calculations for first-passage times in Wiener diffusion models. *Journal of Mathematical Psychology*, *53*, 222–230.

Naveh-Benjamin, M. (2000). Adult age differences in memory performance: Tests of an associative deficit hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1170–1187.

Nelson, A. B., & Shiffrin, R. M. (2013). The co-evolution of knowledge and event memory. *Psychological Review*, *120*(2), 356–394.

Nobel, P. A., & Shiffrin, R. M. (2001). Retrieval processes in recognition and cued recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(2), 384–413.

Nosofsky, R. M., & Zaki, S. R. (2003). A hybrid-similarity exemplar model for predicting distinctiveness effects in perceptual old–new recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1194–1209.

Oberauer, K. (2005). Binding and inhibition in working memory: Individual and age differences in short-term recognition. *Journal of Experimental Psychology: General*, *134*(3), 368–387.

Paivio, A. (1969). Mental imagery in associative learning and memory. *Psychological Review*, *76*(3), 241–263.

Pexman, P. M., Hargreaves, I. S., Siakaluk, P. D., Bodner, G. E., & Pope, J. (2008). There are many ways to be rich: Effects of three measures of semantic richness on visual word recognition. *Psychonomic Bulletin & Review*, *15*(1), 161–167.

Plate, T. A. (2003). *Holographic reduced representations*. Stanford, CA: CSLI Publications.

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.

Raaijmakers, J. G. W. (2003). Spacing and repetition effects in human memory: application of the SAM model. *Cognitive Science*, *27*, 431–452.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*(2), 93–134.

Rae, B., Heathcote, A., Donkin, C., Averell, L., & Brown, S. (2014). The hare and the tortoise: Emphasizing speed can change the evidence used to make decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(5), 1226–1243.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review*, *111*(1), 159–182.

Ratcliff, R., Thapar, A., & McKoon, G. (2011). Effects of aging and IQ on item and associative memory. *Journal of Experimental Psychology: General*, *140*(3), 464–487.

Riordan, B., & Jones, M. N. (2010). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, *3*(2), 1–43.

Roediger, H. L., & Schmidt, S. R. (1980). Output interference in the recall of categorized and paired-associate lists. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(1), 91–105.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573–605.

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(1), 1–17.

Schooler, L. J., Shiffrin, R. M., & Raaijmakers, J. G. W. (2001). A Bayesian model for implicit effects in perceptual identification. *Psychological Review*, *108*(1), 257–272.

Schwartz, G., Howard, M. W., Jing, B., & Kahana, M. J. (2005). Shadows of the past: Temporal retrieval effects in recognition memory. *Psychological Science*, *16*(11), 898–904.

Sederberg, P. B., Howard, M. W., & Kahana, M. J. (2008). A context-based theory of recency and contiguity in free recall. *Psychological Review*, *115*(4), 893–912.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(2), 145–166.

Siedlecki, K. L. (2007). Investigating the structure and age invariance of episodic memory across the adult lifespan. *Psychology and Aging*, *22*(2), 251–268.

Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, *81*(3), 214–241.

Smith, L. B., Suanda, S. H., & Yu, C. (2014). The unrealized promise of infant statistical word-referent learning. *Trends in Cognitive Sciences*, *18*(5), 251–258.

Smith, L. B., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*, 1558–1568.

Starns, J. J., & Ratcliff, R. (2014). Validating the unequal-variance assumption in recognition memory using response time distributions instead of ROC functions: A diffusion model analysis. *Journal of Memory and Language*, *70*, 36–52.

Starns, J. J., Ratcliff, R., & McKoon, G. (2012). Evaluating the unequal-variance and dual-process explanations of zROC slopes with response time data and the diffusion model. *Cognitive Psychology*, *2012*(1), 1–34.

Steyvers, M., & Malmberg, K. J. (2003). The effect of normative context variability on recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(5), 760–766.

Stone, M. (1960). Models for choice-reaction time. *Psychometrika*, *25*, 251–260.

Tuerlincx, F. (2004). The efficient computation of the cumulative distribution and probability density functions in the diffusion model. *Behavior Research Methods, Instruments, & Computers*, *36*(4), 702–716.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology*, *26*(1), 1–12.

Underwood, B. J. (1969). Attributes of memory. *Psychological Review*, *76*(6), 559–573.

Underwood, B. J., Boruch, R. F., & Malmi, R. A. (1978). Composition of episodic memory. *Journal of Experimental Psychology: General*, *107*(4), 393–419.

Unsworth, N. (2010). On the division of working memory and long-term memory and their relation to intelligence: A latent variable approach. *Acta Psychologica*, *134*, 16–28.

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, *114*(1), 104–132.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550–592.

Vandekerckhove, J. (2014). A cognitive latent variable model for the simultaneous analysis of behavioral and personality data. *Journal of Mathematical Psychology*, *60*, 58–71.

Vickers, D. (1970). Evidence for an accumulator model of psychophysical discrimination. *Ergonomics*, *13*(1), 37–58.

Voskuilen, C., & Ratcliff, R. (2016). Modeling confidence and response time in associative recognition. *Journal of Memory and Language*, *86*, 60–96.

Wabersich, D., & Vandekerckhove, J. (2014). Extending JAGS: A tutorial on adding custom distributions to JAGS (with a diffusion model example). *Behavior Research Methods*, *46*(1),

15–28.

Wagenmakers, E.-J., Steyvers, M., Raaijmakers, J. G. W., Shiffrin, R. M., van Rijn, H., & Zeelenberg, R. (2004). A model for evidence accumulation in the lexical decision task. *Cognitive Psychology*, *48*, 332–367.

Yap, M. J., Tan, S. E., Pexman, P. M., & Hargreaves, I. S. (2011). Is more always better? effects of semantic richness on lexical decision, speeded pronunciation, and semantic classification. *Psychonomic Bulletin & Review*, *18*, 742–750.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971–979.

Yonelinas, A. P. (1997). Recognition memory ROCs for item and associative information: The contribution of recollection and familiarity. *Memory & Cognition*, *25*(6), 747–763.

Zacks, R. T., Hasher, L., & Li, K. Z. H. (2000). Human memory. In T. A. Salthouse & F. I. M. Craik (Eds.), *Handbook of aging and cognition* (2nd ed., pp. 293–357). Mahweh, NJ: Lawrence Erlbaum.

Zaromb, F. M., Howard, M. W., Dolan, E. D., Sirotin, Y. B., Tully, M., Wingfield, A., & Kahana, M. J. (2006). Temporal associations and prior-list intrusions in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(4), 792–804.

Appendix A
Method for rotating principal components

We found "factors" representing interpretable dimensions of variability by orthogonal rotation of the principal components of each correlation matrix, as described in the main text. Our goal was twofold: First, to make it easier to interpret the distributions of loadings of each parameter on each dimension (factor); second, to ensure that the structure of the correlations was not distorted during the rotation process.

The first goal was accomplished by selecting the "infomax" rotation criterion, which takes a given set of orthogonal loadings and finds a rotation of those loadings that jointly maximizes the information conveyed about each parameter by each factor *and* the information conveyed about each factor by each parameter (Browne, 2001; McKeon, 1968). Specifically, the infomax criterion to be minimized for a given $P \times M$ matrix of loadings $\mathbf{\Lambda} = [\lambda_{ij}]$ is given by

$$
\begin{aligned}
g(\mathbf{\Lambda}) = {} & \overbrace{\log M}^{\text{Upper bound}} \\
& - \overbrace{\sum_{j=1}^{M} \sum_{i=1}^{P} \frac{\lambda_{ij}^2}{\sum_{k=1}^{P} \lambda_{kj}^2} \log \left(\frac{\lambda_{ij}^2}{\sum_{k=1}^{P} \lambda_{kj}^2} \right)}^{\text{Simplicity of each factor}} \\
& + \overbrace{\sum_{i=1}^{P} \frac{\sum_{j=1}^{M} \lambda_{ij}^2}{\sum_{k=1}^{P} \sum_{l=1}^{M} \lambda_{kl}^2} \log \left(\frac{\sum_{j=1}^{M} \lambda_{ij}^2}{\sum_{k=1}^{P} \sum_{l=1}^{M} \lambda_{kl}^2} \right)}^{\text{Each parameter loads on equal number of factors}} \\
& + \overbrace{\sum_{j=1}^{M} \frac{\sum_{i=1}^{P} \lambda_{ij}^2}{\sum_{k=1}^{P} \sum_{l=1}^{M} \lambda_{kl}^2} \log \left(\frac{\sum_{i=1}^{P} \lambda_{ij}^2}{\sum_{k=1}^{P} \sum_{l=1}^{M} \lambda_{kl}^2} \right)}^{\text{Each factor accounts for equal number of parameters}}
\end{aligned}
$$

where we have labeled each term by which aspect of the loading matrix it quantifies. We seek to find a rotation matrix T that, when applied to the loading matrix $\mathbf{\Lambda}$, yields the minimum $g(\mathbf{\Lambda}T)$.

Because we are dealing not with a single set of principal component loadings, but with a large sample of such loadings, it would be inadvisable to find a separate rotation to each sample. The resulting distribution would be a nonlinear distortion of the original posterior distribution of loadings and would, therefore, not represent the structure of the correlations inferred from the data. Thus, to accomplish our second goal, we found a single rotation matrix T that optimized the *average* infomax criterion across all samples of principal component loadings. Because this single rotation T, once found, is applied uniformly across all samples, the resulting distribution of loadings (on factors) is a linear transformation of the original distribution of loadings (on principal components) and therefore does not distort the posterior.

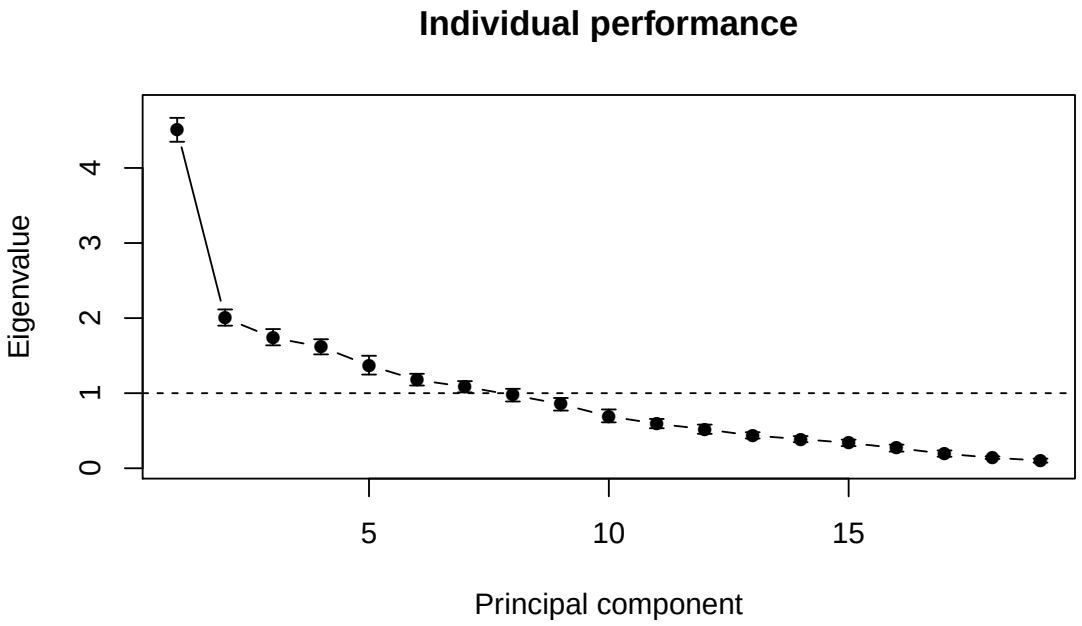


Figure B1. Posterior distribution over eigenvalues of the matrix of correlations between individual participant parameters. Bars depict 95% credible regions and points depict posterior means.

