Variability in Free versus Cued Recall

by

Eric Y. Mah

B.A. (Hons.), Kwantlen Polytechnic University, 2016

MSc, University of Victoria, 2020

A Dissertation Submitted in Partial Fulfilment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Psychology

University of Victoria

We acknowledge and respect the Lək̓ʷəŋən (Songhees and Esquimalt) Peoples on whose territory the university stands, and the Lək̓ʷəŋən and W̱SÁNEĆ Peoples whose historical relationships with the land continue to this day.

Supervisory Committee

Variability in Free versus Cued Recall

by

Eric Y. Mah

B.A. (Hons.), Kwantlen Polytechnic University, 2016

MSc, University of Victoria, 2020

Supervisory Committee

Dr. D. Stephen Lindsay, Department of Psychology

Supervisor

Dr. Adam Krawitz, Department of Psychology

Departmental Member

Dr. Farouk S. Nathoo, Department of Mathematics and Statistics

Outside Member

Abstract

Supervisory Committee

Dr. D. Stephen Lindsay, Department of Psychology

Supervisor

Dr. Adam Krawitz, Department of Psychology

Departmental Member

Dr. Farouk S. Nathoo, Department of Mathematics and Statistics

Outside Member

Two tasks that have been used extensively to study memory are free recall (FR; study a list of words and later attempt to recall as many as possible) and cued recall tasks (CR; study a list of randomly or meaningfully paired cue and target words and later attempt to recall targets given cues). A long tradition of fruitful research using these tasks has resulted in a host of effects that offer insight into human memory. Here we describe one such novel effect and its potential implications for theories of memory. The effect in question–which we refer to as the 'CR:FR variability effect'–is a surprising difference in inter-individual variability in performance between the two memory tasks. Specifically, in an initial experiment we observed greater individual differences in CR accuracy than in FR accuracy among individuals who performed both tasks. This result ran counter to our intuitions about the two memory tasks (e.g., one might expect that the lack of explicit retrieval structure in FR versus CR leaves more room for individual differences), and did not seem to be accounted for by popular formal models of memory (e.g., the *Search of Associative Memory* model; Raaijmakers & Shiffrin, 1981). The vast majority of research using these tasks has focused exclusively on differences in measures of central tendency, and we found no published research comparing individual differences in free and cued recall. Our research project

investigated the CR:FR variability effect and potential explanations for the effect via systematic and incremental manipulations to our stimuli and experimental designs. Specifically, across seven experiments we tested and replicated the effect using a general pool of 'average' English nouns (Experiment 1; *N* = 120 undergraduates), with a forced-recall procedure (Experiments 2A & 2B; *N* = 117 Prolific participants, *N* = 120 undergraduates), with meaningfully-related word pairs (Experiment 3; *N* = 260 Prolific participants), equating the study phases (Experiment 4; *N* = 360 Prolific participants), allowing participants self-paced study (Experiment 5; *N* = 120 undergraduates), and implementing serial recall for CR (Experiment 6; *N* = 211 undergraduates). Having ruled out primarily methodological explanations, we conducted a final experiment in which participants were instructed to use an imagery-based memory strategy previously shown to be effective (Thomas et al., 2023), with a pool of words conducive to imagery (Madan et al., 2010). In this experiment (Experiment 7; *N* = 208 Prolific participants), we did not observe the variability effect. We argue that individual differences in strategy use–due partly to task constraints in FR and CR, and partly to differential individual variability within different strategies–explain the CR:FR variability effect. We consider the implications of these findings for theories of memory and future memory research.

Table of Contents

List of Figures

Acknowledgments

I am profoundly grateful to my supervisor, Dr. Steve Lindsay, for his support & mentorship these past 6 years (time flies!). Your commitment to open, transparent, ethical, and generally good science, your tireless support for graduate and undergraduate students, and your unfailingly sharp insights into tough research problems have been a constant source of learning and inspiration (and aspirations!). It has been a pleasure and privilege working with you.

I would also like to thank the members of my committee–now Lindsay Lab committee veterans Drs. Adam Krawitz and Farouk Nathoo–both for their patience with me as I flitted from topic to topic, and for their invaluable feedback throughout the process. Thanks also go to Henry L. Roediger, Colleen M. Kelley, John Dunlosky, Larry Jacoby, Reed Hunt, and Roger Ratcliff for their helpful insights and suggestions very early on that helped motivate this research.

Finally, I am grateful to the UVic CaBS group for their feedback on various iterations of this work, and for generally fostering a sense of community. Similarly to my cohort of erstwhile classmates and current PGSC comrades–particularly Breanna, Yaewon, and Connor–whose solidarity and support did a great deal for my sanity.

Author Note

Portions of this work were previously published in: Mah, E. Y., & Lindsay, D. S. (2023). Variability across subjects in free recall versus cued recall. *Memory & Cognition*. **https://doi.org/10.3758/s13421-023-01440-4**. This work was supported by an NSERC Discovery Grant (#RGPIN-2016-03944) awarded to Dr. D. Stephen Lindsay.

Introduction & Background

Since the advent of experimental psychology, researchers have sought to understand human memory using tasks in which subjects study, and must later remember, lists of words (Cleary, 1982). Two tasks in particular, *free recall (FR)* and *cued recall (CR*, sometimes also called *paired-associates learning*), have provided valuable insights into memory phenomena. Typical studies of FR have participants study randomly selected related or unrelated words presented one at a time, with a subsequent test where they must recall as many as possible, in any order and without explicit retrieval cues. Such studies have revealed, for instance, that semantically related words tend to be recalled together (Cleary, 2018), that participants can develop false memories for words that were not studied but were related to studied ones (Roediger et al., 2001), and that memory tends to follow a *serial position curve*, where memory is best for earlier-studied items, decreases throughout a list, and improves again for items near the end of the list (Madigan, 1980).

Typical studies of CR, on the other hand, have participants study randomly-paired or meaningfully-related cue-target word pairs, presented one at a time, with targets to be recalled in response to cues at later test. CR involves aspects of both recognition (recognizing a cue) and recall (generating an item from memory; Wilson & Criss, 2017), and has primarily been used as a means of probing associative memory. CR experiments have revealed that providing cues tends to result in better memory performance than no-cue free recall, the strength of the semantic relationship between associates is a powerful determinant of later recall (Cleary, 2018), and that later-learned associations can interfere with earlier-learned associations (Postman & Underwood, 1973).

This long tradition of word memory research has resulted in a catalogue of effects, each purporting to offer a piece of the puzzle that is human remembering. I propose to add another effect to this catalogue. What follows will be a description of the novel and surprising

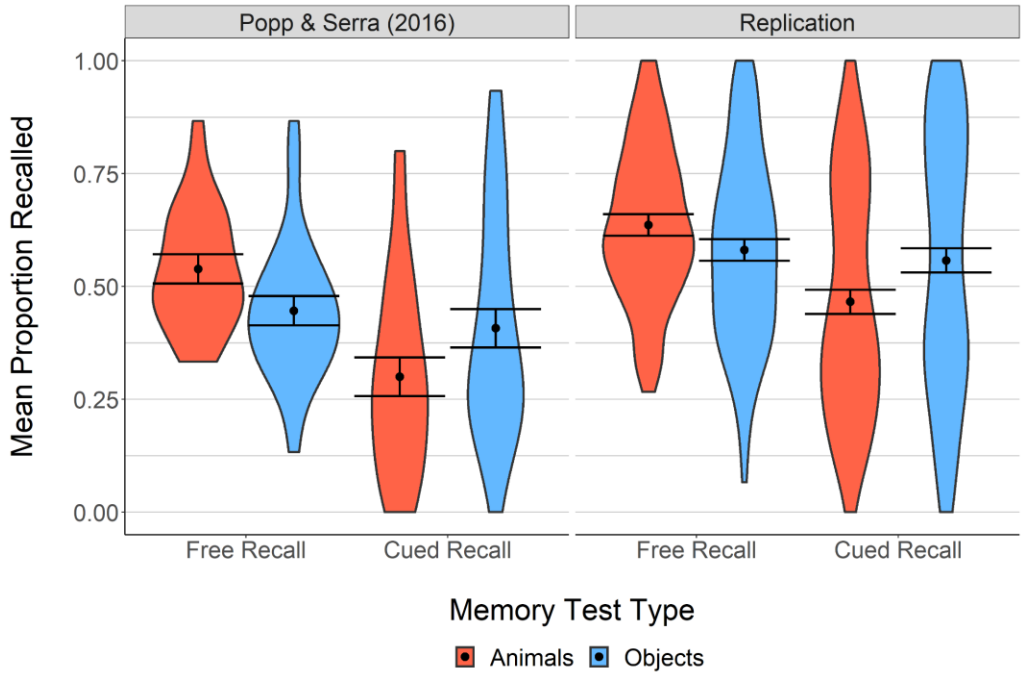
effect, relevant past work on FR/CR, a suite of empirical studies testing the robustness of and potential explanations for the effect, and the implications of the effect for our understanding of memory.

The CR:FR Variability Effect

We (Mah et al., 2023) replicated an experiment by Popp and Serra (2016) in which participants were tested on both FR and CR. Popp and Serra studied the relationship between word category (animals vs. objects) and memory task (FR vs. CR). They found better FR for animal names than for object names (an "animacy advantage," Nairne et al., 2017), but better CR for object names than for animal names (a reverse animacy effect). As shown in Figure 1, we replicated both of those findings.

Figure 1

FR and CR memory performance in Popp and Serra (2016) and replication data from Mah et al., 2023

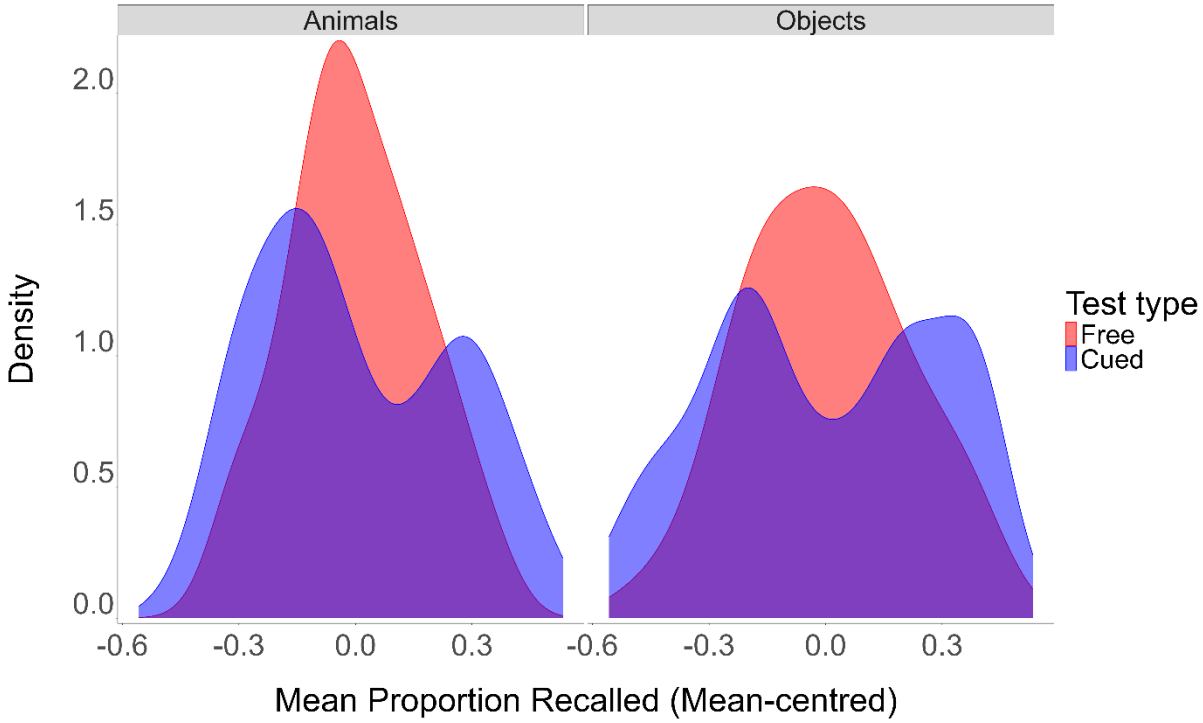


Note: Error bars in this figure are 95% confidence intervals for the within-subject comparison between animals and objects, as per Masson and Loftus (2003).

Looking at this figure, we were struck by the greater variability in CR scores than in FR scores, both in Popp and Serra (2016) and in our replication. Plotting the mean-centred distributions together revealed the dramatic nature of the difference in variability, and indeed in the shapes of the distributions:

Figure 2

Mean-centred overlapping FR and CR memory performance in Mah et al., 2023



Note. Proportion recalled mean-centred using means averaged within memory type and animacy.

Although a less common statistical analysis, there exist a handful of formal tests for comparing the equality of two variances. We opted to use the Pitman-Morgan test of equal variances for paired samples (Morgan, 1939; Pitman, 1939) [1], one of the most-commonly

[1] Tests the null hypothesis that, for two correlated random variables X_1 and X_2 (in this case, *FR proportion recalled* and *CR proportion recalled*), $U_1 = X_1 + X_2$ and $U_2 = X_1 - X_2$. If the null hypothesis of equivalent

used tests for comparing variability of two distributions, to formally confirm our observations of increased CR variability. Although some have reported that this test is sensitive to nonnormality (Wilcox, 2015), others have shown that the test is robust to nonnormality with normal or folded-normal distributions, with adequate power to detect differing variances (García-Pérez, 2013). This test indicated that variance in CR proportion correct was greater than variance in FR proportion correct for both animals (replication $p < .001$, original $p = .004$) and objects (replication $p < .001$, original $p = .004$).

Examinations of variability in cognitive performance (both within- and across-participants) have been applied fruitfully in other domains such as cognitive ageing, where researchers have found increasing variability with age and evidence of links between intra-individual variability and developmental outcomes (e.g., Christensen et al., 1999, LaPlume et al., 2021; Yao et al., 2016). However, we are not aware of research directly comparing variability in performance on standard FR and CR tests. Most studies involving both tasks are primarily concerned with average performance, with participants typically doing better on CR than FR (Cleary, 2018).

One of us (DSL) told several prominent memory researchers about this observation and asked them if they knew of prior research comparing inter-individual variability in accuracy on FR versus CR. Here are their personal communications in reply:

- Henry L Roediger wrote, "If you had asked me to guess beforehand which procedure was more variable, I would have guessed free recall. That task is … prone to various strategies, from forming a story with the words (depending on presentation rate) to rote rehearsal (and many others). Using Craik's logic, paired-associate learning provides more retrieval support (the stimulus or cue at test)

variances (i.e., $\sigma^2_1 - \sigma^2_2 = 0$) holds, then r (the correlation between U_1 and U_2 will be 0. The Pitman-Morgan test evaluates a test statistic derived from the sample size and r against a student's t distribution with $(n - 2)$ degrees of freedom (Mudholkar et al., 2003).

than does free recall (a blank computer screen). I would have thought that the additional retrieval support would have constrained variance."

- Colleen Kelley replied "I am actually surprised to think that there would be more strategies available for paired associates than free recall, wouldn't you think there are more constraints in paired associates, so less room for variation?"

- John Dunlosky responded, "If I hadn't read your note first, I'm pretty sure I would have predicted that individual differences in strategy use would contribute to larger individual differences in free recall than paired associate recall."

- Larry Jacoby wrote "I do not know of any data that shows a difference in variability between free recall and paired-associate learning."

- Similarly, Reed Hunt: "I am not aware of published research directly addressing your finding concerning variability…I cannot think of a specific theoretical approach that speaks to the result."