Appendix B
Analysis of correlations among all participant parameters

As in the analyses in the main text, we obtained the eigenvalues and eigenvectors of each sample of the full correlation matrix of individual parameters. Boundary separation, bias, and residual time parameters all entail transformations, as described in the main text. We computed correlations using the *untransformed* boundary separation and bias parameters, since they were originally estimated on this untransformed scale (logarithmic for boundary separation and logistic for bias). However, because the scale of the residual time parameters is determined by the minimum response time for each participant in each task, these were left on their transformed scale (i.e., on the scale of seconds) when computing the correlation matrices.

As shown in Figure B1, there are eight eigenvalues that are credibly greater than or equal to one, indicating that the pattern of correlations among individual parameters can be satisfactorily accounted for by eight principal components which collectively account for a median of 76 percent (with a 95% credible interval of 75%–77%) of the total correlation among individuals. We again performed an orthogonal rotation on each sample of the loading matrix of these top eight principal components according to the "infomax" criterion to obtain a distribution of "factors", shown in Figure B2, that describe the essential correlational structure among the individual participant parameters.

Individual factors 1, 2, and 3 are quite similar to the same factors identified in the analysis in the main text, while the fourth factor identified in the main text ends up being

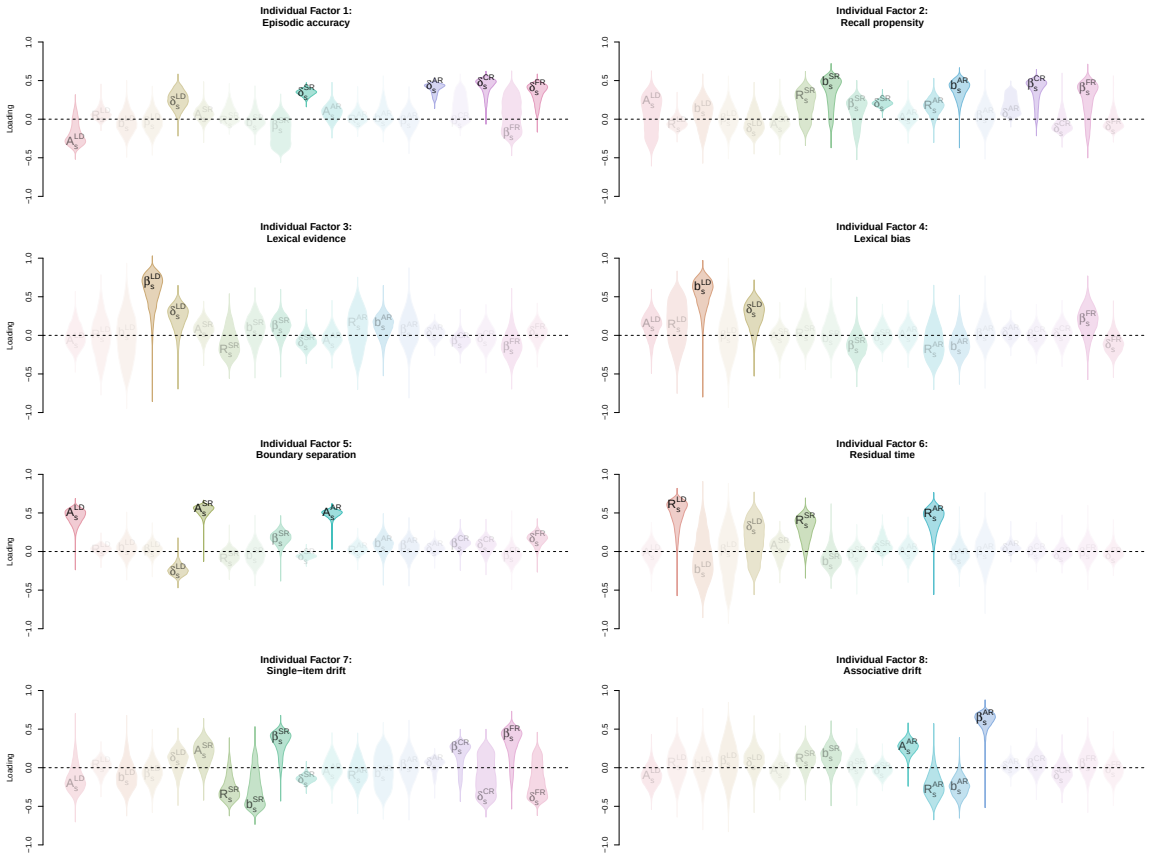


Figure B2. Posterior distributions of the loadings of each individual participant parameter (see Table 2) on factors formed by orthogonal rotation of the top eight principal components. As described in the main text, each factor was assigned a label on the basis of which parameters loaded most strongly on that factor. The saturation of each distribution and label visually indicate the magnitude and uncertainty of each loading.

split into two separate factors, numbered 7 and 8 below. This analysis also finds that boundary separation and residual time form their own factors, suggesting they reflect individual differences that are relatively stable across tasks.

Individual Factor 1: Episodic accuracy. This factor reflects the correlations among the accuracy-related parameters across all episodic memory tasks and their slight correlation with accuracy in lexical memory.

Individual Factor 2: Recall propensity. This factor accounts for rate of responding in both cued and free recall and illustrates their correlation with boundary asymmetry (a form of response bias) in both single-item and associative recognition.

Individual Factor 3: Lexical evidence. This factor accounts for both drift and accuracy in the evidence accumulated for the lexical decision task.

Individual Factor 4: Lexical bias. Separate from factor 3, this factor represents a correlation between lexical decision accuracy and response bias in lexical decision.

Individual Factor 5: Boundary separation. The total amount of evidence a participant needs before committing to a decision is correlated across tasks, indicating that

response caution is a relatively stable property of participants, particularly given that this correlation holds between different types of task, namely, lexical and episodic, which appear to involve different kinds of evidence.

Individual Factor 6: Residual time. As with boundary separation, residual time is correlated across tasks, even between lexical and episodic tasks. Because residual time includes such things as the time taken to execute a motor response, it is sensible that this should be preserved across tasks that use the same response procedures.

Individual Factor 7: Single-item drift. This factor reflects two apparent trade-offs, one between response boundaries and drift rate in single-item recognition and another between overall response rates and response accuracy in both cued and free recall. That said, there remains considerable uncertainty regarding the loadings on this factor (note the very broad distributions), so we hesitate to interpret it too closely.

Individual Factor 8: Associative drift. This factor identifies a trade-off between two sets of parameters in associative recognition: boundary separation and drift, on the one hand, versus residual time and bias on the other. While we again do not wish to over-interpret this factor, we note that in the model proposed by Cox and Shiffrin (2017), item information tends to be retrieved earlier than associative information, such that the information needed to distinguish between intact and rearranged pairs in AR is only available later in retrieval. As a result, accumulating evidence earlier—which would correspond to a shorter residual time prior to the onset of evidence accumulation—would yield more positive recognition evidence—higher drift rate—because the items in both intact and rearranged pairs had been studied. In order to achieve equivalent performance on the AR task (note that *accuracy* in AR does *not* load on this factor), participants would need to adjust their response boundaries to account for this early positive evidence by moving the "yes" response boundary upward (increased boundary separation and decreased bias).

Appendix C

Analysis of correlations among all item parameters, including pseudowords

As described in the main text, our primary analysis incorporated only the LD accumulation rate for words and excluded the rate for pseudowords. Here, we present a parallel analysis that includes both LD bias and accuracy parameters for items, thereby allowing pseudowords to impact the analysis. To anticipate, the only substantial difference from the analysis presented in the main text is that it becomes more difficult to interpret correlations between LD evidence parameters and normative characteristics of words. This is because many of these characteristics become meaningless when applied to pseudowords (e.g., a pseudoword has no concreteness). Nonetheless, we present the full analysis for completeness and to provide some evidence of the robustness of the factors identified in our primary analysis.

The distributions of pairwise correlations among all item parameters are shown in Figure C1. When the eigenvalues and eigenvectors of each sample of the correlation matrix are computed, four of the resulting components are associated with eigenvalues that are credibly greater than or equal to one, as shown in Figure C2, and these components collectively account for a median of 79% (95% CI of 76%–81%) of the correlations among item parameters. The factors formed by orthogonal rotation of these components (Figure C3) preserve factors for overall episodic memory quality and single-item recognition bias. The lexical decision parameters and cued recall rate parameter are now mixed into two factors, due to the uncertainty that arises from the tradeoff between lexical bias (β_i^{LD}, how word-like both a word and its pseudoword form appear) and lexical accuracy (δ_i^{LD}, how easy it is to distinguish between a word and its pseudoword counterpart). The negative correlation between β_i^{LD} and δ_i^{LD} is a function of the fact that a more word-like word, when distorted, tends to yield a more word-like pseudoword, resulting in high bias (high β_i^{LD}) but more difficulty discriminating between the word and its pseudoword counterpart (low δ_i^{LD}). As in the main text, we computed the Kendall's τ rank correlation between each word's score on these factors and its various normative characteristics, with the resulting distribution of correlations shown in Figure C4. The correlations with the two lexicality/cued recall factors differ from those computed as part of the primary analysis, because as mentioned above, these properties either lose meaning or have different interpretations between words and pseudowords.

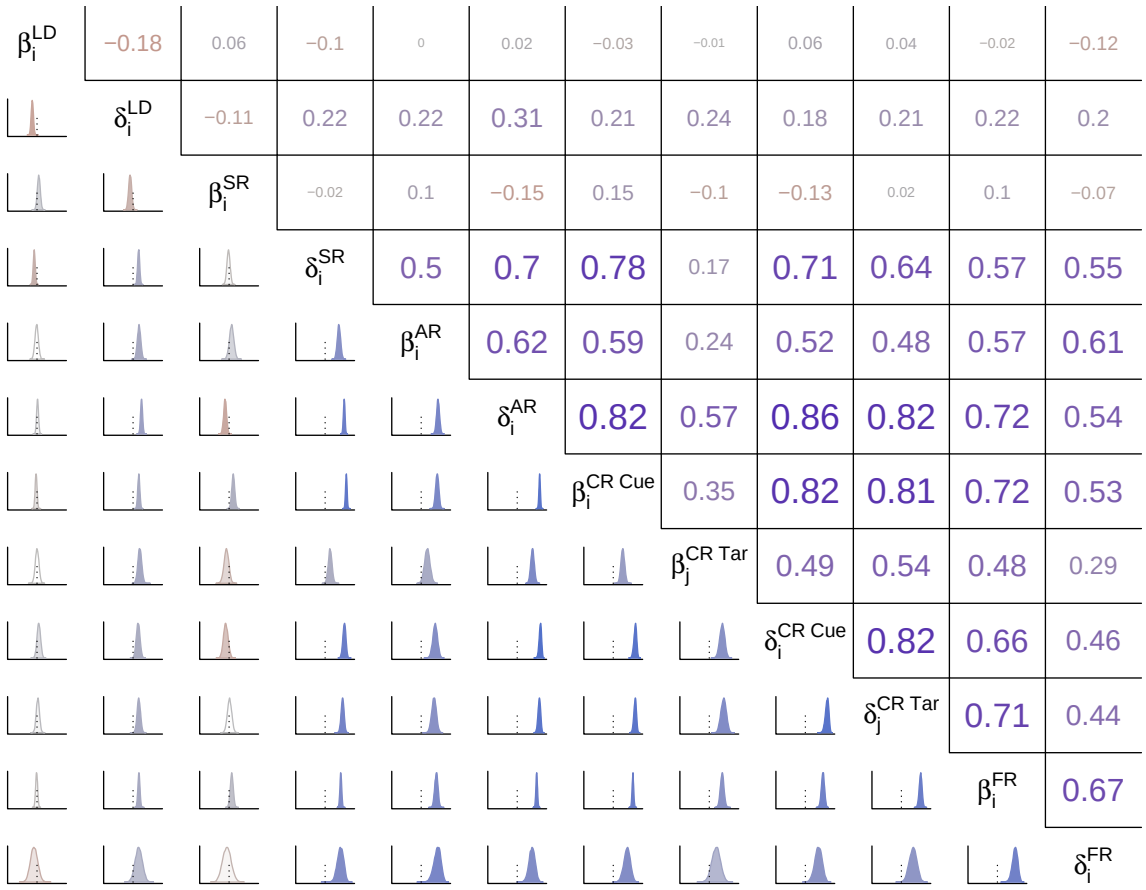


Figure C1. Posterior distributions over correlations between item parameters. Parameter names are given along the diagonal (see Table 3). The lower diagonal depicts the marginal posterior density of each pairwise correlation while the upper diagonal gives the posterior mode of each pairwise correlation. For visualization purposes, colors range between red (negative correlations) and blue (positive correlations) depending on the magnitude of the median correlation and the degree to which the densities in the lower right diagonal are filled reflects the width of the widest highest density interval that excludes zero (smaller for distributions that assign zero a high probability).