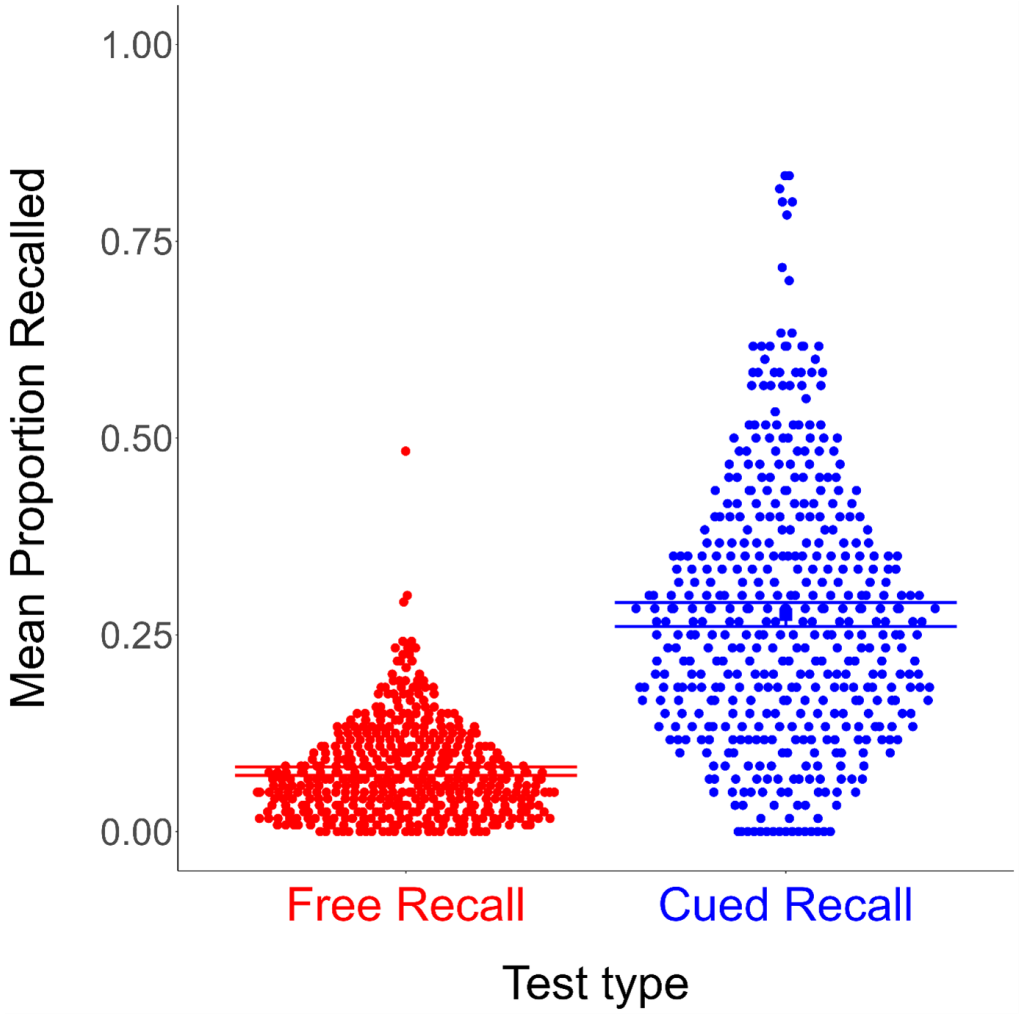
- Finally, Roger Ratcliff indicated interest in the finding. He pointed to Ratcliff et al. (2011), in which subjects were tested on numerous measures, including FR and CR. Looking at Figure 6 in that article, the distribution of scores in CR appears much larger than that in FR.

We searched without success for published experiments directly comparing variability in FR and CR performance. But we noticed some suggestive patterns from studies that

included both CR and FR tasks. Siedlecki (2007) tested adults on both tasks and found a descriptively higher variability in CR relative to FR (*SD* = 2.56 vs. 2.35). Cox et al. (2018) had participants complete CR and FR (along with three other memory tasks) while equating the study phases (all participants studied pairs, and were not told until afterwards whether they would be tested on FR of all words or CR of targets given cues). Although they did not compare variability across tasks their open data permitted a re-analysis and comparison:

Figure 3

FR and CR memory performance in Cox et al. (2018), by participant



Note. Error bars = 95% CIs (between-subjects). Points jittered horizontally based on relative frequency.

From a visual inspection and formal analysis (Pitman-Morgan $p < .001$) of Cox et al.'s (2018) data, these results show greater variability across individuals in CR relative to FR performance. However, the near-floor performance for FR in their data clouds interpretation. Specifically, low baseline FR performance restricts the lower-bound range, which in turn limits the variability (versus CR, which had much more room to vary). Still, the results of these prior studies hint at a pervasive and counterintuitive difference in inter-individual variability across tasks.

Aside from its counterintuitive-ness, why might this effect be of interest? One potential implication is that the mechanisms that determine cued recall performance vary more across individuals than do the mechanisms underlying free recall. Or, it could be that there are *more* processes underlying cued recall (e.g., recall + recognition, item memory + association memory) than free recall, and it is simply compounding variability in these additional processes that explain the difference. Some (e.g., Hockley & Cristi, 1996) have argued that there is a sharp distinction between the *item memory* (e.g., free recall items) and *association memory* (e.g., cued recall pairs), with potential trade-offs between the two. In tests of recognition memory, Hockley and Cristi (1996) found that when item memory was emphasized over association memory, associative memory performance decreased, but also that item memory did not suffer when association memory was emphasized. Applied to the CR:FR variability effect, it could be that association memory varies more across individuals than item memory. A striking difference in the variability and distribution of CR and FR might imply more qualitative differences between the tasks and underlying mechanisms than might be initially assumed. From a purely statistical standpoint, differential distributions in CR and FR performance may have implications for how data from these tasks is analyzed and compared, and could interact with other measures of interest (e.g., age; Schmidt et al., 1992).

Finally, if the variability difference is veridical, existing theories and models of recall must be able to account for it (and if they cannot, modifications must be made).

Potential explanations for the effect

Why might such a difference exist? There are a number of salient differences between FR and CR that are worth considering. The difficulty in evaluating any effect intimately tied to a particular paradigm is disentangling *methodological* explanations (e.g., stimulus set/task-specific effects) and *theoretical* explanations (e.g., mechanisms and processes posited by theories and models of memory). These two general categories informed our investigations of the variability difference. The next section provides a broad overview of findings and theories from these categories that are potentially germane to the novel CR:FR variability effect.

Methodological differences between FR and CR. It is not always possible to cleanly separate methodological and theoretical factors (i.e., design choices like list length ostensibly impact performance because they influence underlying theoretical mechanisms). For the purpose of interrogating the Mah et al. (2023) variability results, by methodological factors I refer mostly to design choices that were not theoretically motivated, and in some sense arbitrary. For instance, the number of studied words, stimulus presentation time, relationships among words, test order, etc. were not theoretically motivated but differed in some sense between FR and CR. In those experiments, words/pairs were presented for 5s, which means that participants had effectively half the time to study each word in a CR pair than they had to study each FR singleton. Study time is a strong predictor of later free and cued recall (de Jonge et al., 2012), and was not controlled for in the Mah et al. (2023) experiments. Similarly, participants studied double the number of words in the CR task

relative to the FR task. Research on *list-length effects* has shown (perhaps unsurprisingly) that recall is better for shorter than longer lists (Ensor et al., 2020).

In the Mah et al. (2023) experiments, participants were free to recall FR targets in any order, whereas CR cues at test were presented in a random order. The former perhaps allowed participants to gravitate to the commonly observed *forward recall* (words recalled in the same order they were studied; Tan et al., 2016). For CR, participants were saddled with whatever random order the program presented, potentially introducing noise to the CR distribution (e.g., some participants by chance ended up with an order more conducive to recall, whereas other unlucky participants did not). Similarly, although the studied words were *generally* related (i.e., they were all animals/objects), the FR lists and CR pairs were random. Relative properties of words in CR pairs have been shown to modulate recall performance, e.g., pairs where both words were high in imageability were better recalled than low-high mixed pairs, which were better recalled than pairs where both words were low in imageability (Madan et al., 2010). For FR on the other hand, *semantic* properties of individual words such as *animacy*, *usefulness*, and *size* strongly predict recall. And some properties, like *arousal*, improve memory for items but degrade memory for pairs (Madan et al., 2012). Although the words in Mah et al. (2023, originally drawn from Popp & Serra, 2016) were chosen to be roughly equivalent across animal and object categories in terms of recall-relevant words characteristics like frequency, imageability, and concreteness, this control was at the level of individual words, *not* particular word pairs.

In sum, there are a number of possible incidental methodological factors that could have some bearing on the observed variability difference. Some, like recall order, have clearer intuitive connections to variability, while others, like study time, may not have obvious a priori connections to recall variability but nonetheless represent uncontrolled differences between the tasks. Critically, *none* of these factors have been examined in terms

of their effects on recall variability – almost all have been examined in the context of mean recall performance.

Models of free and cued recall. As mentioned, there is not always a clear distinction between incidental methodological factors and theoretical explanations. Indeed, most of the previously described methodological factors have accompanying theoretical explanations for their effects on recall. And there are several prominent general theories of recall – formalised as computational models – that seek to provide explanations for the gamut of observed methodological effects. What are these models, how do they conceptualise the processes and mechanisms of recall, and critically, do they offer any clues about the novel variability effect?

The Search of Associative Memory (Raaijmakers & Shiffrin, 1980). The *Search of Associative Memory* (SAM) model is an influential model of recall and recognition that has proven able to explain a variety of memory effects (e.g., primacy/recency effects, serial position curves, interference effects, part-list cuing, list length/exposure duration effects; Raaijmakers & Shiffrin, 1980, 1981). SAM was the first *global-matching model*, a class of models that assume that retrieval cues (e.g., list context, specific cues) are used to probe the entire contents of memory, with a 'best-match' being chosen from amongst the competing memories (Huber et al., 2015). Matches in SAM are determined by the strength of association between the retrieval cue probes and the various candidate memory traces (i.e., representations of the study list items; Osth & Dennis, 2020). In the simple example of a free recall list, studied words build associations with the general list context, and with other words in the study list. Words that are studied in close proximity (i.e., held in a limited-capacity short-term "buffer" representing working memory) develop stronger associations with one another, while words further apart in the list develop weaker associations. The degree of association is also a function of study time (words that are studied longer build more

associative strength with the context and with other words in the buffer). At test/retrieval, participants initially only have the list context as a general cue for memory, and this context is used to search among memories for the studied words. SAM assumes that during the eponymous search, memories may be probabilistically *sampled* but not necessarily *recovered* (e.g., 'tip-of-the-tongue' state). If a word is sampled and recovered, it is recalled, and it (and the context, and prior recalled words) serves as a probe for subsequent memory searches until search is terminated. This basic SAM model also incorporates processes representing retrieval-induced incrementing of associations, short-term versus long-term storage, pre-existing semantic relationships among studied words, rechecking, and stopping rules for individual probe searches and the overall search (Kahana, 2020).

Although it has most often been applied to FR data, SAM has also been extended to cued recall tasks like the one in which we observed the variability effect. In cued recall, it is assumed that each pair clears the short-term buffer, and while in short-term memory builds up item-context associative strength and interpair-strength (Raaijmakers & Shiffrin, 1980). At retrieval, each cue serves as a joint probe with the general list context for each memory search. Like in FR, SAM's formulation of CR allows for extra-list intrusions (e.g., unstudied but highly semantically similar words). However, SAM-CR also allows for the possibility of *same-list-non-target* intrusions (i.e., a non-paired cue or target recalled because of its association with list context).

SAM does not make any explicit predictions about variability differences in FR and CR. But it does point to task differences in encoding and retrieval that could explain the variability effect. One difference I have already alluded to – the possibility of within-list interference for CR but not for FR. In SAM, within-list interference cannot occur in FR because all studied words are valid recall targets. But for CR, other cues and targets can interfere with recall of a particular pair. So it is possible that this extra source of interference

(and individual variation in source monitoring or the ability to resist within-list intrusions) introduces added variability. At encoding, FR words build stronger associations with many other words in the short-term buffer, whereas CR items in a pair only build associations with one another. It could be that increased CR variability lies in the *nature* of these individual associations. If an individual does not develop a strong association between a particular cue and target (e.g., by adopting an effective association strategy, or if the particular pair is conducive to association), correct recall will be unlikely. For FR, individuals have the opportunity to associate each word with many other words, potentially reducing the importance of idiosyncratic inter-word relationships or association strategies. Indeed, this pattern is consistent with our prior results showing higher performance for FR than for CR (although there are other plausible explanations for the observed FR performance advantage, like the fact that the CR lists involved twice as many total words, CR study time per word was lower, etc.).

SAM does not speak directly to the variability question, and nor is it obvious *where* in the model a variability difference could arise. However, the above coverage of the model highlights some potentially promising avenues for study. It could be that differences in the source and nature of interference at FR versus CR retrieval contribute to the effect. Or, it could be that the increased influence of particular word pairings in CR (versus multi-word groupings in FR) and the strategies adopted to form pair-based associations in CR (versus the strategies adopted to form group-based associations in FR) might explain why CR is more variable.

The Adaptive Control of Thought Model (Anderson, 1983; Anderson et al., 1998). Anderson's Adaptive Control of Thought (ACT; 1983) model has been successfully used to model a wide variety of higher-level cognitive processes, including recall and recognition. The latest iteration – ACT-R (Anderson, 1993) – has been able to account for effects in free

recall (e.g., latency and error patterns, serial position, list length, delay effects; Anderson & Matessa, 1997) and in cued recall (e.g., spacing effects; Delaney et al., 2018; interference effects on RT; Anderson, 1981). ACT-R posits that memory traces reside in a *declarative memory* system, and the levels of activation of different traces determine the probability that they will be processed by a *production system* (Anderson et al., 1998). Production rules in the latter system coordinate the retrieval of information from the declarative system. Critical for recall is the activation of declarative memory traces. The activation of a particular memory trace is a function of its base-level activation (e.g., increased by repeated exposures at study or recency) and the *associative* activations of other elements currently in attention (e.g., presented cues). Like SAM and other global-matching models, ACT-R assumes that multiple traces are activated in response to a given cue. Which traces are activated depends on the degree of associative strength between the presented cue and a particular trace. The degree of association between a particular cue and a trace is determined by the co-occurrence probability of the cue and the trace (i.e., the proportion of times that trace *i* occurs and cue *j* is also present). Given a particular cue, the memory trace with the highest overall activation will be retrieved if that activation exceeds an internal threshold.

How does ACT-R specifically conceptualise free and cued recall? Applying the model to FR data, Anderson et al. (1998) assume a short-term buffer (similar to SAM) where studied items are held for rehearsal and replaced by new items as the list progresses. Items in the buffer were randomly chosen for rehearsal. At test, items whose activation exceeded a threshold would be recalled – first those still in the buffer would be recalled, and then all other items with sufficient summed activation. In this case, activation increases with the number of encodings/rehearsals, and decreases with study-test delay and list length. Unlike SAM, ACT-R's model of free recall does not incorporate associations between studied words. The only associations that ACT-R considers are between the individual words and the

general list context. These associations rely only on the list length, with longer lists implying less association between each individual word and the context (i.e., a "fan effect"; Anderson & Reder, 1999). Despite this simplification, ACT-R has been able to predict a wide variety of results in FR experiments (Anderson et al., 1998).