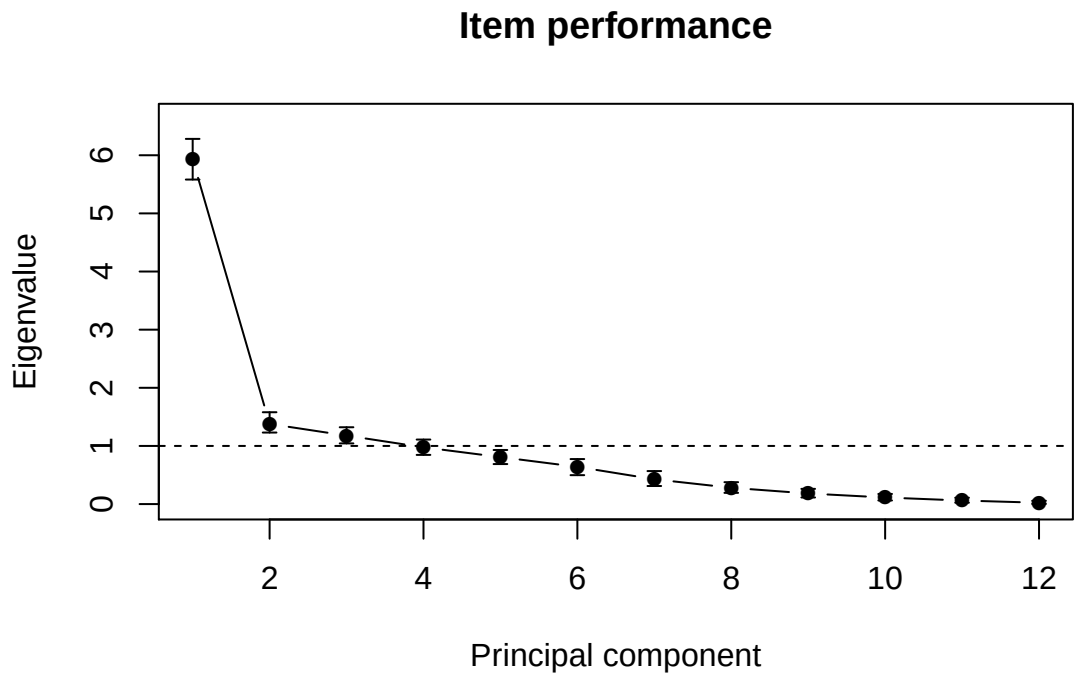


Figure C2. Posterior distribution over eigenvalues of the correlations between item parameters. Bars depict 95% credible regions and points depict posterior means.

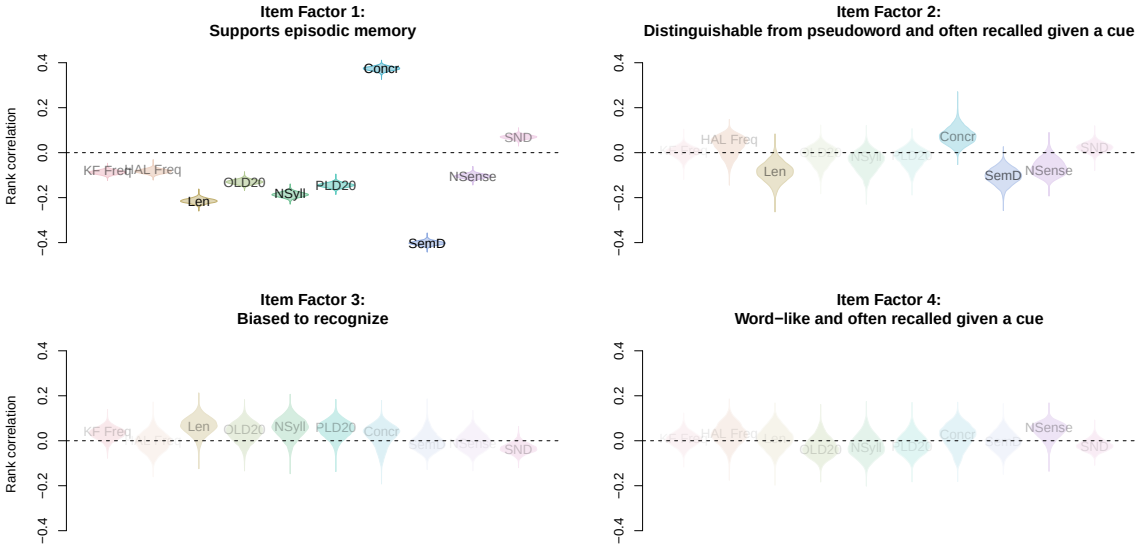


Figure C3. Posterior distributions of the loadings of each item parameter (see Table 3) on factors formed by orthogonal rotation of the top four principal components. As described in the main text, each factor was assigned a label on the basis of which parameters loaded most strongly on that factor. The saturation of each distribution and label visually indicate the magnitude and uncertainty of each loading.

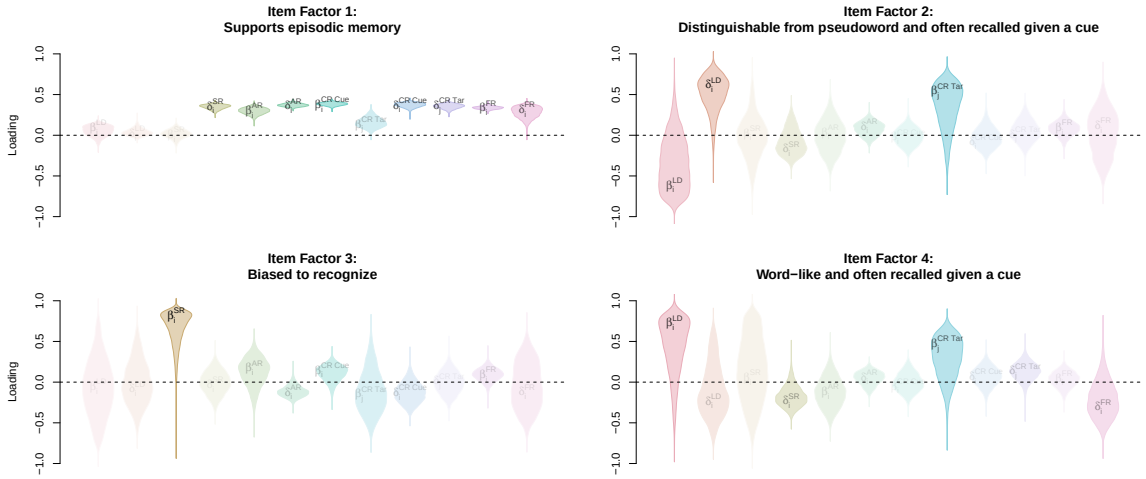


Figure C4. Posterior distributions over Kendall's τ rank correlations between each item's score on each factor (as depicted in Figure 16) and its normative lexical characteristics (as described in Table 1). For visualization purposes, the lightness of each distribution reflects how strongly it deviates from zero.

Appendix D
Robustness of exploratory correlation analyses

In this section, we explore the consequences of two important choices on the factors that emerge from our analyses: first, the choice of the number of principal components used to construct factors via orthogonal rotation; second, the choice of which tasks to include in the study. In each case, we are interested in whether different choices would have led to different conclusions. We show that, in general, the factors that emerge from these different choices have interpretations that are identical or very similar to those that we discuss in the main text, with any deviations being reasonable.

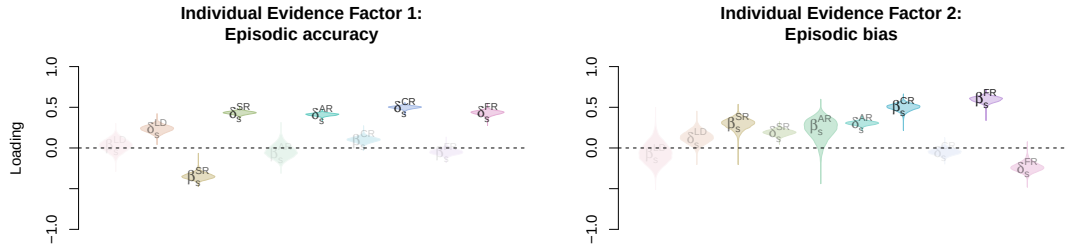
Number of components used for factor construction

In the analyses presented in the main text, we used rotated principal components to help understand the structure in the correlations among both item and individual parameters, as estimated from the data. By looking at factors formed by rotation of the principal components of each correlation matrix, we get a picture of the "most important" dimensions along which individuals and items may vary, where "importance" is indicated by the eigenvalues associated with each principal component (which govern the proportion of variability that each component can explain) and "most important" hinges on the analyst's choice of what counts as a sufficiently large eigenvalue. In other words, there is a degree of subjectivity in the choice of how many components deserve to be focused on: Some researchers might prefer a smaller number, which may not capture as much total variability but would restrict the focus to only the most prominent features of the correlation matrix. Other researchers might prefer a larger number, which would provide a more intricate picture of the correlation matrix at the cost of explanatory clarity (in the limit, of course, using *all* principal components would leave us back where we started, namely, the full matrix!). Coupled with our use of orthogonal rotation, different choices of how many components were "important" could lead different researchers to different conclusions.

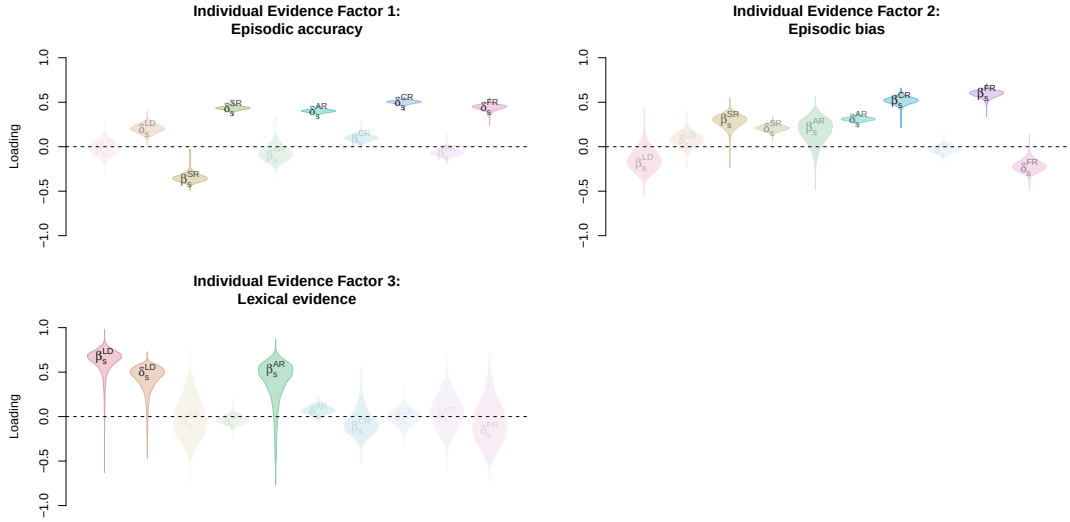
In the main text, we chose to focus on the components associated with eigenvalues greater than or equal to one. This is a commonly used criterion that says, in essence, that a component is "important" if it accounts for the variability in *at least* one parameter (the magnitude of an eigenvalue of principal component of a covariance/correlation matrix corresponds to the effective number of variables the variation of which is accounted for by that component). In the following, we demonstrate the consequences of selecting different numbers of components for individual participants' evidence parameters or item parameters, thereby demonstrating which dimensions remain consistent as factors are added or subtracted.