For cued recall, context cues, base-level activation, and associations between particular words in pairs determine recall probability. The first two aspects are identical to free recall – the new element is the association between words in pairs. Thus, ACT-R essentially posits that *more* sources of activation contribute to cued recall than free recall. If associative activation differs across pairs, or across individuals, this additional variation could contribute to more overall variation in recall for CR than for FR. Like with SAM, ACT-R does not make explicit predictions about the variability effect. However, like SAM, ACT-R seems to point to the associations between cues and targets as a key difference between cognitive processes in the different tasks.

Other models. SAM and ACT-R are not the only models of recall. Other models that have been applied to recall (with varying degrees of success) include the *theory of distributed associative memory* (TODAM; Murdock, 1995), *MINERVA II* (Hintzman, 1984), and the *composite holographic associative recall model* (CHARM; Metcalfe, 1990), among others. However, these models either a) do not have obvious application to the kinds of free and cued recall tasks of interest or b) have core assumptions that are similar to SAM and/or ACT-R (and in some cases identical mathematical predictions; Osth & Dennis, 2020). SAM and ACT-R are particularly attractive models because they seek to explain memory generally, describe free and cued recall tasks within the same theoretical framework, and provide clear formulations of the processes thought to be involved in both tasks.

That is not to say that these other models do not offer potential insight into the variability question. Some models, such as TODAM, Pike's *matrix model* (Pike, 1984), and

CHARM represent the associations formed during paired-associates learning as a combinative transformation of representations of the items that make up each pair (Osth & Dennis, 2020). It is these *transformed representations* that are compared to test probes. This is qualitatively different from free recall, where items are compared to probes untransformed. So here again, the theoretical difference between free and cued recall seems to lie in whether and what kinds of representations are formed during learning. In *MINERVA II*, paired associates are learned as vectors containing features representing the study context, cue, and target. At test, the context and cue features serve as probes. The similarity of these features to features of the traces stored in memory (including the trace containing the correct target item) determines recall (Atkins, 2001). Thus, *MINERVA II* cued recall performance relies on a combination of features of the study context, as well as the cue and target items, whereas free recall performance depends on encoding and recall of only context and item features. This lends credence to the idea that cued recall may involve either more processes or qualitatively different processes (like the combination of items in to-be-remembered pairs), which could translate into greater performance variability.

Taking stock

What can we say thus far? First, it is possible that inter-individual variability is greater for cued recall than for free recall. This result, if robust, is potentially counterintuitive, and does not seem to be explicitly predicted or straightforwardly implied by dominant formal theories of memory (although clues point to the nature and role of formed associations in cued versus free recall, and how they might quantitatively and qualitatively differ). Second, this result may have important implications for our understanding of memory. At the broader end, greater cued recall variability might imply that the processes underlying associative memory (i.e., the abilities and strategies that people use to form specific associations) differ more across individuals than the processes underlying item memory (i.e., the abilities and

strategies that people use to memorize lists). In this case, investigations of recall variability could prove useful in further uncovering the different processes that underlie both memory types. At the narrow end, greater cued recall variability might simply imply that one or more incidental methodological features of common cued and free recall tasks – as opposed to intentional, theoretically-informed manipulations – result in greater observed variability in performance for the former. This result, although to my mind less interesting, might still be informative for psychometric or statistical reasons.

Finally, to our knowledge, there have not been specific empirical comparisons of memory task variability across individuals. Researchers have examined individual differences on particular tasks – e.g., linking individual strategy use to free recall performance (Unsworth et al., 2019), examining variation in serial recall patterns (Unsworth et al., 2011) – but have not used performance variability as a means to gain insight into similarities and differences between common memory tasks. The goal of the current research was threefold: first, to establish the robustness of the novel and surprising variability effect, second, to probe potential explanations for a variability difference, and third, to determine what (if anything) the results we obtain mean for our understanding of free and cued recall, and memory in general.

An empirical investigation of the variability effect

To test the novel variability effect, we conducted a series of experiments in which we systematically manipulated methodological factors, one-by-one, to test the robustness and generality of the effect, and to consider various potential explanations – both experimental-design and theoretically-motivated in nature. All experiments were preregistered, with materials, data, and analysis code available on the Open Science Framework (https://osf.io/3tra5/).

1. Experiment 1: "Cued vs. Free Nouns"

Our incidental finding of greater inter-individual variability in CR relative to FR (Mah et al., 2023) was obtained in an experiment using animal and object words, which behave in different and specific ways in those memory tasks (e.g., Popp & Serra, 2016). As such, we thought it wise to test for the replicability of that pattern using a more representative set of words than those we had used in our replication of Popp and Serra (2016). To that end, we preregistered and conducted an initial experiment (registration viewable at https://osf.io/xfj6a). We hypothesized that we would observe greater inter-individual variability in CR than FR performance with our new materials.

Method

Materials. We constructed a pool of 120 concrete English nouns designed to be "average" on a number of memory-relevant characteristics (frequency, age of acquisition, concreteness, imageability, and familiarity). Briefly, we began with the MRC Psycholinguistic Database (Wilson, 1988) of 21,561 nouns and then in several steps selected from that pool a set of nouns that are 'average' on multiple dimensions (i.e., within a central mass of the database-wide distribution).[2] The experiment program itself was a modified

[2] See the preregistration (https://osf.io/xfj6a) for the final word pool and details of the word selection procedure.

version of the Livecode program used in Popp and Serra (2016) and Mah et al. (2023). The experiment program and word list can be found at https://osf.io/xfj6a.

Procedure. Participants downloaded and ran the experiment program on their own computers, completing two FR and two CR study-test cycles. Order of FR and CR was counterbalanced across subjects, and each test phase occurred directly after its corresponding study phase. Each list consisted of either 15 words (FR) or 15 word pairs (CR) randomly sampled for each participant from the word pool. At study, each word or word pair was presented on-screen for 5s. At FR test, participants typed in as many words as they could remember before proceeding. At CR test, studied cues were presented one at a time in a random order with a prompt to enter the correct associated target. After completing the four study-test cycles, participants were asked open-ended questions about strategies that they used when studying the FR and CR lists, the subjective difficulty of FR and CR (0 = *Very Easy* - 100 = *Very Hard*), an estimate of the percentage of words they understood (0%, 25%, 50%, 75%, 100%), their age, and whether they encountered any distractions (Major, Minor, None) or technical difficulties.

Sample. Via a priori power simulations, we determined that an $N = 120$ would be sufficient to detect a difference in FR and CR variability at least as large as the lower-bound 95% percentile bootstrap CI on the variability difference observed in Mah et al. (2023). To reach a post-exclusion N of 120, we collected data from 165 undergraduate participants who received bonus course credit for participating. From our total sample of 165, we excluded 45 participants based on preregistered exclusion criteria. Specifically, 6 participants indicated experiencing a major distraction during the study, 22 reported understanding fewer than 75% of the studied words, 24 did not get at least one correct on all four tests, and 20 participants had a CR list on which 50% or more responses were skips with RT < 1 s (note that some participants were excluded on multiple criteria). Our final sample included 120 participants

ages 17-39 (M = 21.4, SD = 4.18). Most (86.7%) of our sample reported English as a first language (4.2% as a second language, 9.2% English bilingual).

We manually checked and coded participant commission errors on CR and FR, counting errors we deemed "close enough" as correct (e.g., minor spelling errors, plural versions of words). In total, 350 FR errors (out of 2,862 total FR responses) and 871 CR errors (out of 3,256 total CR responses) were manually checked by two independent coders. Of these errors, the coders disagreed on 32 FR errors (122 accepted corrections) and 41 CR errors (89 corrections accepted). All disagreements were resolved by a third coder.

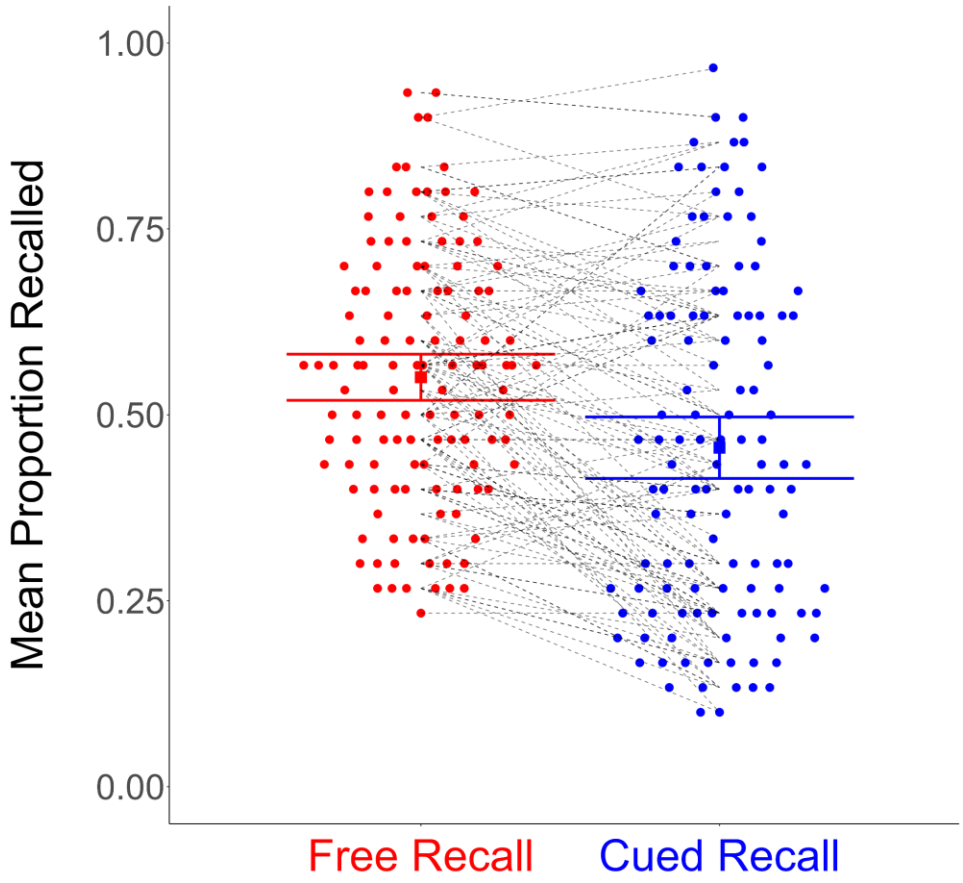
Results & Discussion[3]

Confirmatory analyses. Our critical hypothesis was that inter-individual variability would be greater for CR than for FR. Figure 4 depicts the means, within-subjects 95% CIs, and distributions of CR and FR performance in our sample.

[3] All analyses were conducted in R (R Core Team, 2021). Data files and analysis scripts (including computational model files) are available at https://osf.io/274qd/.

Figure 4

Experiment 1: Memory performance as a function of recall test type



Note. Error bars = 95% CIs (between-subjects). Points jittered horizontally based on relative frequency. Dashed lines connect FR and CR performance for individual participants.

A preregistered paired Pitman-Morgan test indicated that the null hypothesis of equal CR and FR variability was rejected, $t(118) = 4.27$, $p < .001$. Because there is some evidence that the Pitman-Morgan test can be sensitive to nonnormality and may not control the probability of a Type I error when distributions have heavy tails (Wilcox, 2015), we also estimated the critical effect size (i.e., the ratio of CR:FR variance) via nonparametric bootstrap (1000 resamples of the data, estimating the variance ratio with each resample), a method that requires fewer assumptions. The estimated ratio of CR:FR variance was 1.35 (95% percentile

bootstrap CI [1.18, 1.55]), that is, inter-individual variability in memory performance was about 1.35 times greater for CR than FR.[4]

We also evaluated inter-individual variability via generalized mixed-effects logistic regressions. Estimates of inter-individual variability (i.e., random effects on FR and CR performance) also provided evidence against equal CR and FR variability (see Supplementary Material 3B).[5] Thus, we found evidence for the CR:FR variability effect with a general noun wordset, with converging results obtained from a variety of analyses (Pitman-Morgan, bootstrap, GLMM, interocular test)

Exploratory analyses. We conducted several exploratory analyses aimed at uncovering potential mechanisms underlying the variability difference. We examined self-reported FR and CR study strategies (e.g. *Rehearsal/Repetition*, *Imagery*). Of 330 coded responses[6] (i.e., one FR and one CR response for all participants in the full sample), the two coders agreed on 177. The remaining 163 disagreements were put to a third coder. Final coded strategies were those that were reported by at least two out of three coders for a given participant response. The report proportions for each strategy (restricted to the final $N = 120$) are shown in the figure below, with separate proportions for FR and CR:

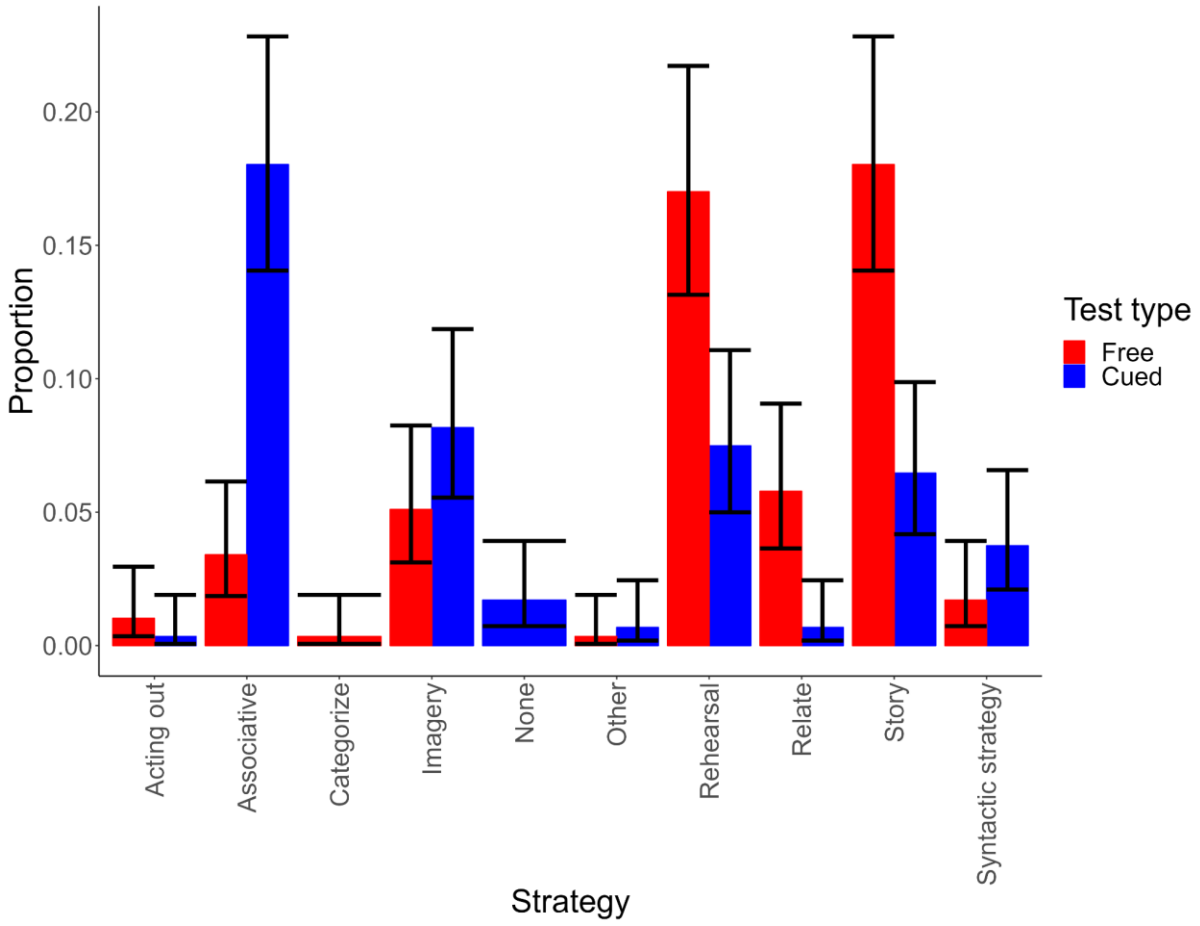
[4] Results were similar when looking at accuracy separately by test order (i.e., for those who did CR first vs. second), see Supplementary Material 3C.