Individual evidence parameters. We present the results of applying our orthogonal rotation method based on several potential choices for the number of components used to describe the correlations among individual evidence parameters.

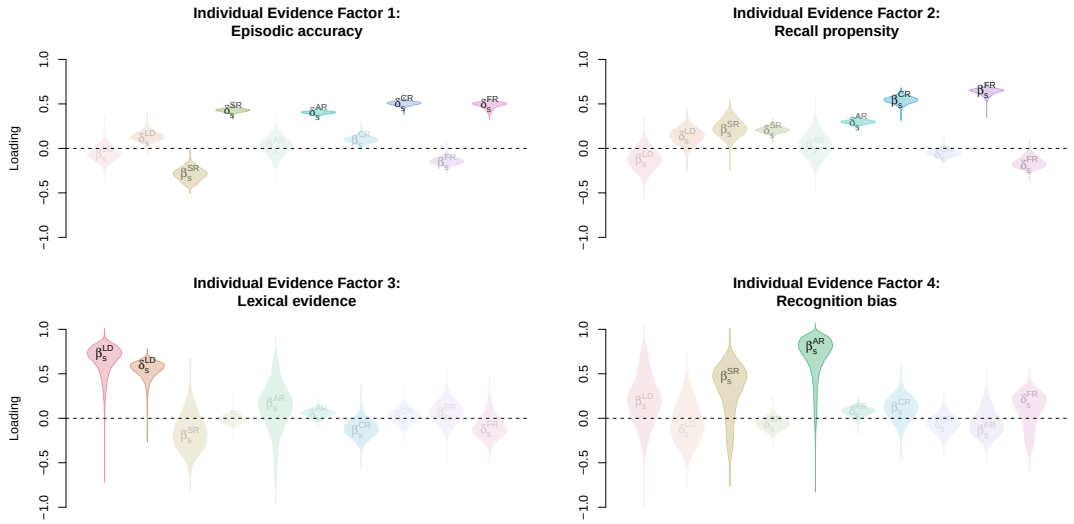
2 Factors These two factors are essentially the same as the first two identified in the main text, one relating to episodic accuracy (albeit with a larger loading for lexical accuracy) and another to the propensity to respond in recall tasks (which now involves a stronger loading of recognition drift parameters).



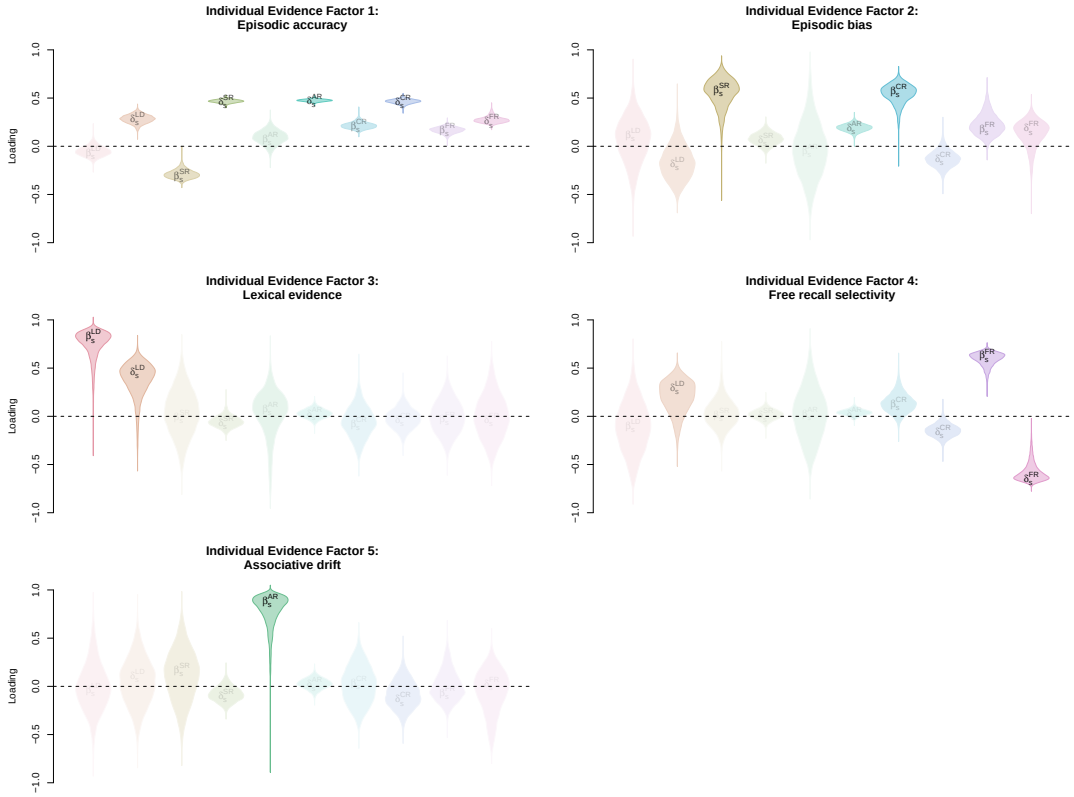
3 Factors Adding a third factor leaves the first two essentially unchanged, with the third component accounting for evidence (both drift and accuracy) in lexical decision as well as drift in associative recognition.



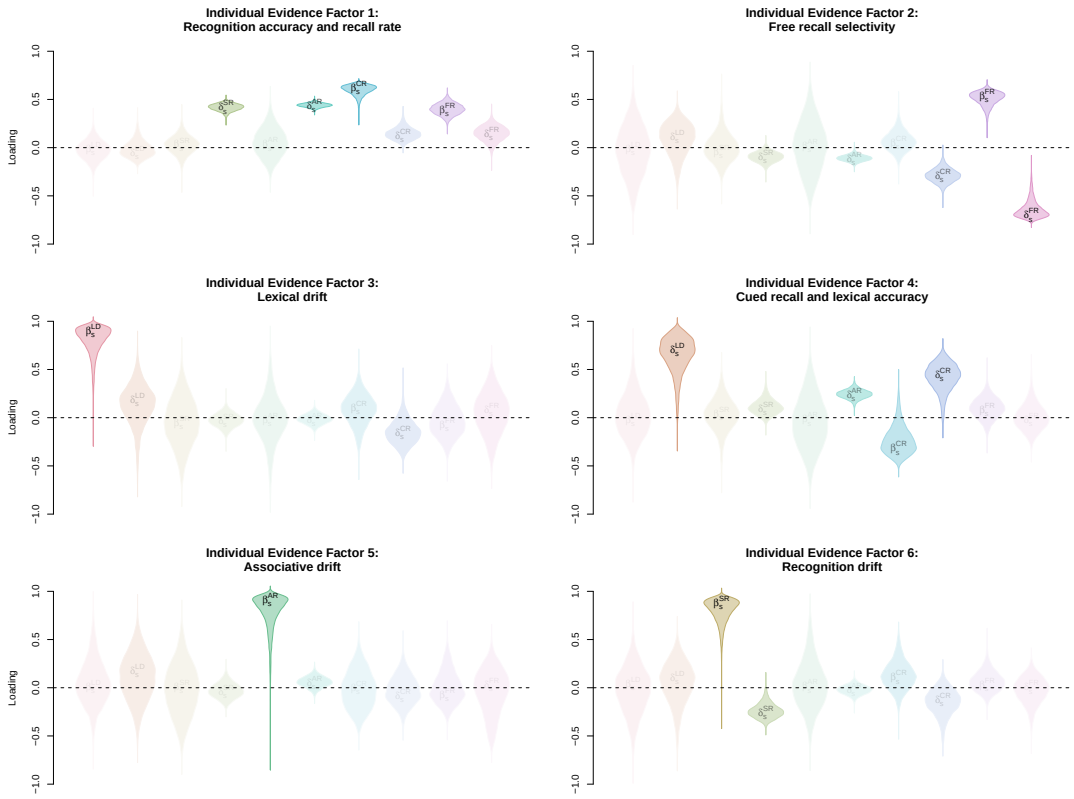
4 Factors This is the number we selected for the purposes of our main analyses. The first two factors again remain unchanged, with factor 3 now exclusively related to lexical decision and factor four accounting for drift in both single-item and associative recognition.



5 Factors Going from four to five factors, drift in AR now loads exclusively on its own factor (5) while drift in SR and the two lexical decision evidence parameters are split between two factors (3 and 4).



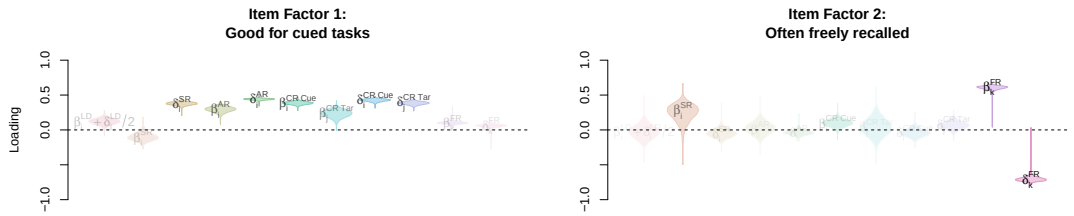
6 Factors The episodic accuracy and recall propensity factors (1 and 2) persist, with the remaining factors now devoted entirely to single parameters, one for each evidence drift parameter in LD, SR, and AR and one for accuracy in LD.



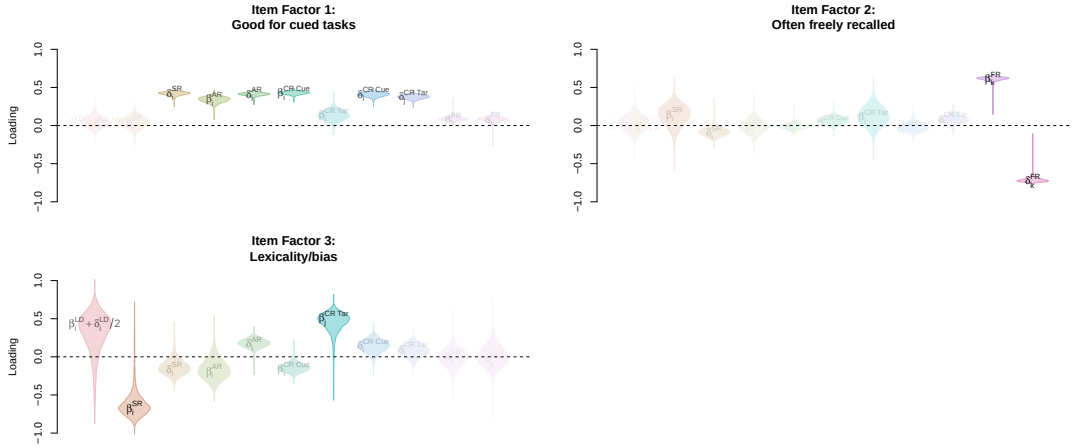
It is clear that different choices of the number of components used to describe correlations among different evidence parameters lead to factors with similar interpretations, with those factors generally splitting up as more components are added. Factors corresponding to episodic accuracy and recall propensity were found throughout, as was a distinction between lexical and episodic task parameters. The drift-related bias parameters split off into their own factors as more components were added, suggesting that these drift/bias parameters play less of a role in defining the structure of the correlations among individual evidence parameters than do accuracy-related parameters.

Item parameters. We now conduct a similar exercise with regard to the item parameters, where we eschew evidence rates associated with pseudowords in lexical decision, as in the main text.

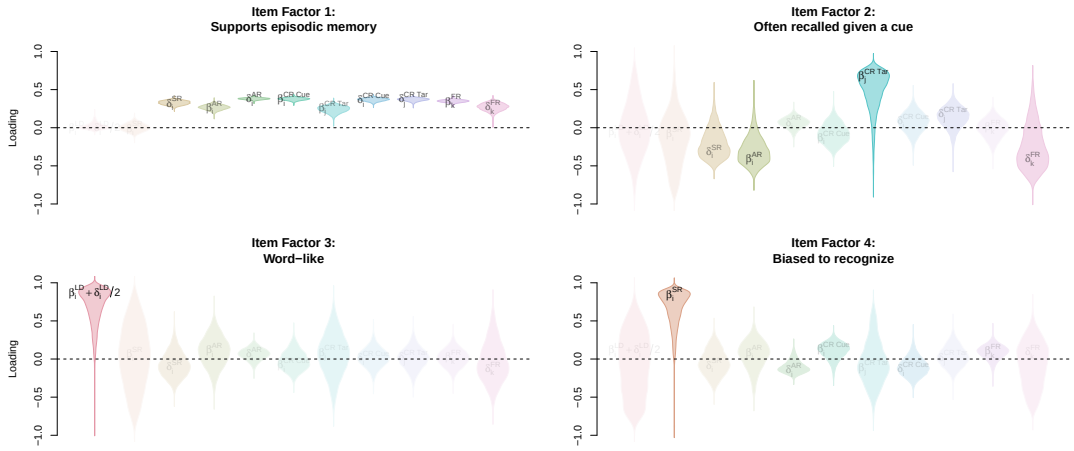
2 Factors The first factor corresponds to the first factor identified in the main text, namely, one that relates most of the parameters in episodic tasks, including all episodic accuracy parameters. The second factor acts as a "grab-bag" for the remaining parameters, including lexical evidence, item recognition bias, and response propensity in cued recall.



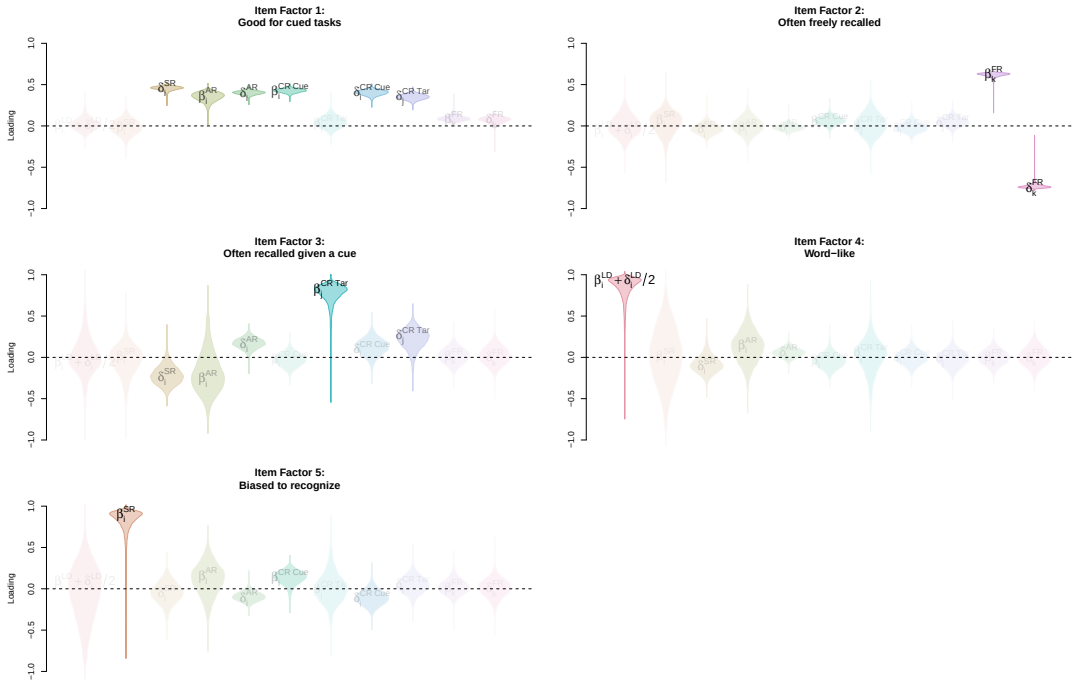
3 Factors While the first factor remains unchanged, the second has now split into two factors, one accounting for lexical evidence and cued recall response rate and another primarily accounting for item recognition bias.



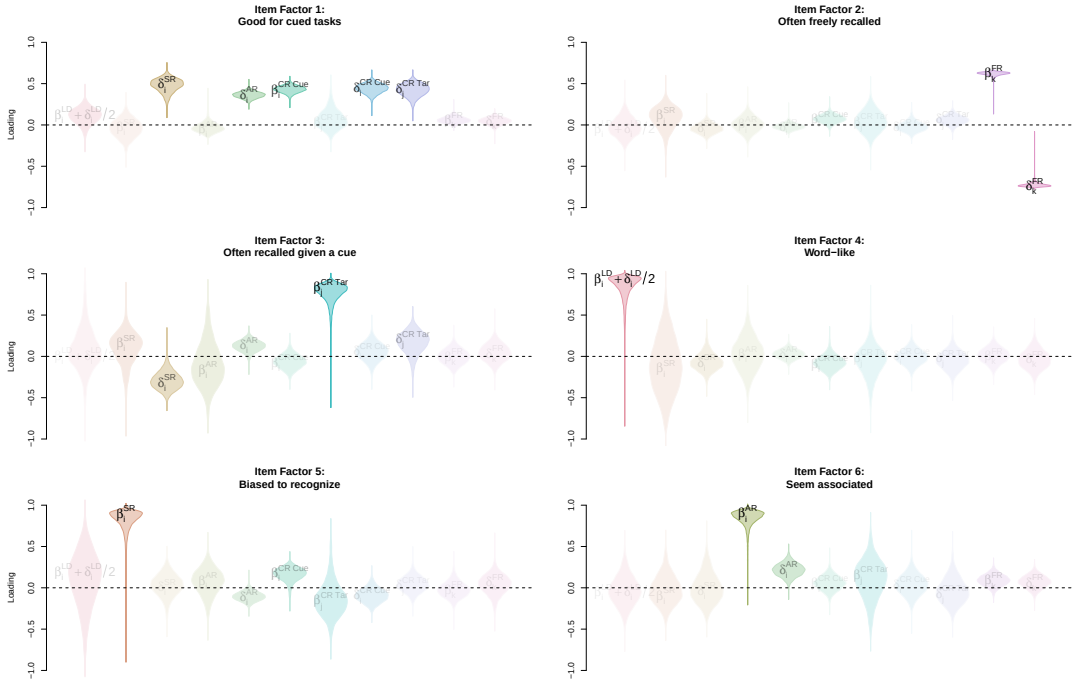
4 Factors This is the number chosen for the analysis in the main text. The main effect of going from three to four components is to split lexical evidence into its own parameter, separate from either cued recall response rate or SR bias, which are accounted for by their own respective factors.



5 Factors This is the point at which the first factor (related to episodic tasks in general) begins to break apart, this time into a factor for cued episodic tasks (recognition and cued recall) and one for free recall (along with bias in associative recognition).



6 Factors Going from five to six factors sees parameters related to free recall in their own factor, now separated from drift in associative recognition.



The dimensions describing the correlations among item parameters remain consistent even with different numbers of dimensions. As with parameters describing individual memory evidence, factors corresponding to episodic quality, recall rate, and lexical evidence are

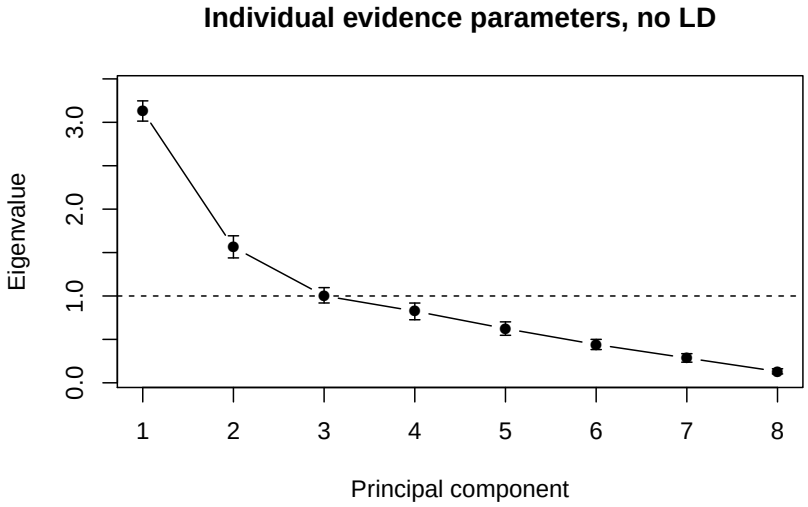
apparent across a wide range of choices for number of components, while several bias/drift parameters do not have a consistent pattern of loadings, suggesting that they play a less important role in defining the structure of the correlations among item parameters.

Excluding tasks from analysis

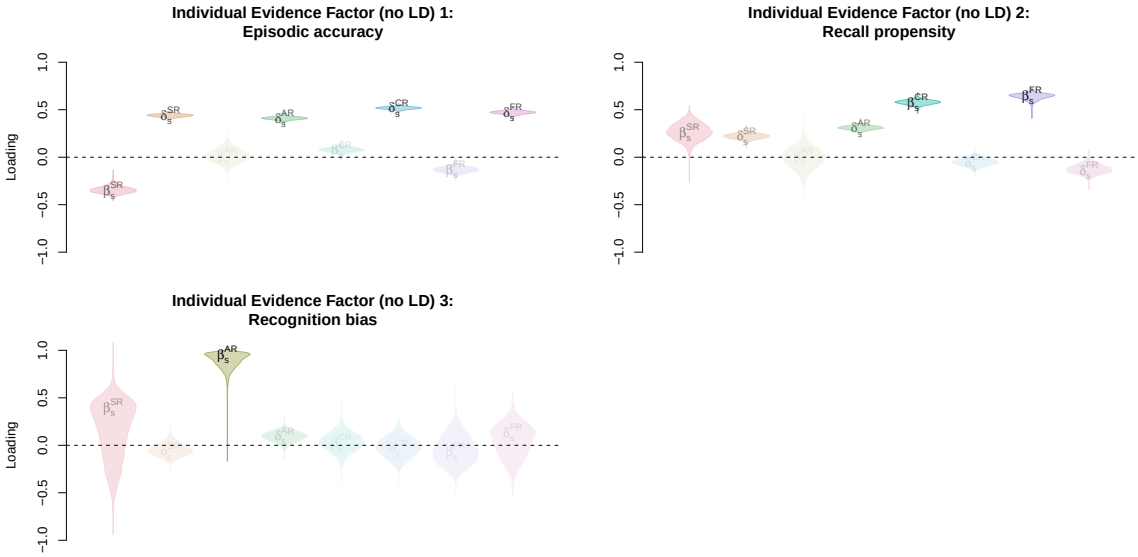
In addition to the choice of number of components used for factor construction, there is the design choice of which tasks to include in the experiment and subsequent analyses. While we cannot (yet) assess the consequences of *adding* tasks to the experiment, we can consider whether the same factors would have emerged had we excluded particular tasks. For example, we found that lexical decision (LD) tended to lie on a separate factor for both items and individuals; would this factor then disappear if this tasks were excluded?

Individual evidence parameters. Excluding parameters related to evidence in lexical decision results in 3 components with eigenvalues credibly greater than or equal to one, as shown in Figure D1a. After applying orthogonal infomax rotation to these components, as described above, the distribution of loadings on these three factors (Figure D1b) is similar to that reported in the main text, with factors corresponding to episodic accuracy, recall rate, and bias in recognition (particularly AR). Given the importance of bias in AR to the recognition bias factor, excluding parameters related to associative recognition results in 3 components with eigenvalues credibly greater than or equal to one (Figure D2a) which when rotated yield a distribution of parameter loadings (Figure D2b) that reflect factors for episodic accuracy, recall rate, and lexical evidence, but no separate recognition bias factor.