[5] We also fit and compared Bayesian computational models of FR and CR performance (for this an all subsequent experiments). These analyses generally agreed with the ones reported here (see SOM 2 and our preregistration for more details about these analyses)

[6] See Supplementary Material 1 for details on the coded categories and criteria

Figure 5

Experiment 1: Coded study strategy as a function of recall test type



Note. Error bars = proportion 95% CIs (using Wilson's (1927) method).

One cannot estimate the variability of categorical data (i.e., to compare the variability in strategies across test type), but one can compute *unalikeability*, which is an analogue measure that measures the probability of pulling two unequal categorical variables from the sample (Kader & Perry, 2007). Unalikeability ranges from 0 to 1, with higher values indicating greater unalikeability (2007). Unalikeability was similar for CR (.77, 95% percentile bootstrap CI [.72, .81]) and FR (.76, 95% percentile bootstrap CI [.72, .79]). and found a non-significant difference in variability in strategy use. We also looked at self-

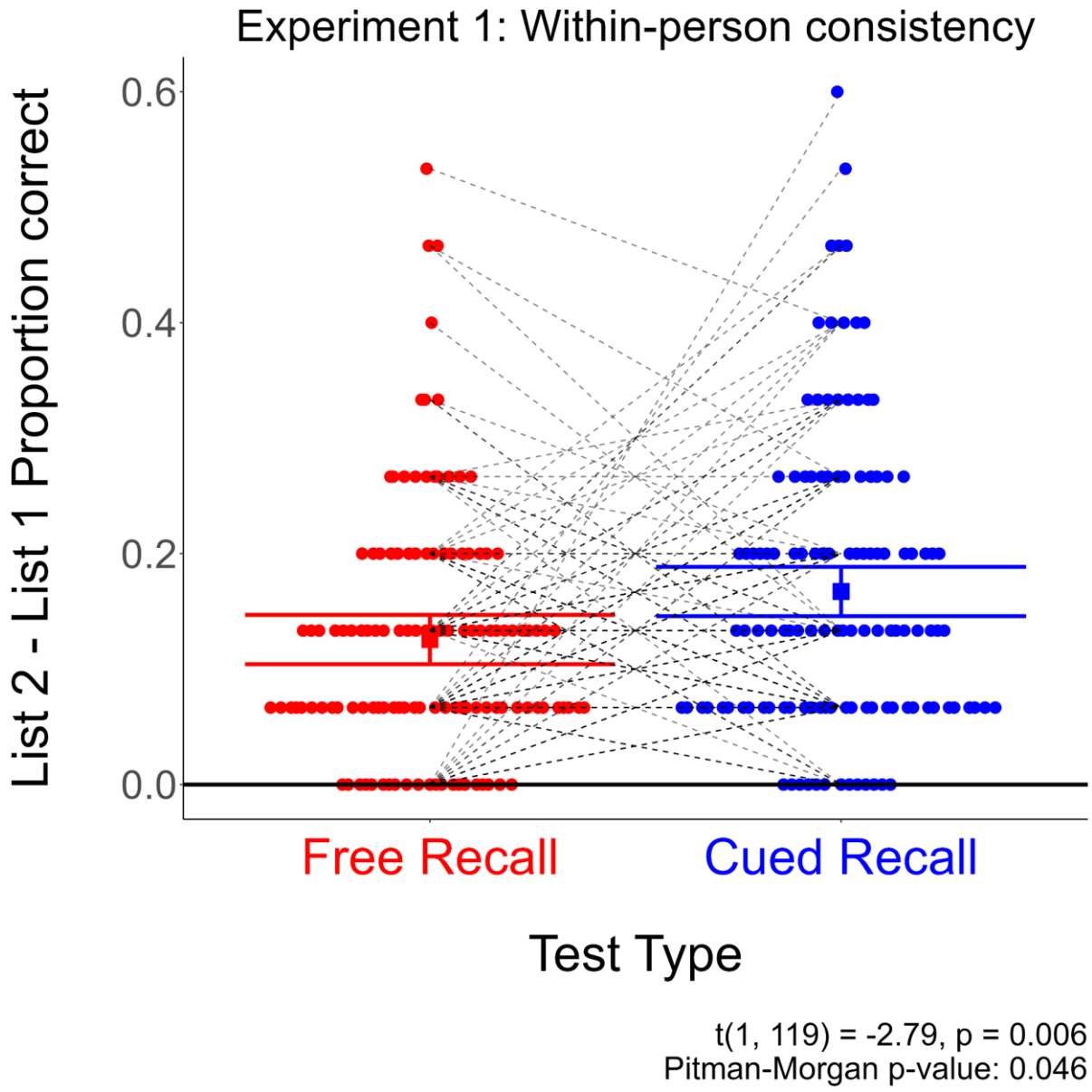
reported recall difficulty, and though participants rated CR as more difficult than FR, there was no evidence for a variability difference (See Supplementary Material 3D).

The inclusion of two within-subjects study-test cycles for FR and CR permitted the investigation of an additional question: do participants differ from *themselves* more for CR than for FR? That is, does the CR:FR variability effect extend to *intra*-individual differences? We had data that permitted an exploratory analysis of this hypothesis: both from this experiment and from the animacy experiments conducted in Mah et al. (2023).

First looking at Experiment 1, which had participants complete two study-test cycles each of FR and CR, we computed the absolute difference in performance between first and second lists for both memory types. Figure 6 below plots these differences:

Figure 6

Experiment 1: Within-person consistency across study lists for FR and CR



Note. Error bars = 95% CIs (within-subjects). Points jittered horizontally based on relative frequency. Dashed lines connect FR and CR between-list differences for individual participants.

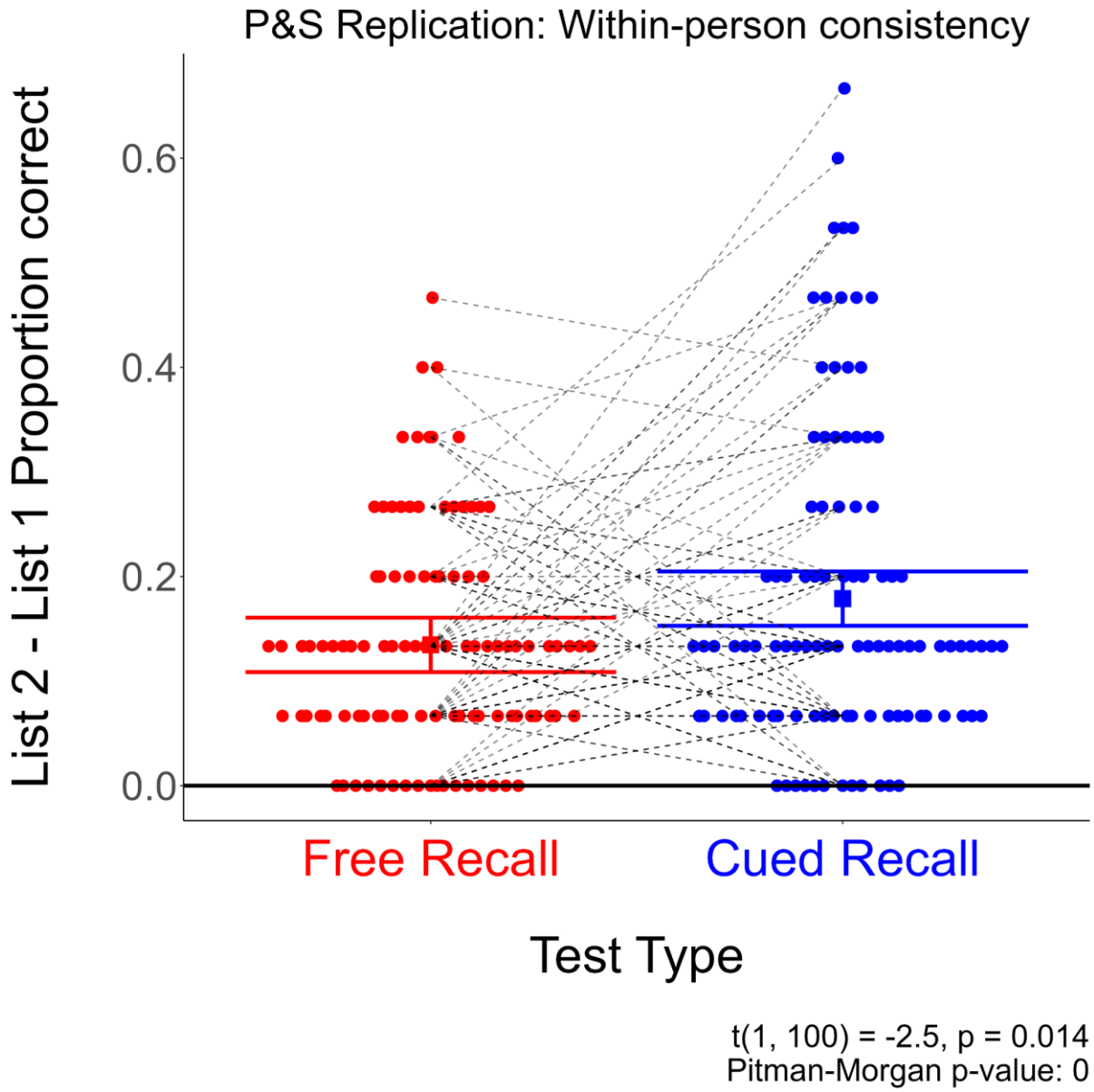
If participants are less self-consistent for CR than for FR, then the absolute difference between lists should be higher for CR for FR. As indicated in the figure, this was the case.

Additionally, a Pitman-Morgan test of the inter-individual variability in self-consistency showed that not only were participants less self-consistent for CR than FR, self-consistency itself varied more for CR than for FR. That is, the degree to which participants performed similarly across tests was more variable for CR than for FR. Interestingly, cross-list changes in FR did not predict cross-list changes in CR, $F(1) = 2.35$, $p = .13$. That is, those who improved (worsened) on FR from the first to second list did not necessarily improve (worsen) on CR from the first to second list.

We also investigated within-person variability in two datasets from animacy experiments (Mah et al., 2023). In these experiments, participants studied lists of animal and object words or pairs, and like our Experiment 1, completed two FR and two CR study-test cycles. In the first of these experiments (the replication of Popp & Serra, 2016 mentioned in the introduction, $N = 101$), the results were similar:

Figure 7

Popp & Serra Replication: Within-person consistency across study lists for FR and CR



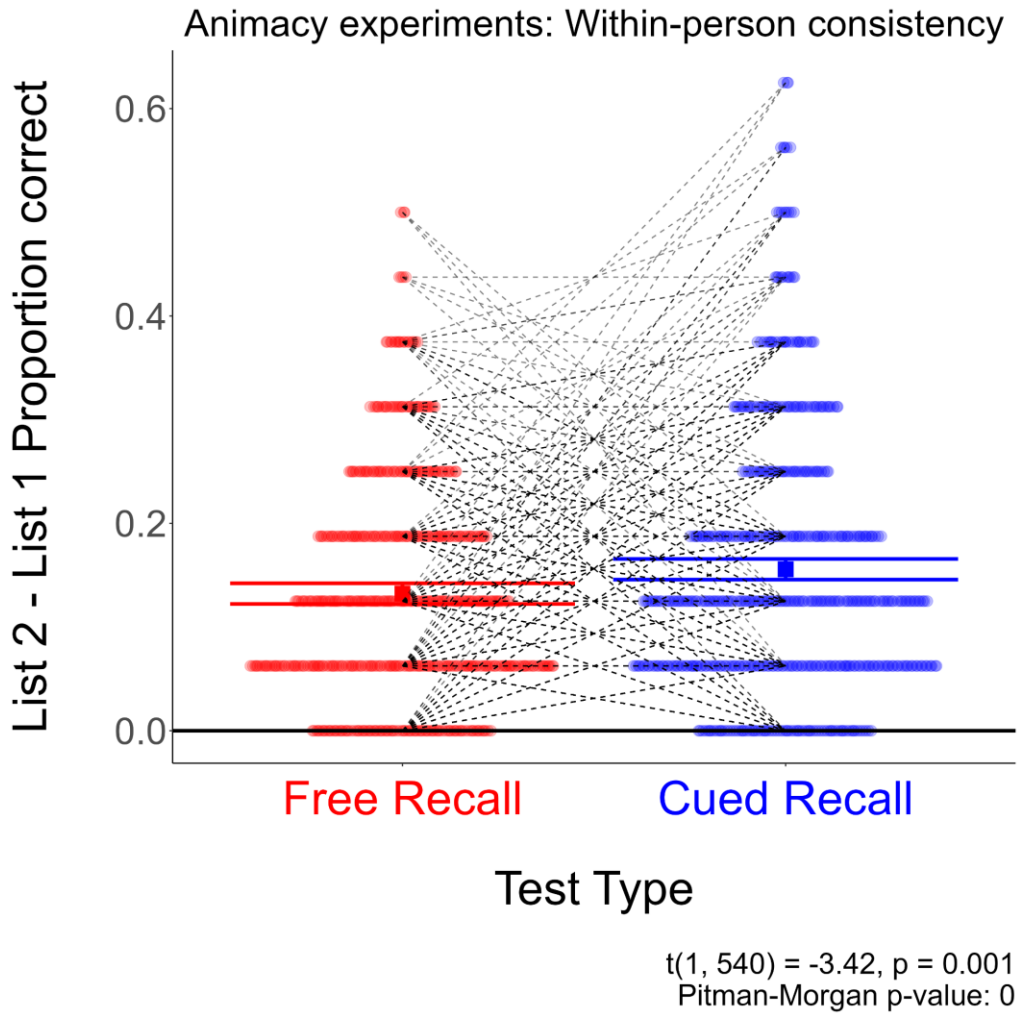
Note. Error bars = 95% CIs (within-subjects). Points jittered horizontally based on relative frequency. Dashed lines connect FR and CR between-list differences for individual participants.

Again, self-consistency was significantly lower for CR than FR, and also significantly more variable for CR than FR. Consistent with the previous dataset, across-list change for FR did not predict cross-list change for CR, $F(1) = .08$, $p = .78$.

Finally, we applied these analyses to the combined sample of four additional animacy experiments ($N = 541$) with the same basic design. In this, our largest sample, we observed the same results:

Figure 8

Animacy experiments: Within-person consistency across study lists for FR and CR



Note. Error bars = 95% CIs (within-subjects). Points jittered horizontally based on relative frequency. Dashed lines connect FR and CR between-list differences for individual participants.

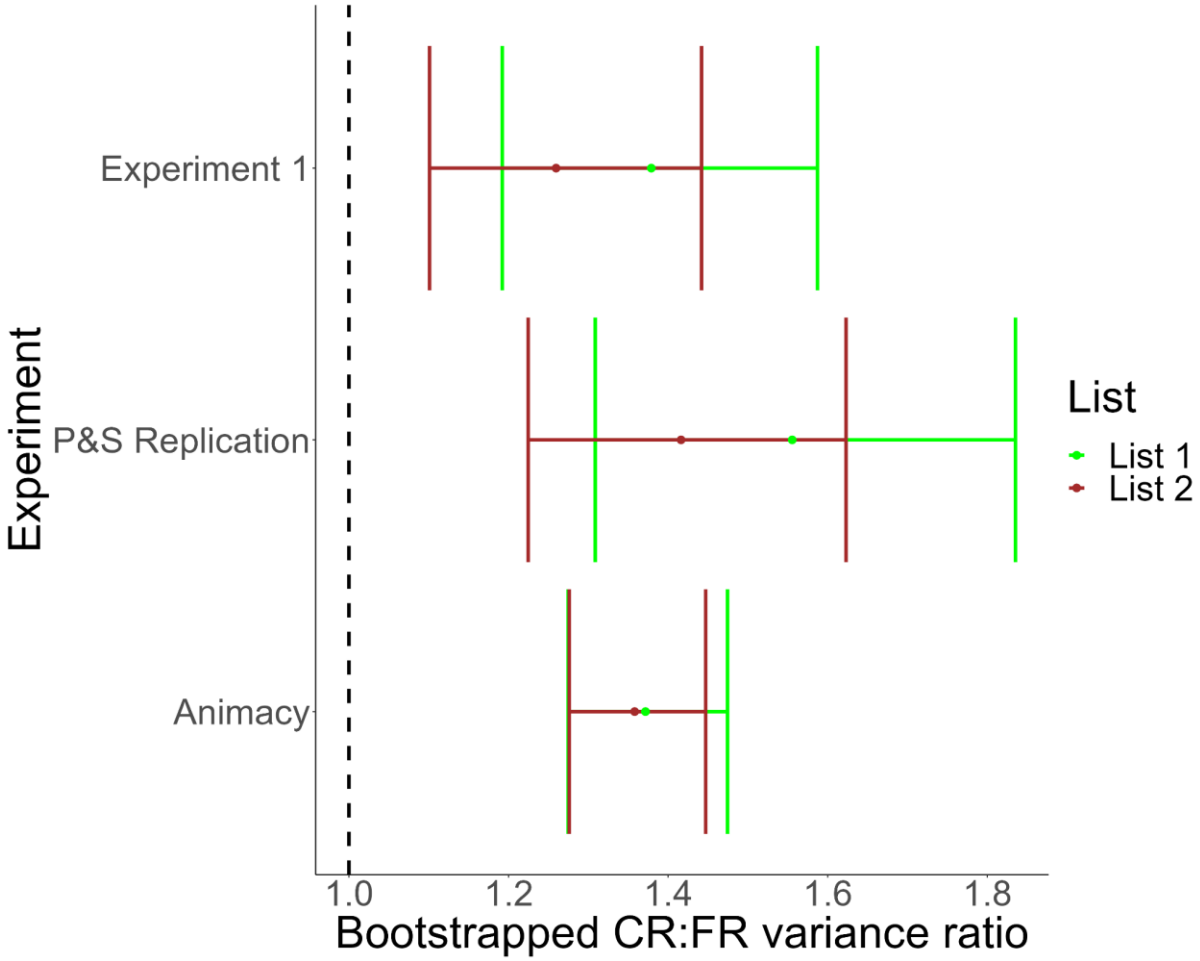
Similar to the previous datasets, cross-list change in FR did not predict cross-list change in CR.

What does evidence of greater *within*-person variability imply? When considering the between-subjects variability effect, it makes sense to say (for instance) that the processes or encoding/retrieval strategies involved in CR vary more person-to-person than the processes involved in FR. However, this explanation seems less tenable for within-person variability–it seems unlikely that individuals change their strategies across multiple CR tests, or that the processes underlying CR vary from test to test. What then explains this?

One possibility we had not considered thus far is that participants may simply be less familiar with CR-type tasks. This would explain the fact that while participants tended to hover around the same performance level for FR, more participants improved than worsened for CR (see SOM 10). If this is the case, then one might expect the CR > FR variability effect to be smaller (or gone) on second lists relative to first lists. Although the bootstrapped CR:FR variance ratios were slightly smaller on second lists in all the multi-list experiments (see figure below and SOM 11), the ratios were similar and well above 1:

Figure 9

Bootstrapped CR:FR variance ratios for multi-list experiments



Note. Estimates obtained using 1000 bootstrap resamples. Error bars = bootstrapped 95% quantile intervals.

So, although it is possible that some of the CR variance could be due to varied understanding of or familiarity with the task it seems unlikely that this fully explains why participants varied more across lists for CR than for FR.

Finally, we speculated that greater variability in accuracy on CR than FR may be due to participants' interpretations of standard CR task requirements being more variable than their interpretations of standard FR task requirements, particularly regarding how to respond when unsure of an answer. For example, on FR participants may be more inclined to

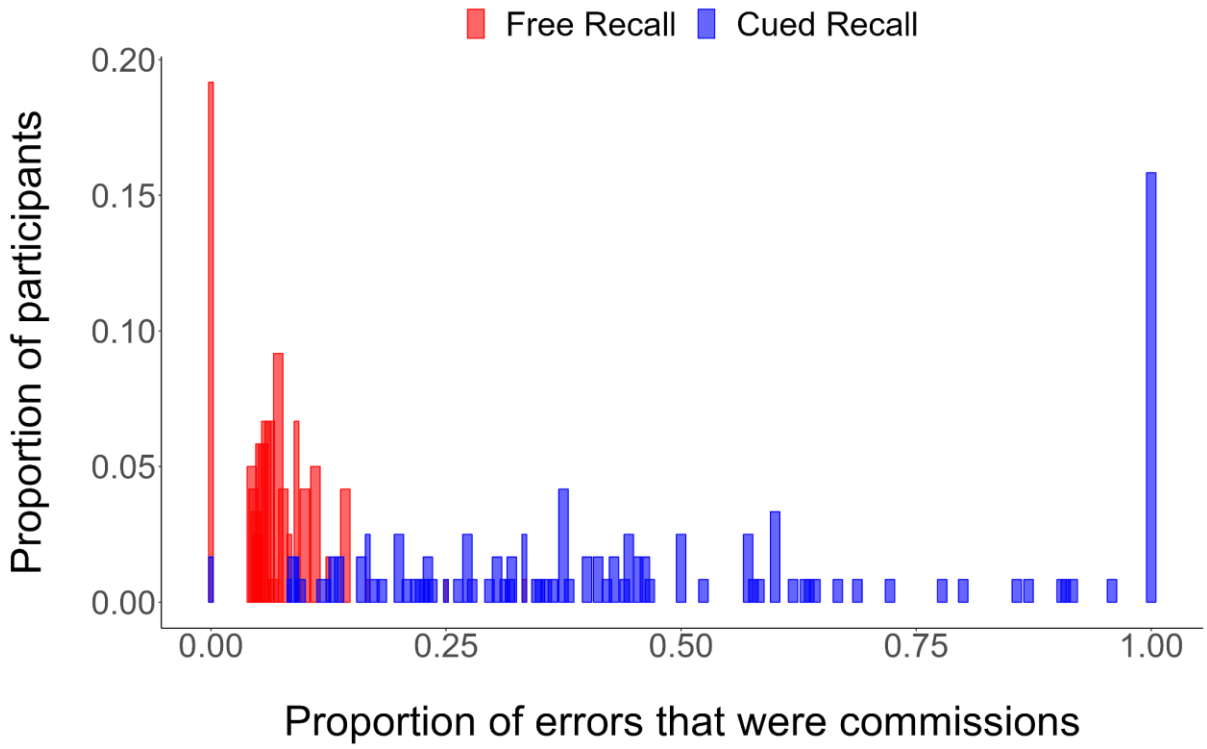
terminate the test than to make additional guesses. On CR, in contrast, the cue might nudge some participants to guess and others to leave it blank. In fact, the CR instructions stated that participants could "guess or leave the box blank," while the FR instructions merely told participants to end the test when they "cannot think of any more words."

To investigate this possibility, we conducted an exploratory analysis of participants' tendency to make errors of *omission* (i.e., a failure to recall a given target) versus errors of *commission* (i.e., an incorrectly recalled target – from a current list, previous list, or new intrusion not previously studied). Looking first at what kinds of commission errors participants made, the vast majority (91%) of FR commission errors were new intrusions (i.e., a word not studied in a previous FR or CR list), with the rest (9%) coming from previous lists. For CR, the majority of commission errors were also new intrusions (65%), with some coming from the same list (30%) and previous lists (5%, FR or CR). The majority of CR same-list commission errors (69%) were studied targets recalled with the incorrect cue (versus 31% where a cue was recalled in place of a target).

Next, for each participant we computed the proportion of FR and CR errors that were commission errors. Thus, a "commission proportion" of 0 means that none of a participant's errors were commission errors, while a commission proportion of 1 means that all a participant's errors were commission errors. A histogram of these commission proportions is shown in Figure 10.

Figure 10

Experiment 1: Commission error proportion by recall type



There was a striking difference between the distribution of commission proportions for FR and CR. For FR, commission proportions are closely clustered around lower values, but for CR, commission proportions were spread more evenly. A Pitman-Morgan test confirmed this, with greater variability in commission proportions for CR than for FR, $t(118) = 33.73$, $p < .001$. Thus, it appears that variability across participants in "propensity to guess" was greater for CR than FR. Alternatively, due to the non-trivial proportion of CR commission errors that came from the same list (30%), the proportions above could also represent greater variability in "propensity for same-list intrusions."

Exploratory analyses: Applying the Search of Associative Memory (SAM) Model (Raaijmakers & Shiffrin, 1981). Although an exhaustive evaluation of the variability effect across the various computational models previously described is beyond the scope of the current project, we were curious whether and how one popular model (SAM) could handle

the anomalous results we found. To that end, we applied the basic SAM model to the Experiment 1 data–specifically the first list each of CR and FR. To supplement the brief overview of the model we gave previously, here we will describe the important model components and parameters in the basic formulation of SAM. These can be broadly broken down into *encoding* components/parameters and *retrieval* components/parameters.

Encoding. During encoding, single FR items or CR pairs enter a short-term store buffer that holds up to *m* items. For FR, as an item enters the store, it moves to the first position, shifting all other items in the buffer down. When the buffer is full and a new item enters, a random item is moved out of short-term store. For CR, each new pair that enters the buffer clears the buffer of the previous pair. While items/pairs are in the short-term store, associations *S* develop between each item *i* and context *c* (i.e., the general study context)–that is, $S(i, c)$ increases by amount *a* for the amount of time *t* the items are in the buffer. Similarly, items *i* in the store develop associations with all other items *j* in the store (item-item strength for FR, interpair-strength for CR)–that is, $S(i, j)$ for all i,j increases by amount *b* for each unit of time the items co-occupy the buffer. For FR, the increase in associative strength also incorporates total time in the buffer and the number of words *n* in the buffer–that is, for any individual word the total increase in $S(i, j) = b * (t / n)$. For CR, because there is only ever one pair of words in the buffer at a time, the increase in interpair-strength is simply $S(i, j) = b * t$. As words/pairs are studied, the item-context and item-item associations populate separate matrices that represent long-term store. Words that never co-occupied short-term store are still considered to have a small amount of residual association *d* with one another in virtue of being studied as part of the same list. Sometimes (and in our current application of the model), the preexisting semantic relationships between words are entered into a separate matrix that modifies the item-item matrix (i.e., weighting the item-item associations by item-item semantic similarity). These matrices are used to model search and recall during retrieval.

Retrieval. For FR, the contents of the short-term store (i.e., *m* items) is output, capturing recency effects. Then, retrieval proceeds as a search process–first, LTS is probed using only context as the cue. Items are probabilistically sampled in proportion to their item-to-context strength (from the item-context matrix). So, items that spent more time in short-term store will have a higher chance of being sampled from the pool of studied items. However, just because an item is sampled does not mean it will be successfully recalled. Specifically, the first sampled item will be recalled with a probability corresponding to its strength of association with the current cue (i.e., context alone). If recall is not successful, recall is attempted *l* times with the current cue up to a maximum of l_{MAX}. If recall succeeds on repeated attempts, context *and* the successfully-recalled item serve as a compound probe cues for the next search. If recall fails after repeated attempts, then the item cue is discarded and context alone is used for the next search. Each failed retrieval attempt contributes to a tally *k*, and at k_{MAX} (a 'failure criterion') retrieval is terminated. SAM also incorporates learning-during-retrieval via a set of parameters that increment when a word is successfully recalled: *e*, an increase in the strength of association between the item and context, and if there is item cue(s) present, *f*, an increase in the strength of association between the recalled item and the item cues in the probe set, and *g*, an increase in the strength of association between an item and itself.

For CR, the process is similar but simpler–for each test item, the probe cue is always context and the current cue. Items are again probabilistically sampled in proportion to their item-to-context and item-to-cue strength, and recalled with probabilities corresponding to their compound strength of association. Each context + item compound cue is used to probe search *l* times until either successful recall or l_{MAX} failures. If an item is successfully recalled, its strength of association with context is incremented by *e* and its strength of association with the presented cue by *f* (though largely not relevant as each cue is presented only once).

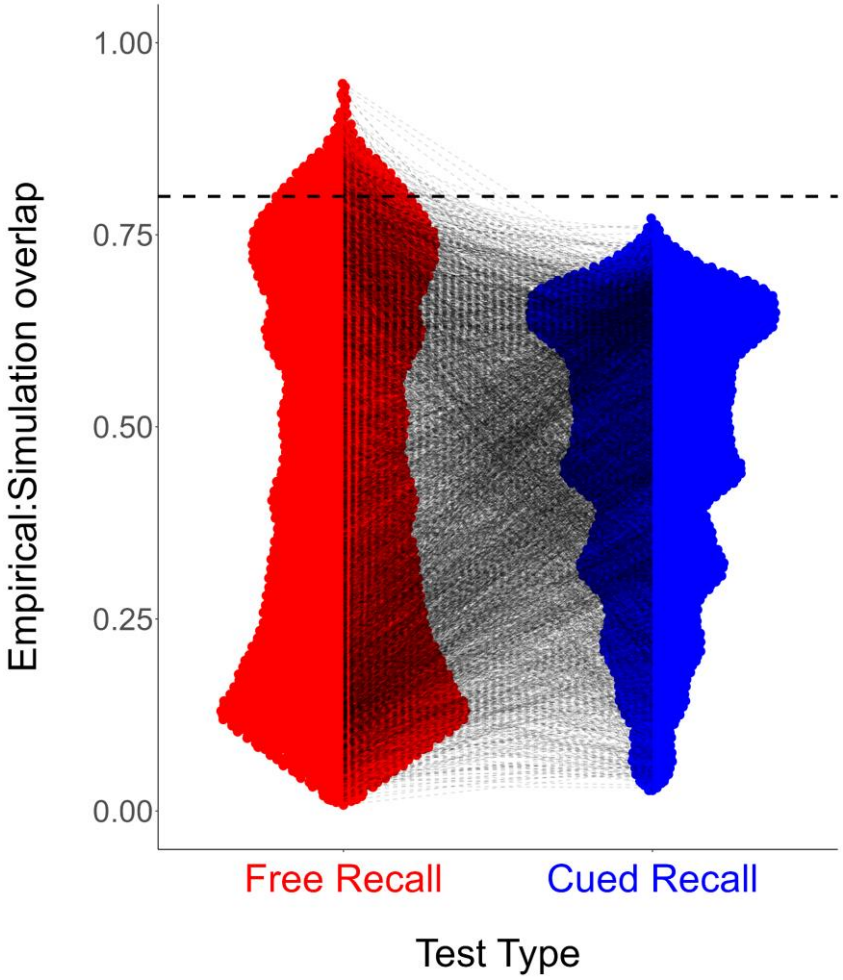
Overall search is terminated once all cues are presented.