These results are in contrast to when SR-, CR-, or FR-related parameters are excluded. In each case, there are *four* components with eigenvalues credibly greater than or equal to one (Figures D3a, D4a, and D5a). The resulting factors are generally similar to those found when all tasks are included: When single-item recognition is excluded, the recognition bias factor only includes a loading for bias in associative recognition (since the other recognition task was excluded; Figure D3b). When cued recall is excluded, the recall rate factor now emphasizes a trade-off in free recall between overall response rate and the correctness of recall responses (Figure D4b). Finally, when free recall is excluded, an interesting pattern emerges in which response rate in cued recall loads both with other episodic accuracy parameters as well as bias in SR. This is notable in that it suggests that, while the best predictor of recall rate in one task is recall rate in another (e.g., β_s^{CR}), the tendency to recognize an item as having been studied (β_s^{SR}) also provides information about how likely one is to produce a response when given a cue.

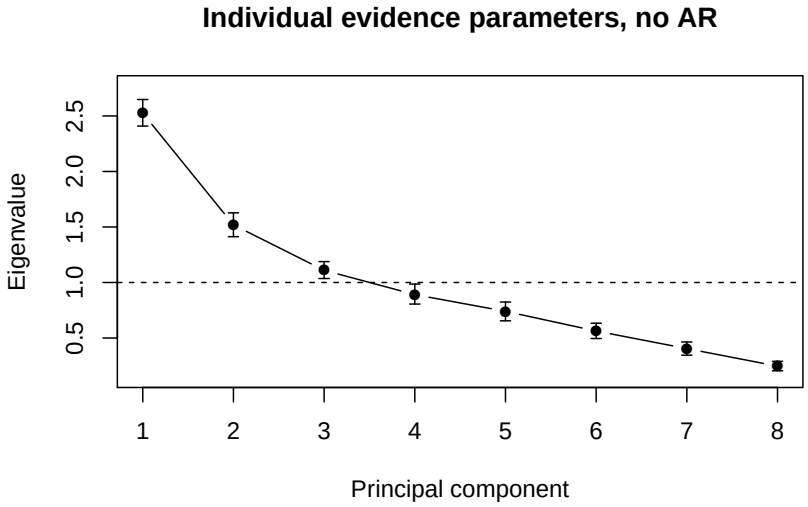


(a) Posterior distribution over eigenvalues of the matrix of correlations. Bars depict 95% credible regions and points depict posterior means.

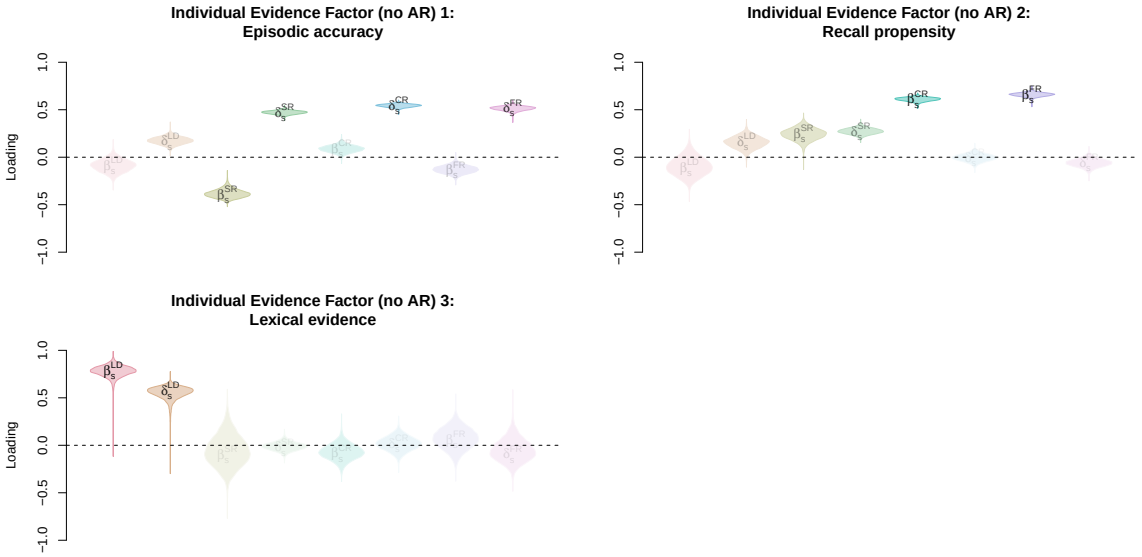


(b) Posterior distributions of loadings on factors formed by orthogonal rotation of the top three principal components. The saturation of each distribution and label visually indicate the magnitude and uncertainty of each loading.

Figure D1. Analysis of patterns of correlation among individual participant evidence parameters (β's and δ's; see Table 2), excluding those related to lexical decision.

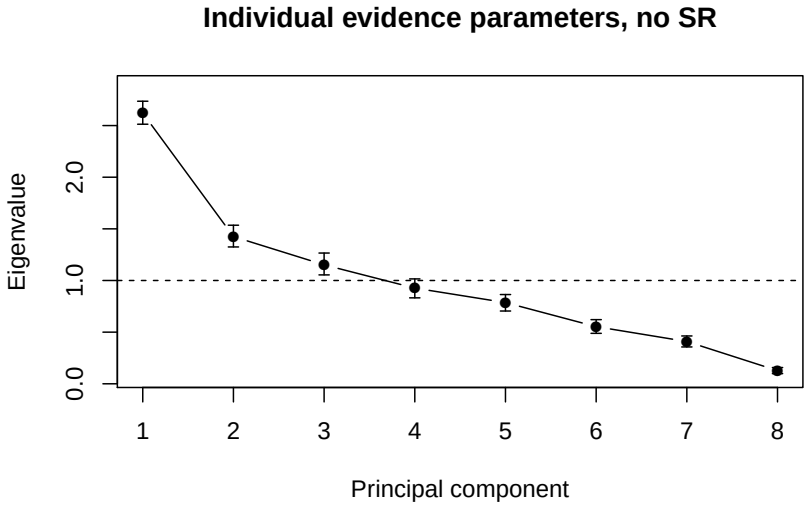


(a) Posterior distribution over eigenvalues of the matrix of correlations. Bars depict 95% credible regions and points depict posterior means.

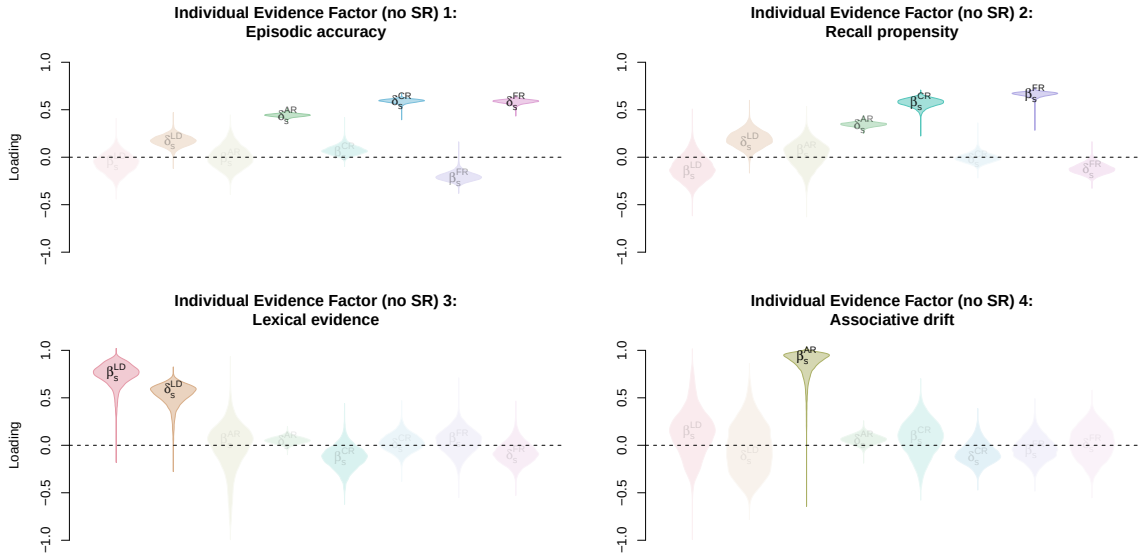


(b) Posterior distributions of loadings on factors formed by orthogonal rotation of the top three principal components. The saturation of each distribution and label visually indicate the magnitude and uncertainty of each loading.

Figure D2. Analysis of patterns of correlation among individual participant evidence parameters (β's and δ's; see Table 2), excluding those related to associative recognition.

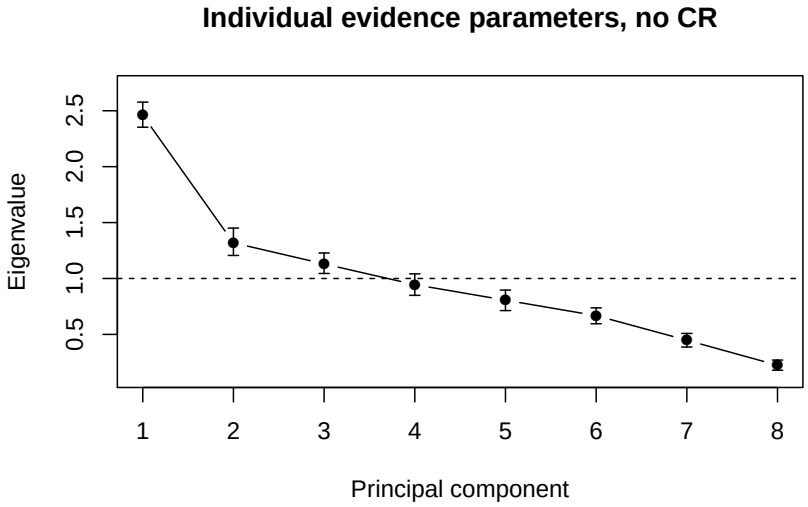


(a) Posterior distribution over eigenvalues of the matrix of correlations. Bars depict 95% credible regions and points depict posterior means.

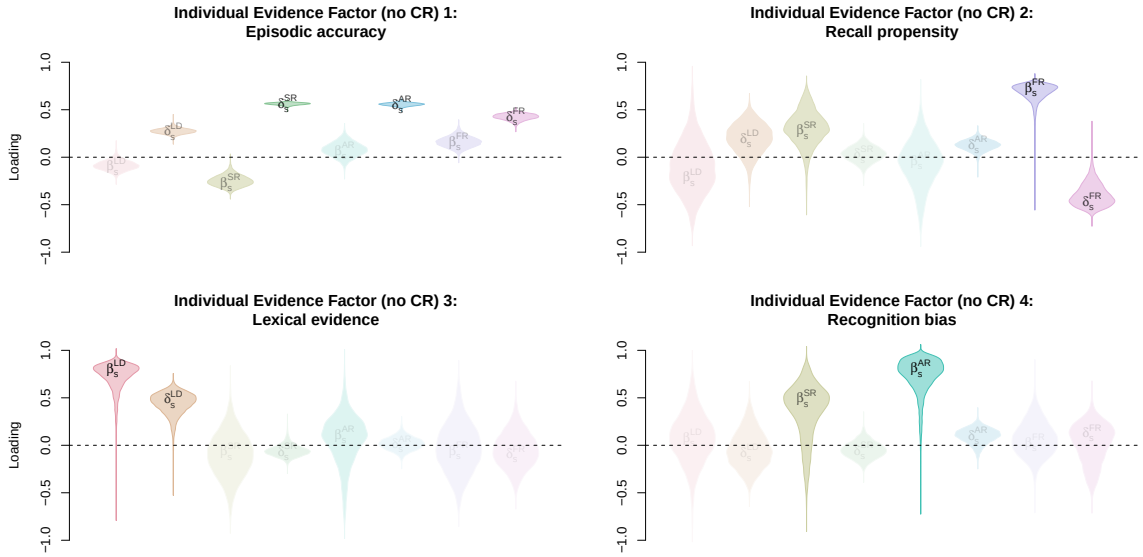


(b) Posterior distributions of loadings on factors formed by orthogonal rotation of the top three principal components. The saturation of each distribution and label visually indicate the magnitude and uncertainty of each loading.