The basic SAM model. Using this initial model, we conducted the first batch of simulations–4,604 generated datasets using a simplified version of the design of Experiment 1 as the basis (simulated accuracy for 120 participants completing one FR and one CR list comprising 15 words/pairs each). Parameters *a, b, c, d, e, f,* and *g* were randomly sampled from a 0-1 bounded uniform distribution, and parameters k_{MAX} and l_{MAX} were randomly sampled from a 1-20 bounded uniform distribution *for each dataset*. Buffer size *m* was set to 4. That is, in this initial run all participants shared the same parameter values, which were informed by Raaijmakers & Shiffrin's (1981) results and recommendations for broadly applicable parameter values. Note that although participants shared the same parameter values, each participant (and indeed each test item) were simulated independently. That is, although the *probability* of individual item recall was determined by the parameters, the actual outcome of a given recall attempt was stochastic—randomly drawn from a binomial distribution. The primary objective of this simulation was to see if the most basic version of the model could simulate data in which there was a CR > FR variability effect.

The simulated datasets were then compared to the empirical data from Experiment 1 (specifically, accuracy on the first FR and CR lists). Model fit to empirical data was quantified in two ways. First, via the degree of distributional overlap (specifically, the integral of the minimum between the two densities) between the simulated FR/CR data and the empirical FR/CR data (Pastore et al., 2022). The higher the degree of overlap, the better-fitting the simulation. The figure below shows overlap values for each memory test type:

Figure 11

Basic SAM model: Distributional overlap between simulated and empirical FR and CR data



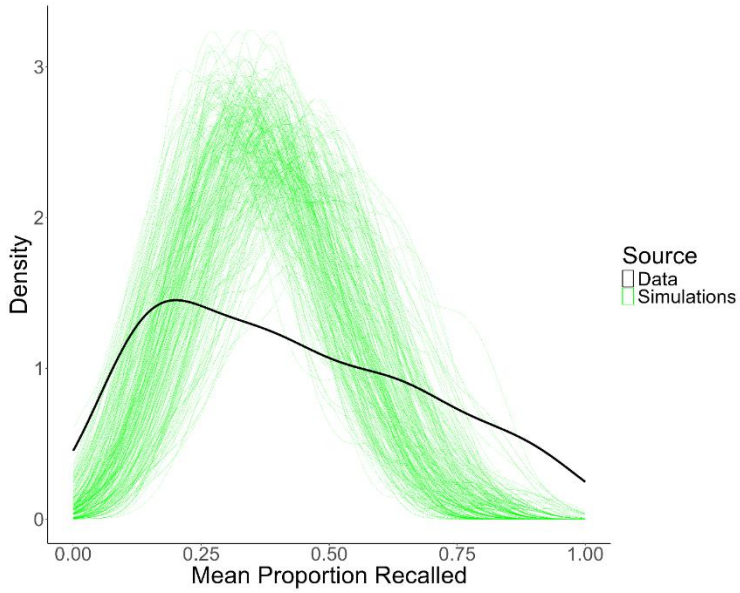
Note. Points jittered horizontally based on relative frequency. Dashed lines connect FR and CR overlap values for a particular simulation.

The solid horizontal dashed line indicates an '.80 overlap cutoff', a criterion based on the second quantification of model fit–Kolmogorov-Smirnov tests of the equality of the simulated and empirical distributions. In these simulations, the .80 (80%) overlap cutoff corresponded roughly to Kolmogorov-Smirnov *p*-values greater than .05, indicating a failure to reject the null hypothesis of distributional equality.

Immediately evident is the fact that although some simulations resulted in adequate fits to the Experiment 1 FR data, *no simulations resulted in well-fitting data for the CR data*. In fact, the highest overlap/*p*-value obtained for CR was .76/.01. This suggests that, without modification, the basic SAM model may not be able to account for the CR:FR variability effect. To determine what the nature of the model failure might be (e.g., does SAM fail to capture the *shape* of the empirical CR distributions, the *variability* of the distributions, or both?), we examined simulated data from 500 of the fits with the highest CR overlap. Distributions from these simulations (along with the empirical distribution) are plotted below in Figure 12.

Figure 12

Basic SAM model: Best CR fits



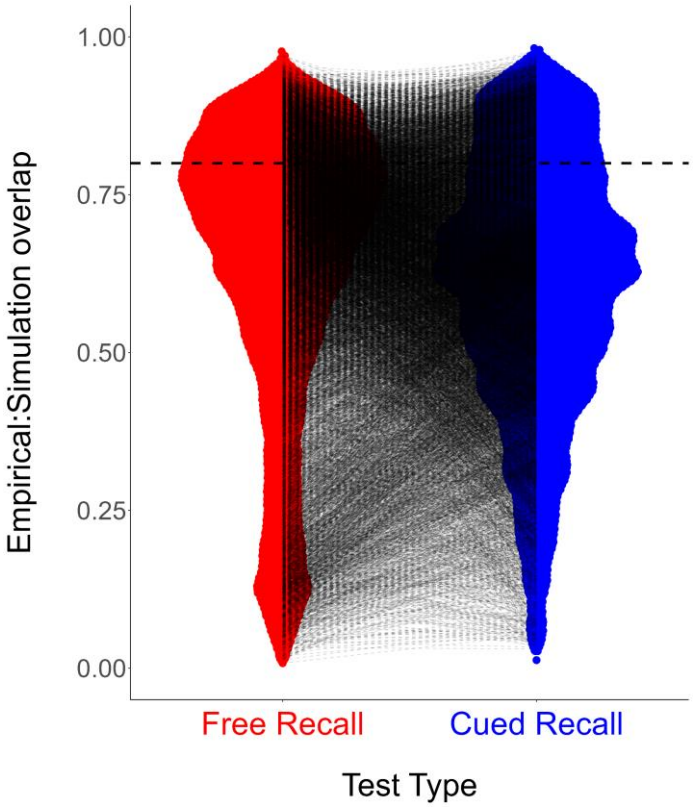
Note. Each green line represents a unique simulation.

From these results, it seems clear that SAM is able to capture multimodality—its weakness seems to be an inability to capture the *spread* of CR performance observed in the empirical data. In other words, the original SAM model fails to successfully predict the high degree of individual-to-individual variability in the data.

The modified SAM model. In the basic SAM model that we used, the same parameter values were used for each subject. The very nature of the CR:FR variability effect and the observed failure of SAM to model the full range of CR performance suggests that the key lies in *individual differences*. So an obvious extension to the basic SAM model is to estimate each of the parameters separately for each participant (but still using the same parameter values across test type). Using the mean best-fitting parameter values from the first search, we conducted an additional 9,908 simulations in which the parameter values were chosen separately for each participant, with the variability in a particular parameter value across subjects chosen anew each simulation. We again examined overlap between the empirical and simulated distributions, including results from both simulation runs:

Figure 13

Extended SAM model: Overlap between simulated and empirical FR and CR data

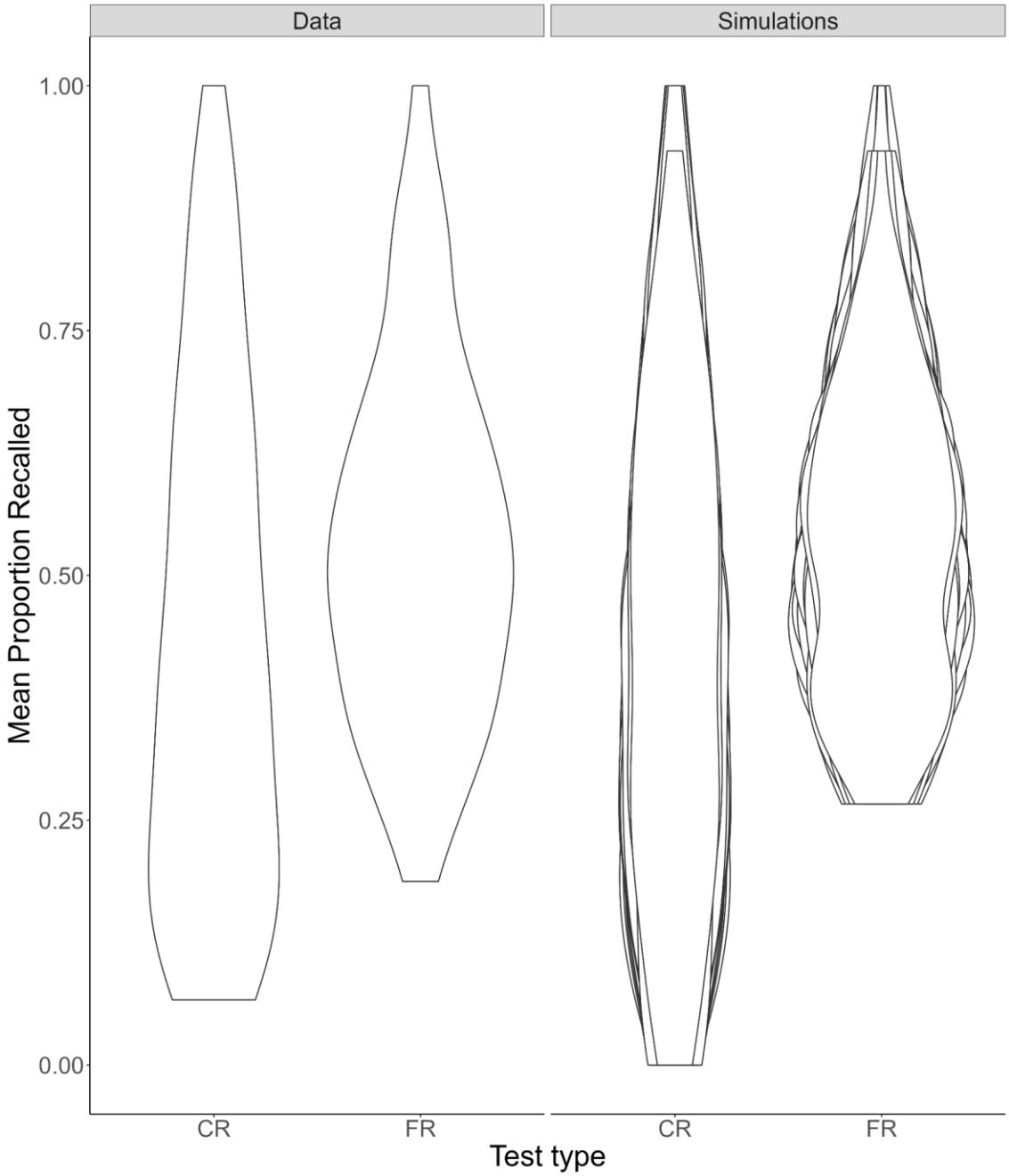


Note. Points jittered horizontally based on relative frequency. Dashed lines connect FR and CR overlap values for a particular simulation.

The estimation of separate parameters for each participant resulted in a number of parameter configurations that fit the data well. To better visualize model fit, we plotted a sample (n = 90) of the best-fitting simulations:

Figure 14

Extended SAM model: Comparison of best-fitting simulations and empirical data

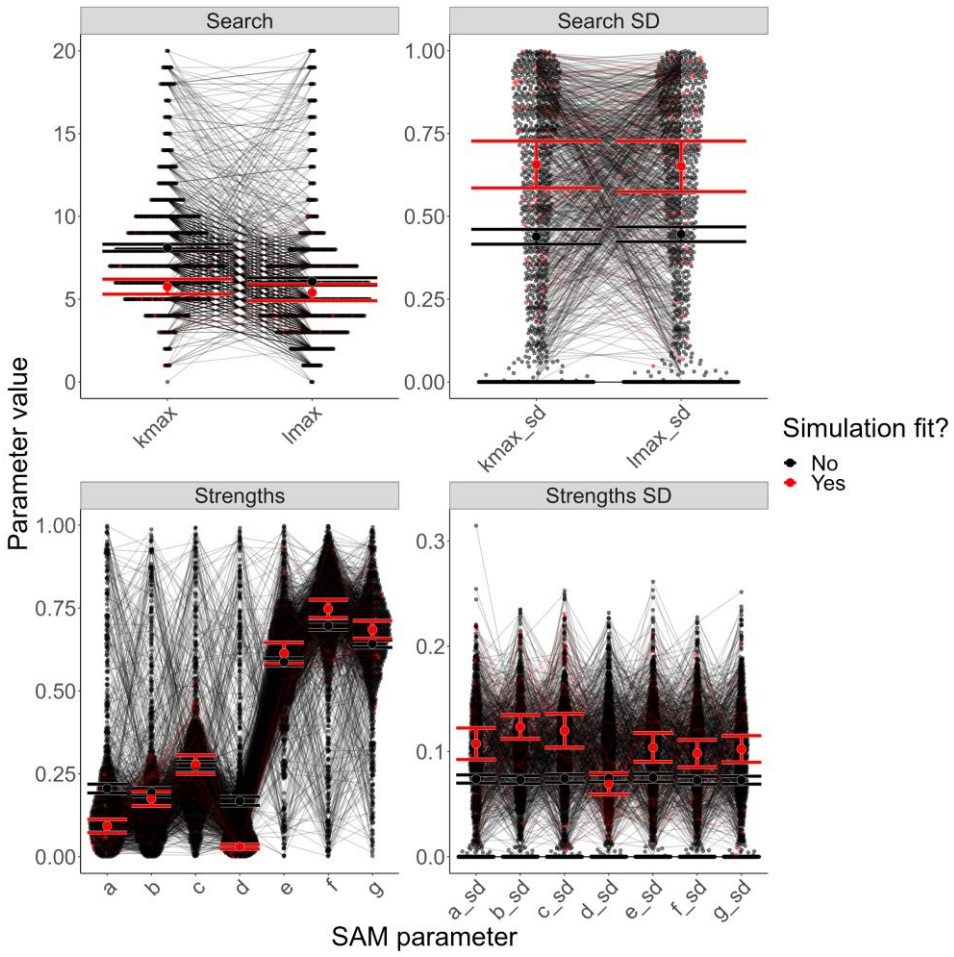


Note. Each violin plot in the second panel represents one of the 90 best-fitting simulations.

As we can see, these simulations captured the qualitative data patterns quite well. The question is, a) is there a consistent pattern of parameter values that result in this pattern (i.e., greater CR than FR variability), and b) can we make sense of the best-fitting parameter values? To investigate these questions, we compared parameter estimates (both parameter values and variability in those parameter values) in well-fitting (i.e., non-significant K-S tests for both CR and FR) and poorly-fitting simulations (i.e., significant K-S tests for either CR or FR):

Figure 15

Extended SAM model: Parameter values for well-fitting and poorly-fitting simulations



Note. Small points represent individual parameter estimates (jittered horizontally by frequency), lines connect estimates from the same simulations. Large points and error bars: Mean estimates and 95% CIs for simulations that either fit (red) or did not fit (black).

The key patterns that we sought were cases in which parameter estimates differed for fitting versus non-fitting models–the logic being that those parameters (and obtained values) were more likely to drive the good fit of the best models. The key differences were:

a) *Lower k_{MAX} (search termination criterion) and higher individual variability in search termination criterion*
b) *Lower a (item-to-context increment) and d (residual item strength)*
c) *Higher variability in all association strength parameters (except residual strength)*

Do these differences hint at possible explanations for the CR:FR variability effect? There do not seem to be obvious links. Lower search termination criteria could lead to lower variability in FR than CR, because k_{MAX} does not factor into the simulation of CR, and reducing the chance of more individual-differences-driven searches could reduce FR variability. But *higher* individual variability in k_{MAX} works against this point. Lower item-to-context incremental associative during strength has some theoretical attractiveness to it–by lowering the role that general context plays as a retrieval cue, there is perhaps more emphasis on inter-list and inter-pair associations, a possible source of the variability effect that we previously alluded to. But we did not also see any differences in *b*–item-to-item associative strength–between well-fitting and poorly-fitting models. Similarly, it is not clear how lower residual item strength (i.e., the small associations formed between all words in a study list) might explain higher CR variability. Finally, larger individual differences in the association-strength parameters for well-fitting versus poorly-fitting models suggests that broadly, individual differences are important in accounting for the CR:FR variability effect, but it is hard to link any parameter in particular (e.g., the seemingly-important *b* parameter) to the variability effect.

It is possible that a further extended version of SAM–a hierarchical model in which participants have separate parameters for FR and CR–could provide clearer clues into the

CR:FR variability effect. However, without more empirical tests of the effect, it is not clear which parameters to selectively vary, which parameters to vary more for CR than FR, etc. It is also still not clear whether the effect is due to processes that can be described by SAM (or other models), or due to more incidental experimental design factors. What does seem to be clear is that at least in its commonly used form, SAM does not seem to be able to account for the observed CR:FR variability difference in a theoretically satisfying way. We turn our focus now to further empirical manipulations aimed at probing the effect, but will return to consider the implications of our experiments for theoretical models of memory at the end.

2. Experiment 2A & 2B: "Forced Recall"

As was suggested by some of the exploratory analyses in Experiment 1 (within-person variability, commission error patterns), it could be that the CR recall task is inherently more ambiguous than the FR one. That is, increased CR variability may simply be due to CR instructional ambiguity and/or greater variability in how participants interpreted the CR task. Alternatively, there could have been differential regulation of reporting in the CR task (e.g., Goldsmith & Koriat, 2008). We attempted to address these possibilities in Experiment 2 by implementing a *forced recall* procedure in which participants at test had to provide a word for each target they had studied (in the vein of Roediger & Payne, 1985). If the variability difference in Experiment 1 was due to greater variability in interpretation of CR test instructions, then forced recall should eliminate that difference, especially when performance is measured in terms of number of targets recalled irrespective of whether they are reported in response to the correct cue. Specifically, we hypothesized that we would *not* observe a CR:FR variability difference in memory performance, both when memory performance was measured in terms of targets reported in response to the correct cue and targets correctly recalled (even to the wrong cue). We preregistered these hypotheses (viewable at

https://osf.io/3w6fm) and tested them in two samples (Experiment 2A: Prolific, Experiment 2B: undergraduates).

Method

Materials. We made several changes to the materials for Experiment 2. First, we re-examined and reduced the set of 120 nouns used in Experiment 1, excluding any words with salient non-noun meanings and any words we thought might be unfamiliar to participants (e.g., HIND). Word exclusions were based on the subjective ratings of three research team members.[7] The reduced wordset contained 83 words. The reduced wordset and experiment program (now made in PsychoPy & run via Pavlovia) can be found at https://osf.io/yv3b7/.

Procedure. In Experiment 2, participants completed one FR and one CR study-test cycle (order counterbalanced), each consisting of 15 words/word-pairs. As in Experiment 1, words/word-pairs were presented for 5s each at study, with standard FR/CR study instructions. Participants were given standard FR/CR study instructions, but at test had to provide a response for each target they had studied. That is, on the FR test participants had to provide 15 words before they could continue, and on the CR test participants had to provide a word for each cue that appeared. The exact instructions were as follows, for FR:

> On the next page, you will be tested on the word list that you just studied. You will try to recall as many of the studied words as possible, and will need to recall one word for every word that you studied (15 words total). So, if you recall less than 15 words you will need to make your best guesses for the remainder. Each word you enter will be displayed on the screen after you enter it. Remember that the order of the words you

[7] If two out of three raters considered a word to have a salient non-noun meaning or to be too obscure, that word was removed from the pool.

recall does not matter for them to be counted correct.

…and for CR:

> On the next page, you will be tested on the word pairs that you just studied. For this test, each of the left-side words from each of the pairs that you studied will appear one at a time, and in a random order. For each left-side word, your task is to recall the right-side word that went with it. So if you studied "guitar – spoon", you would have to recall "spoon" when presented with "guitar". You will need to attempt to recall a right-side word for each left-side word presented, even if you can't recall the correct target. If you can't recall a particular correct target, give your best guess for it.

Entered words had to be at least two letters long. Participants were not told about the forced recall manipulation at study. After completing both study-test cycles, participants completed the same questions as in Experiment 1, with the only differences being the addition of a cheating question ("Did you take notes?"), self-reported frequency of withholding a given CR target because of certainty that it was studied but uncertainty that it was paired with the given CR cue (never, once/twice, several times, very often), and self-reported frequency of recalling a given CR target for multiple CR cues due to realizing that the target actually went with the later presented CR target (never, once/twice, several times, very often). .

Sample.

Experiment 2A. We once again had a target $N = 120$, and collected this sample from

a total sample of N = 150 Prolific participants[8], from which we excluded: 12 participants who didn't get at least one correct on both lists, 14 who didn't report understanding at least 75% of words, 2 who reported a major distraction, 4 who reported cheating, 7 who reported completing a prior version of the study (i.e., on mTurk), and 2 reported technical difficulties. Our final sample included 120 participants aged 18-68 (M = 33.97, SD = 12.49). Participants received $3 USD for participating in the +/- 15-minute study.

As with the previous experiments, commission errors were manually checked by two coders (930 FR errors out of 2,250 total FR responses, 1334 CR errors out of 2,250 total CR responses). Coders disagreed on 20/930 FR errors (38 accepted corrections) and 6/1334 CR errors (28 accepted corrections). Disagreements were resolved by the 2nd coder.

Experiment 2B. For our undergraduate sample, we used the same target N = 120 as in previous experiments, and ended up with slightly more[9], N = 127, after making exclusions from a total sample of N = 207 participants. From our initial sample, we excluded: 31 participants who didn't get at least one correct on both lists, 31 who didn't report understanding at least 75% of words, 23 who reported major distraction, 5 who reported cheating, 12 who reported completing a prior version or similar version of the study (i.e., on SONA), and 4 who reported technical difficulties. Our final sample included participants

[8] Before conducting Experiment 2A, we pilot tested the procedure on Prolific (N = 16). This testing revealed a high rate of exclusions (14/16 participants reported not understanding at least 75% of words), so we pre-registered an additional inclusion criterion for this sample: English as a first language (self-reported on Prolific), in addition to self-reported English fluency. This had the added benefit of making our Prolific sample more comparable to our student samples in terms of language status.

[9] This was due to our sampling procedure (i.e., opening more study slots than we needed to maximize data collected while trying to anticipate exclusions), but the results were the same when including/excluding the seven additional participants.

aged 16-40 (*M* = 19.65, *SD* = 3.57). Participants received bonus course credit for participating.

Coders manually checked 1419 FR errors (out of 3,105 total FR responses) and 2138 CR errors (out of 3,105 total CR responses), disagreeing on 18 FR errors (91 accepted corrections) and 12 CR errors (42 accepted corrections). Disagreements were resolved by a 3rd coder.

Results & Discussion

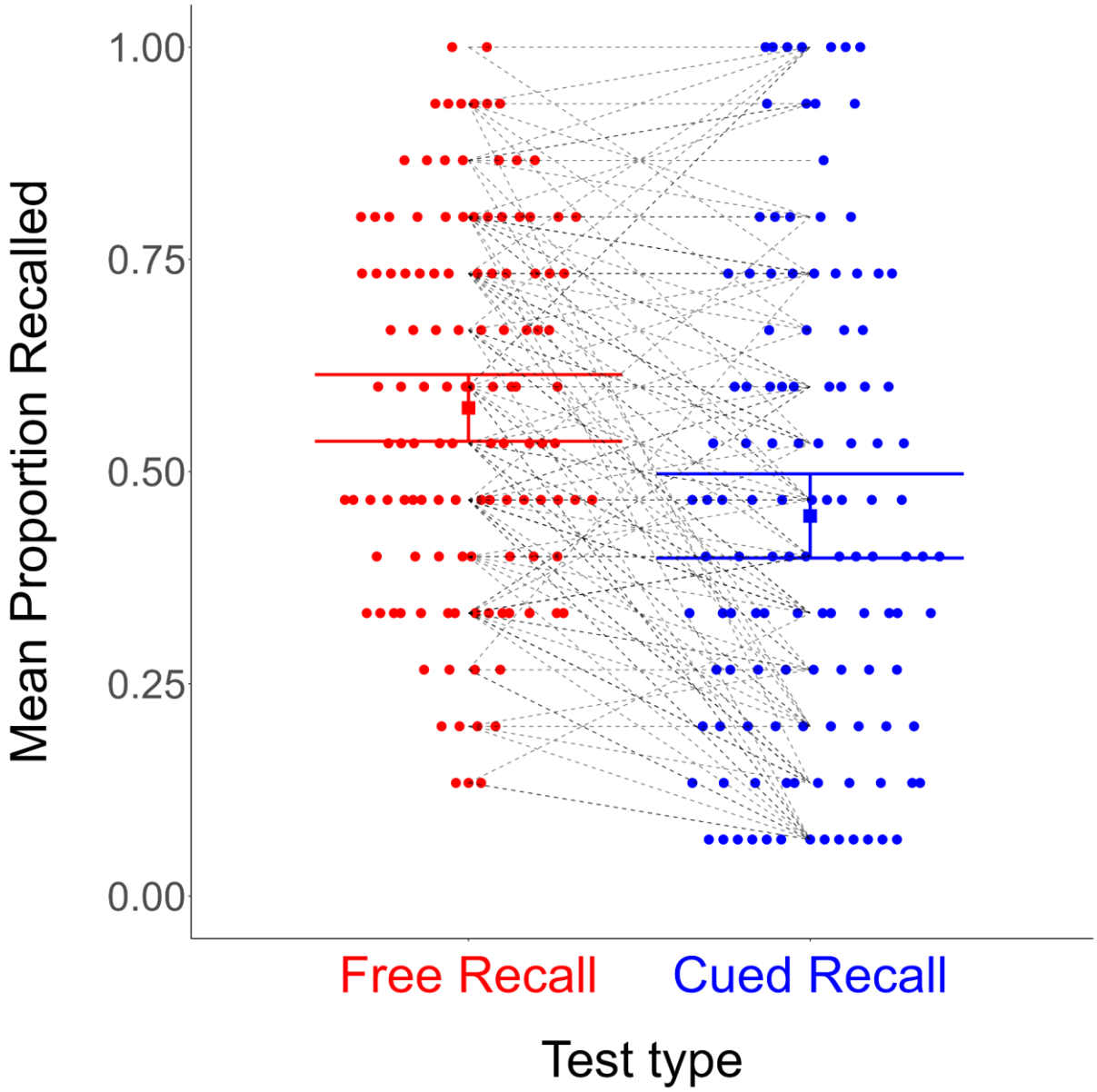
Data files and analysis scripts for Experiments 2A and 2B are available at https://osf.io/yv3b7/.

Confirmatory analyses.

Experiment 2A. We first compared FR and CR variability as we did in the previous experiments–treating only target responses to the matching cue as correct. Memory performance by test type is shown in Figure 16:

Figure 16

Experiment 2A: Memory performance as a function of recall test type



Note. Error bars = 95% CIs (between-subjects). Points jittered horizontally based on relative frequency. Dashed lines connect FR and CR performance for individual participants.

Variability was significantly higher for CR proportion correct than FR proportion correct, Pitman-Morgan $t(118) = 3.10$, $p = .002$, with a bootstrapped CR:FR variance ratio of

1.27, (95% percentile bootstrap CI [1.10, 1.45]).[10] This ratio is slightly smaller than that observed in Experiment 1 (by about 5%).[11] The generalized mixed-effects logistic regression also provided evidence for a CR:FR variance difference (see Supplementary Material 4C). The results were largely similar when treating CR responses as correct if they matched any studied target (i.e., treating the CR task like an FR one, see Supplementary Material 4B). It appears that forced recall reduced but did not eliminate the variability difference. This suggests that instructional ambiguity or variability in interpretations of the CR task only partly account for the greater inter-individual variability in CR performance relative to FR performance.

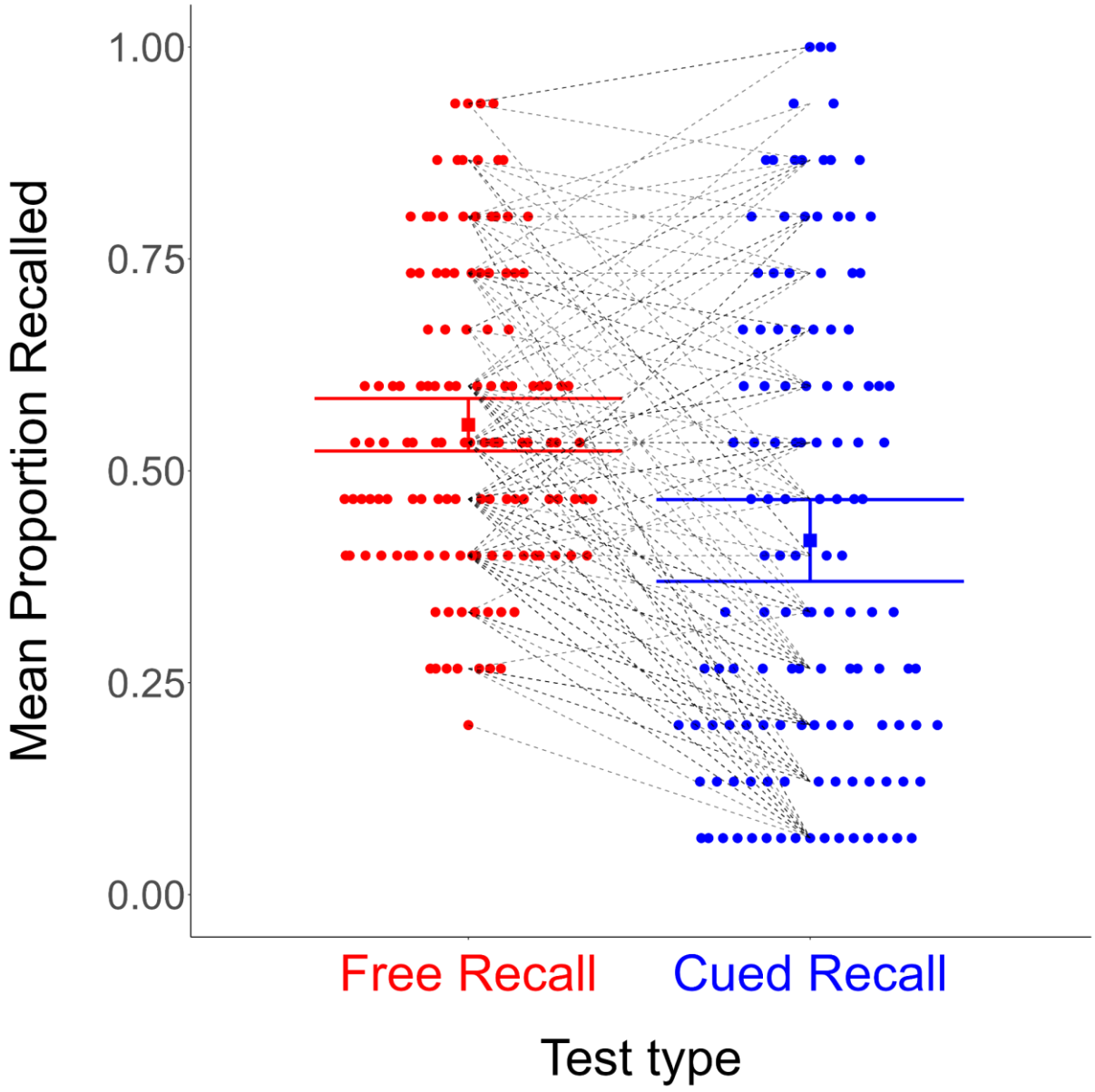
Experiment 2B. We conducted the same analyses for Experiment 2B, in our undergraduate sample. Memory performance as a function of recall test type (only treating targets recalled with matching cues as correct) is shown in Figure 17.