Figure D3. Analysis of patterns of correlation among individual participant evidence parameters (β's and δ's; see Table 2), excluding those related to single-item recognition.

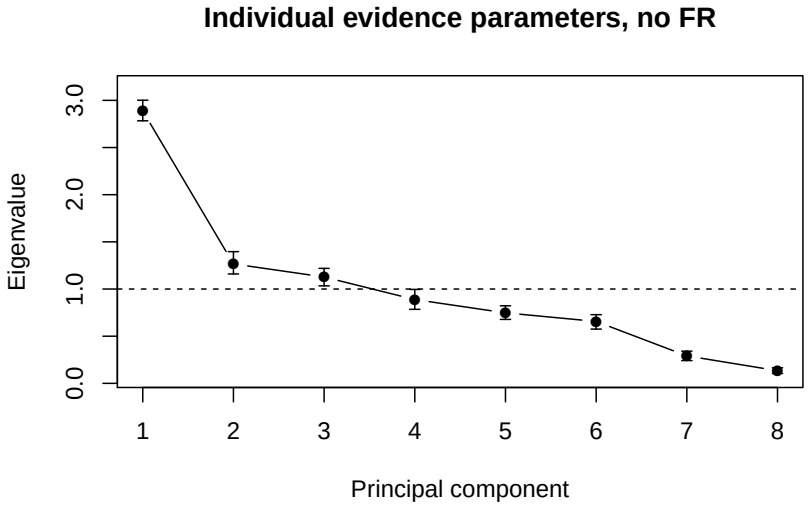


(a) Posterior distribution over eigenvalues of the matrix of correlations. Bars depict 95% credible regions and points depict posterior means.

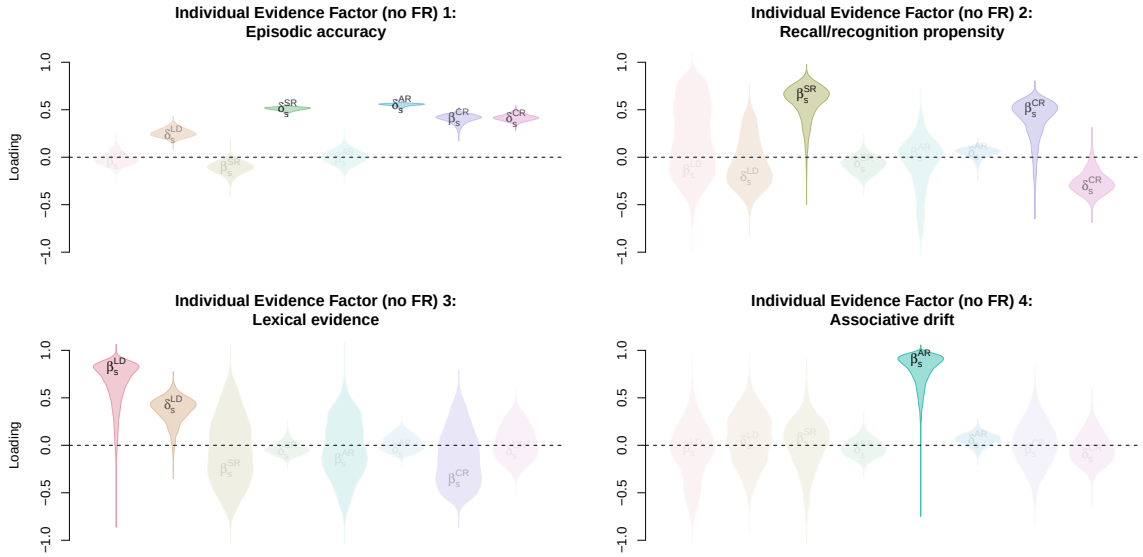


(b) Posterior distributions of loadings on factors formed by orthogonal rotation of the top three principal components. The saturation of each distribution and label visually indicate the magnitude and uncertainty of each loading.

Figure D4. Analysis of patterns of correlation among individual participant evidence parameters (β's and δ's; see Table 2), excluding those related to cued recall.



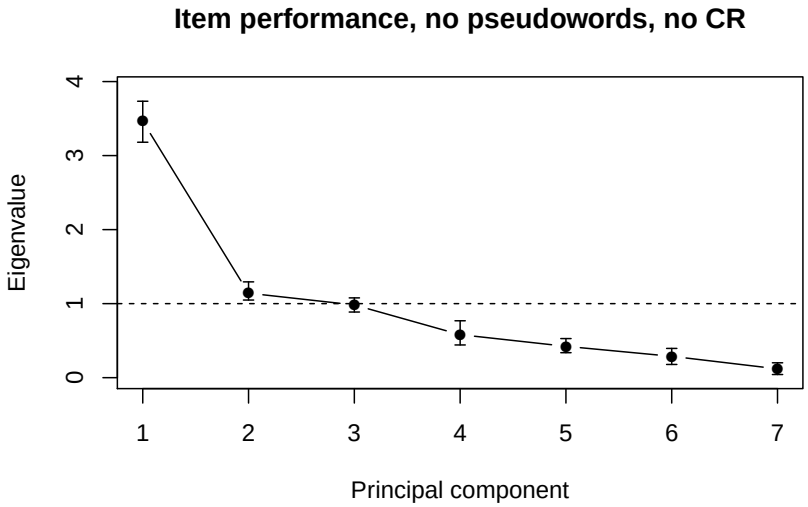
(a) Posterior distribution over eigenvalues of the matrix of correlations. Bars depict 95% credible regions and points depict posterior means.



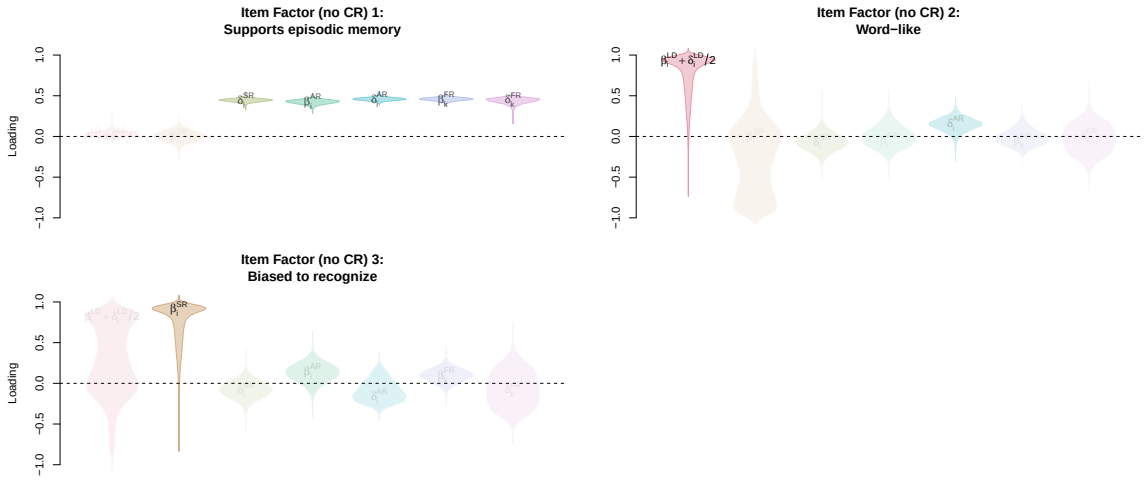
(b) Posterior distributions of loadings on factors formed by orthogonal rotation of the top three principal components. The saturation of each distribution and label visually indicate the magnitude and uncertainty of each loading.

Figure D5. Analysis of patterns of correlation among individual participant evidence parameters (β's and δ's; see Table 2), excluding those related to free recall.

Item parameters. Doing the same thing with item-level parameters (where, again, we exclude the pseudoword accumulation rates), we first note that there are three tasks for which one of their parameters defined a factor in the main analysis: cued recall (specifically, β_j^{CR}), lexical decision ($\beta_i^{LD} + \frac{\delta_i^{LD}}{2}$), and single-item recognition (β_i^{SR}). Therefore, we should expect that excluding any of these tasks would effectively eliminate their corresponding factor. As shown in Figures D6a, D7a, and D8a, eliminating either CR, LD, or SR results in a set of eigenvalues for which only three are credibly greater than or equal to one and, as expected, the remaining factors correspond to those from the main text that were not defined by a parameter from the excluded task (Figures D6b, D7b, and D8b). We note some important points: excluding LD leaves the remaining factors essentially unchanged, consistent with a distinction between the information relevant for episodic memory and that relevant to lexical access; excluding CR results in some residual uncertainty between the word-likeness factor and recognition bias factor; and excluding SR results in a recall rate factor that also includes some aspects of free recall accuracy and associative drift. This latter point echoes what happened for individuals when SR was excluded, namely, that other tasks "pick up the slack" when information from SR is unavailable, suggesting that SR performance is important for characterizing both items and individuals. In contrast, excluding parameters related to either associative recognition or free recall results in four components with eigenvalues greater than or equal to one (Figures D9a and D10a) which each yield factors (Figures D9b and D10b) with identical interpretations to those reported in the main text. In all, this provides a strong validation for the roles of each task in defining the structure of correlations within items.

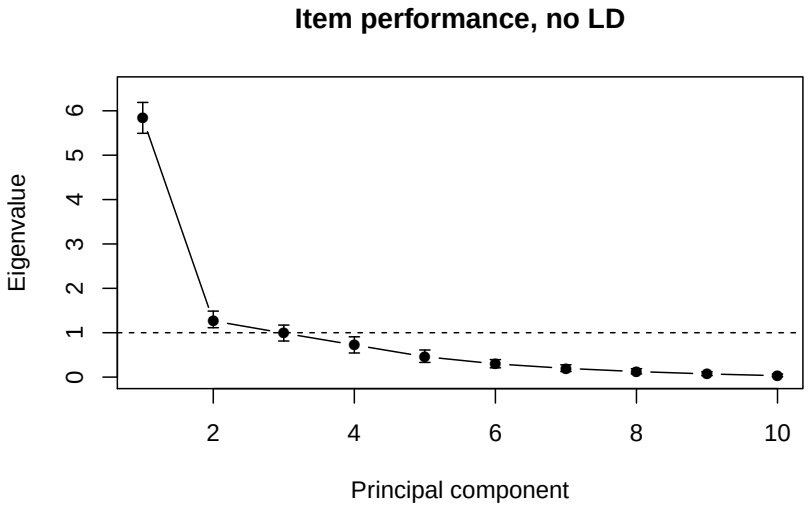


(a) Posterior distribution over eigenvalues of the matrix of correlations. Bars depict 95% credible regions and points depict posterior means.

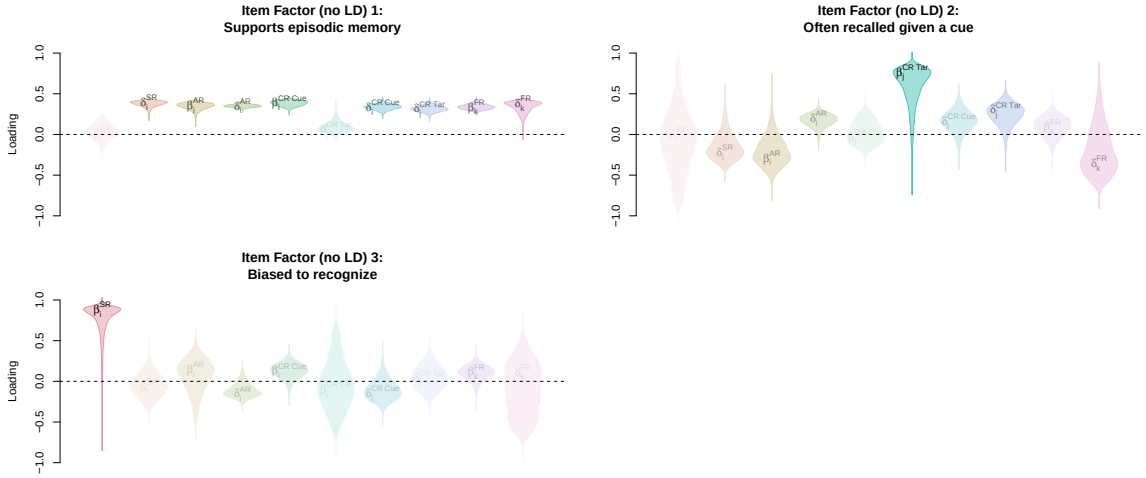


(b) Posterior distributions of loadings on factors formed by orthogonal rotation of the top three principal components. The saturation of each distribution and label visually indicate the magnitude and uncertainty of each loading.

Figure D6. Analysis of patterns of correlation among item parameters (see Table 3), excluding those related to cued recall.

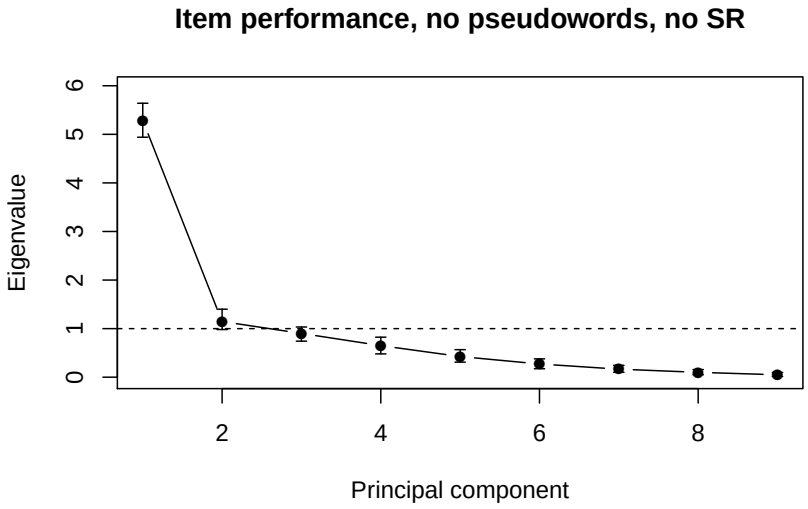


(a) Posterior distribution over eigenvalues of the matrix of correlations. Bars depict 95% credible regions and points depict posterior means.

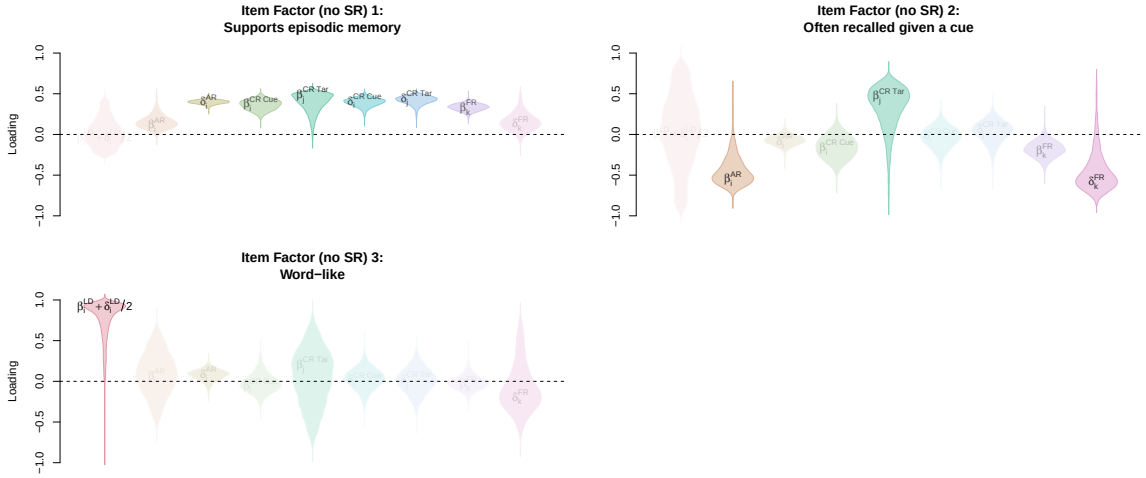


(b) Posterior distributions of loadings on factors formed by orthogonal rotation of the top three principal components. The saturation of each distribution and label visually indicate the magnitude and uncertainty of each loading.

Figure D7. Analysis of patterns of correlation among item parameters (see Table 3), excluding those related to lexical decision.

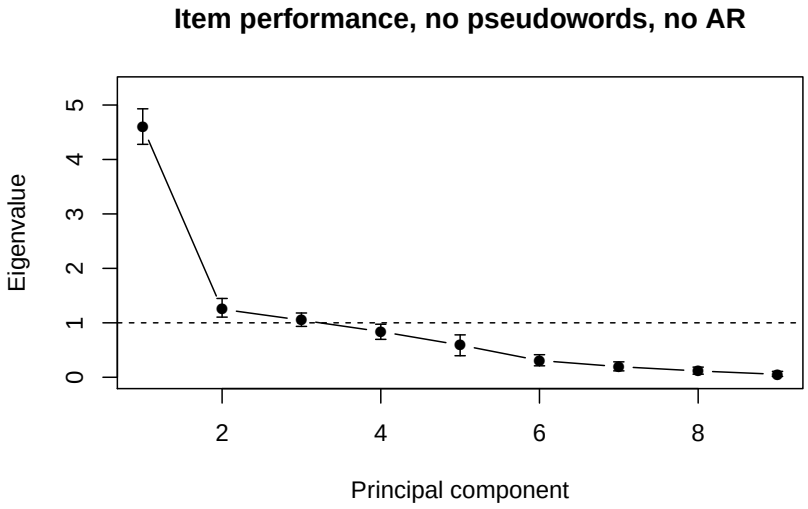


(a) Posterior distribution over eigenvalues of the matrix of correlations. Bars depict 95% credible regions and points depict posterior means.

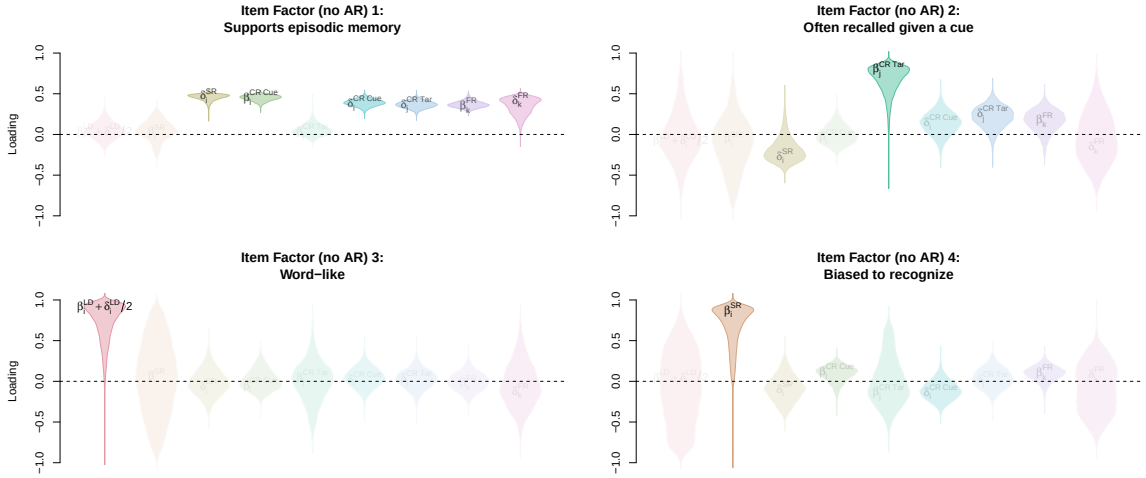


(b) Posterior distributions of loadings on factors formed by orthogonal rotation of the top three principal components. The saturation of each distribution and label visually indicate the magnitude and uncertainty of each loading.

Figure D8. Analysis of patterns of correlation among item parameters (see Table 3), excluding those related to single-item recognition.

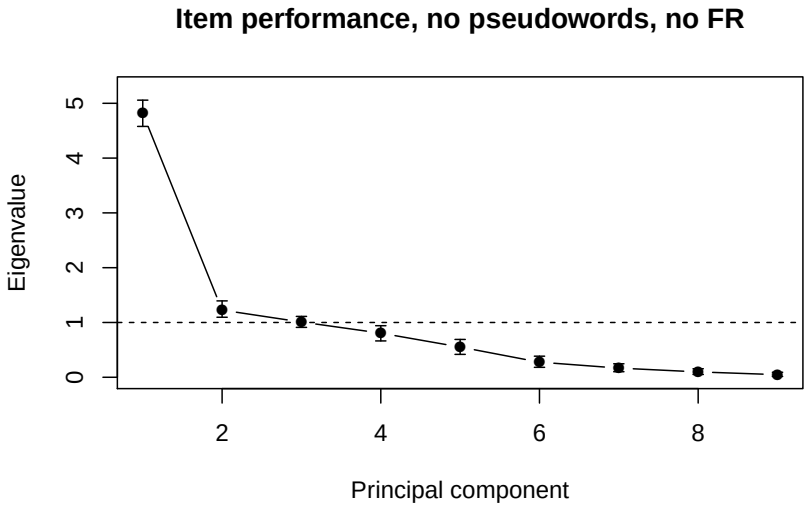


(a) Posterior distribution over eigenvalues of the matrix of correlations. Bars depict 95% credible regions and points depict posterior means.

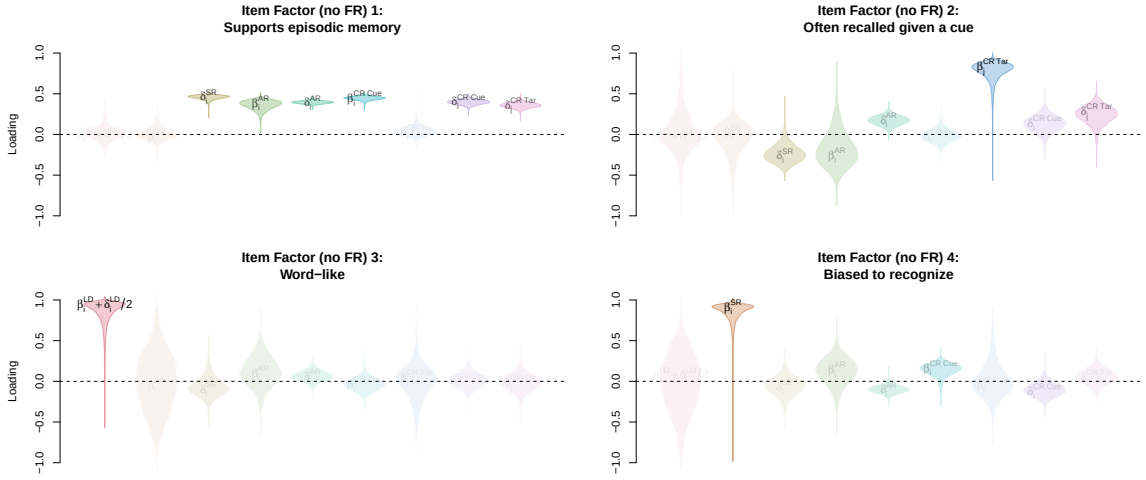


(b) Posterior distributions of loadings on factors formed by orthogonal rotation of the top three principal components. The saturation of each distribution and label visually indicate the magnitude and uncertainty of each loading.

Figure D9. Analysis of patterns of correlation among item parameters (see Table 3), excluding those related to associative recognition.



(a) Posterior distribution over eigenvalues of the matrix of correlations. Bars depict 95% credible regions and points depict posterior means.



(b) Posterior distributions of loadings on factors formed by orthogonal rotation of the top three principal components. The saturation of each distribution and label visually indicate the magnitude and uncertainty of each loading.

Figure D10. Analysis of patterns of correlation among item parameters (see Table 3), excluding those related to free recall.