[10] Results differed as a function of test order–the Pitman-Morgan test was significant for those that did FR before CR, but not for those who did CR before FR. This may be due to slightly lower CR performance in the latter group constraining CR variance (see Supplementary Material 4D).

[11] Perhaps this is why the corresponding Bayesian analysis did not provide compelling evidence for a CR:FR difference (see Supplementary Material 4A).

Figure 17

Experiment 2B: Memory performance as a function of recall test type



Note. Error bars = 95% CIs (between-subjects). Points jittered horizontally based on relative frequency. Dashed lines connect FR and CR performance for individual participants.

Here, the variability difference is apparent and striking, and confirmed via Pitman-Morgan test, $t(125) = 5.76$, $p < .001$. The bootstrapped CR:FR variance ratio was 1.57 (95%

percentile bootstrap CI [1.37, 1.78]), larger than the ratio observed in Experiment 1 by 16%.[12] The corresponding generalized mixed-effects logistic regressions also provided clear evidence of a CR:FR variance difference (See Supplementary Material 5C).[13] These results held when treating CR responses as correct if they matched any studied target. (See Supplementary Material 5B).

Why the difference in findings between Experiments 2A and 2B? One thing that is apparent from Figures 16 and 17 is the greater FR variability in the Prolific sample relative to the undergraduate sample. There are myriad reasons why this difference might exist (e.g., Prolific draws from a broader population; students may have more homogenous FR strategies), but the important point is that higher FR variability necessarily constrains the CR:FR variance ratio one can observe. Thus, it may be that in Experiment 2A the reduction in the variability effect is due not to our forced recall manipulation but instead to sample characteristics. This explanation makes sense, especially in light of the *increased* variability effect in Experiment 2B relative to Experiment 1. Although further investigation of the sample differences is beyond the scope of this paper, these differences further support the value of examining differences not only in means, but in variability between conditions. Overall, our confirmatory analyses for Experiments 2A and 2B suggest that the CR:FR variability effect cannot be explained purely in terms of instructional ambiguity or differences

[12] Results were generally similar when comparing those who did CR before FR and vice versa, though the variability difference and CR accuracy were greater/higher for those who did FR before CR (see Supplementary Material 5D). It is possible that doing FR first (easier task) better prepares participants for CR, and this increase in accuracy (i.e., off of CR floor) serves to increase CR variability.

[13] Results were nearly identical when excluding the excess seven participants above our target *N*. Specifically, the Pitman-Morgan $p < .001$, bootstrapped CR:FR variance ratio = 1.54 (95% percentile bootstrap CI [1.34, 1.77]).

in the ways participants interpreted the CR task.

As with the previous experiment, we conducted several exploratory analyses. We examined qualitative self-report strategy data and found a non-significant difference in variability in strategy use, although variability was directionally greater for CR than FR (See Supplementary Material 4E (Experiment 2A) and 5E (Experiment 2B)). We also looked at self-reported recall difficulty, but again failed to find compelling evidence for a variability difference (Significant in Experiment 2B but not Experiment 2A; see Supplementary Material 4F and 5F). Finally, due to the "forced recall" nature of the task, we were interested in a) the frequency of repeated responses, b) participant self-reports of repeating answers, and c) participant self-reports withholding a cue as an answer because they remembered it but weren't sure that it matched the current tested target. In both experiments, the vast majority of participants did not repeat answers on the FR test (See Supplementary Material 4Ha. and 5Ha.). For CR, in both experiments the modal number of repeats was zero, but the majority of participants repeated at least one response (See Supplementary Material 4Hb. and 5Hb.). The majority of CR repeats were studied targets (71.1% in Experiment 2A, 57.4% in Experiment 2B). These results largely matched the self-report data (See Supplementary Material 4G and 5G), but are not particularly illuminating.

3. Experiment 3: "Highly-related DRM words"

Another possible explanation for the CR:FR variability difference lies in the design of our previous experiments. In all of our experiments, our cued recall task involved randomly paired cues and targets nouns. In CR, the strength of cue-target associations is a powerful determinant of cue effectiveness (Cleary, 2018). Although the free recall lists were similarly randomly constructed, and within-list relatedness is associated with free recall performance (e.g., semantic clustering; Cleary, 2018), it is possible that cue-target relatedness matters more for CR performance than within-list relatedness matters for FR performance. In our

examination of computational models of memory, it seemed to be the developed associations that distinguished hypothesized FR and CR processes. The fact that participants performed worse on CR than FR in all our experiments suggests that it was difficult to form effective cue-target associations with our materials. In the qualitative self-reports of study strategies, participants often reported making imaginative or unorthodox associations to connect otherwise unrelated cue-target pairs (e.g., "I tried to find the connection between the words or make a little story. E.g., Somersaulting through a pasture…", "I tried to make an association between the two words. Like "the spoon in the guitar"..."). It could be that if CR pairs and FR lists are constructed such that words are meaningfully related, associative CR strategies would become more homogenous across participants and the variability effect would disappear. To test this possibility, in Experiment 3 we used a new, fixed set of word pairs with cues and targets meaningfully related to one another but not to other cues and targets. We did not have a firm prediction as to whether the effect would persist or disappear with related word pairs, but preregistered our experiment design and analyses (https://osf.io/v67gy).

Methods

Materials. In this experiment, we drew our words and word pairs from normed DRM (Deese-Roediger-McDermott) word lists (Roediger et al., 2001). Specifically, we chose 20 DRM critical lures to serve as targets. We then determined via piloting that using the corresponding probes with the 10th strongest backwards associative strength (as measured by Roediger et al., 2001) as cues resulted in CR performance away from floor and ceiling. We chose another four word pairs to serve as primacy/recency buffers (two primacy, two recency). For FR, we simply used the lures/targets from each CR pair. The final wordset (along with more details about the word selection process) can be viewed in our preregistration for this experiment (https://osf.io/v67gy).

Procedure. In Experiment 3, participants were randomly assigned to complete either a single CR study-test cycle consisting of 20 tested word pairs (+ four primacy/recency buffers) or a single FR study-test cycle consisting of 20 tested words (+ buffers). Although all participants received the same 20 words/pairs and the primacy/recency buffers always appeared in the same position, the order of the tested words/pairs was randomized for each participant. Like the previous experiments, each word/pair was displayed on-screen for 5s during the study phase, and participants were given unlimited time for the test phase. We chose to use a between-subjects design because a) it reduced the possibility of order effects, b) we had a smaller word pool, and c) we observed the CR:FR variability difference when performing "between-subjects" analyses in our other datasets (i.e., comparing CR and FR variability on only the first list that each participant completed). We did not include qualitative strategy questions or ratings of CR/FR difficulty–both to keep the experiment length short and because these questions did not yield particularly enlightening data in previous experiments.

Sample. Via power simulations based on the between-subjects effect observed in Experiment 1, we determined that an *N* of 260 would be sufficient to detect a variability difference of that magnitude with a power of .80. Although we were agnostic in our hypotheses, we were only interested in whether variability was greater for CR than for FR. Thus, our power simulations (and subsequent analyses) used one-sided tests (CR > FR). We collected data from *N* = 306 Prolific participants, from which we excluded (based on preregistered criteria): 15 participants who didn't get at least three correct on both lists, 19 who didn't report understanding at least 75% of words, 2 who reported a major distraction, 1 who reported cheating, and 14 who reported completing a prior version of the study (i.e., on mTurk). Our final sample included 260 participants aged 18-84 (*M* = 38.57, *SD* = 13.78). Participants received $3 USD for participating. Unlike in previous experiments, we did not

manually code commission errors (due to the small proportion of responses that coding affected in previous experiments).

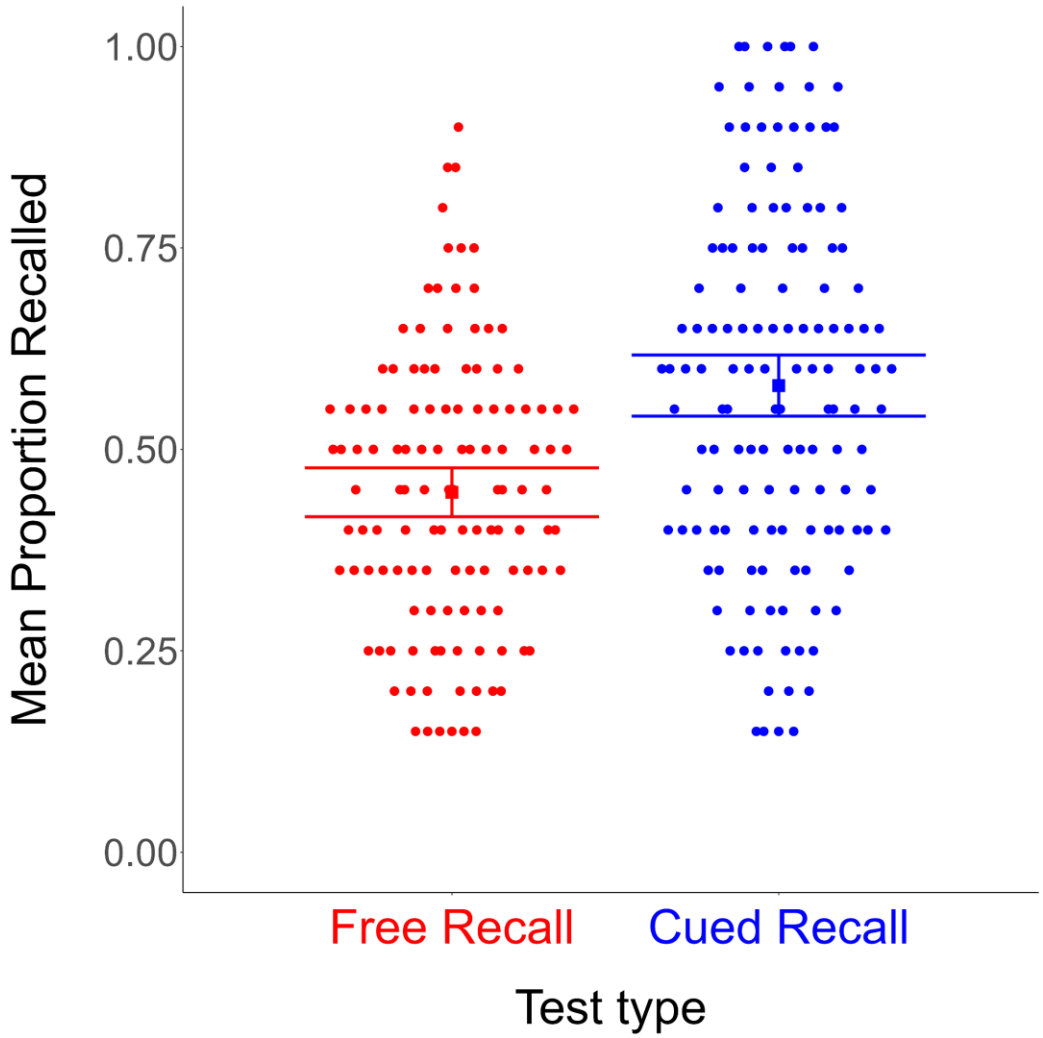
Results & Discussion

Data files and analysis scripts for Experiment 3 are available at https://osf.io/pfhu9/.

As in previous experiments, we compared variability in CR and FR memory performance. Memory performance by test type is shown in Figure 18.

Figure 18

Experiment 3: Memory performance as a function of recall test type



Note. Error bars = 95% CIs (between-subjects). Points jittered horizontally based on relative frequency.

To compare variability *between*-subjects, one cannot use the Pitman-Morgan test, but there are several alternative options (e.g., Levene's, Bartlett's, O'Brien's; Hatchavanich, 2014; Othman et al., 2007). Hatchavanich found that the tests did not differ in terms of robustness and power (2014), and Othman et al. (2007) report that the popular Levene's test is robust to nonnormality, so we opted to use the widely-used Levene's test as our primary analysis of between-subjects variability (although we again supplemented this test with a nonparametric bootstrap of the variance ratio). Via our sole preregistered confirmatory analysis, variability was significantly higher for CR proportion correct than FR proportion correct, Levene's $F(1) = 10.81$, $p = .001$, with a bootstrapped CR:FR variance ratio of 1.33, (95% percentile bootstrap CI [1.14, 1.54]). This ratio is nearly identical to that observed in Experiment 1. Thus, participants differed more from one another on CR than FR even when CR pairs are meaningfully related. This might imply that the CR:FR variability difference is not due to variation in relatedness (or a lack of relatedness) of the CR cues and targets we used in previous experiments. This experiment was also the first in which we observed higher average performance for CR relative to FR. The fact that we have observed higher CR than FR variability when CR < FR performance and when CR > FR performance suggests that the variability effect is not confounded with levels of performance. In this experiment in particular, CR and FR performance were both quite close to .50, allowing ample room to vary in either direction.

4. Experiment 4: "Equivalent study lists"

In the following two experiments, we considered possible methodological explanations for the CR:FR variance effect. In all prior experiments participants studied twice as many CR as FR words. This difference in encoding could have contributed to the observed effects (e.g., participants vary more for longer study lists). To address this possibility, we conducted an experiment similar in design to Cox et al. (2018)—all participants studied word pairs, then

were tested on FR (for all words) or CR (for targets in response to presented cues). We predicted that the CR:FR variability effect would persist even when equating the number of studied words, and preregistered our materials, hypotheses, and analyses: https://osf.io/de7bu

Methods

Materials. In an attempt to obtain accuracy levels that were not too high or low, we pilot tested a number of wordsets (see the wiki page at https://osf.io/tnspg/wiki/home/). Ultimately, we chose a pool of 80 English object words (that we had used in prior animacy experiments). The complete wordset can be viewed at the link above.

Procedure. Like Cox et al. (2018), all participants studied word pairs. Memory test type (CR, FR) was between-subjects. Unlike Cox et al., participants assigned to the CR condition were told *before* study that they would complete a CR test (and vice versa for participants assigned to the FR condition). We hoped that providing instructions at study would increase performance on the subsequent test while otherwise keeping encoding as similar as possible across conditions. Participants studied a single list of seven test pairs (randomly sampled from the overall pool) for 5s each, with one fixed primary and one fixed recency buffer pair. Cues were color-coded black and targets were color-coded red; FR participants were told they would later have to recall as many black and red words as possible, while CR participants were told they would be presented with the black words in a random order and would have to recall the corresponding red word. The experiment was programmed in PsychoPy and administered online via Pavlovia.

Sample. Via power simulations, we set a post-exclusion target $N = 360$. We collected data from a total of $N = 492$ Prolific.co participants and based on preregistered criteria excluded: 82 participants who did not get at least one (two) correct on CR (FR), 52 participants who reported understanding less than 75% of the words, 19 who mentioned completing a previous version of the study (e.g., on mTurk), 2 who reported a major

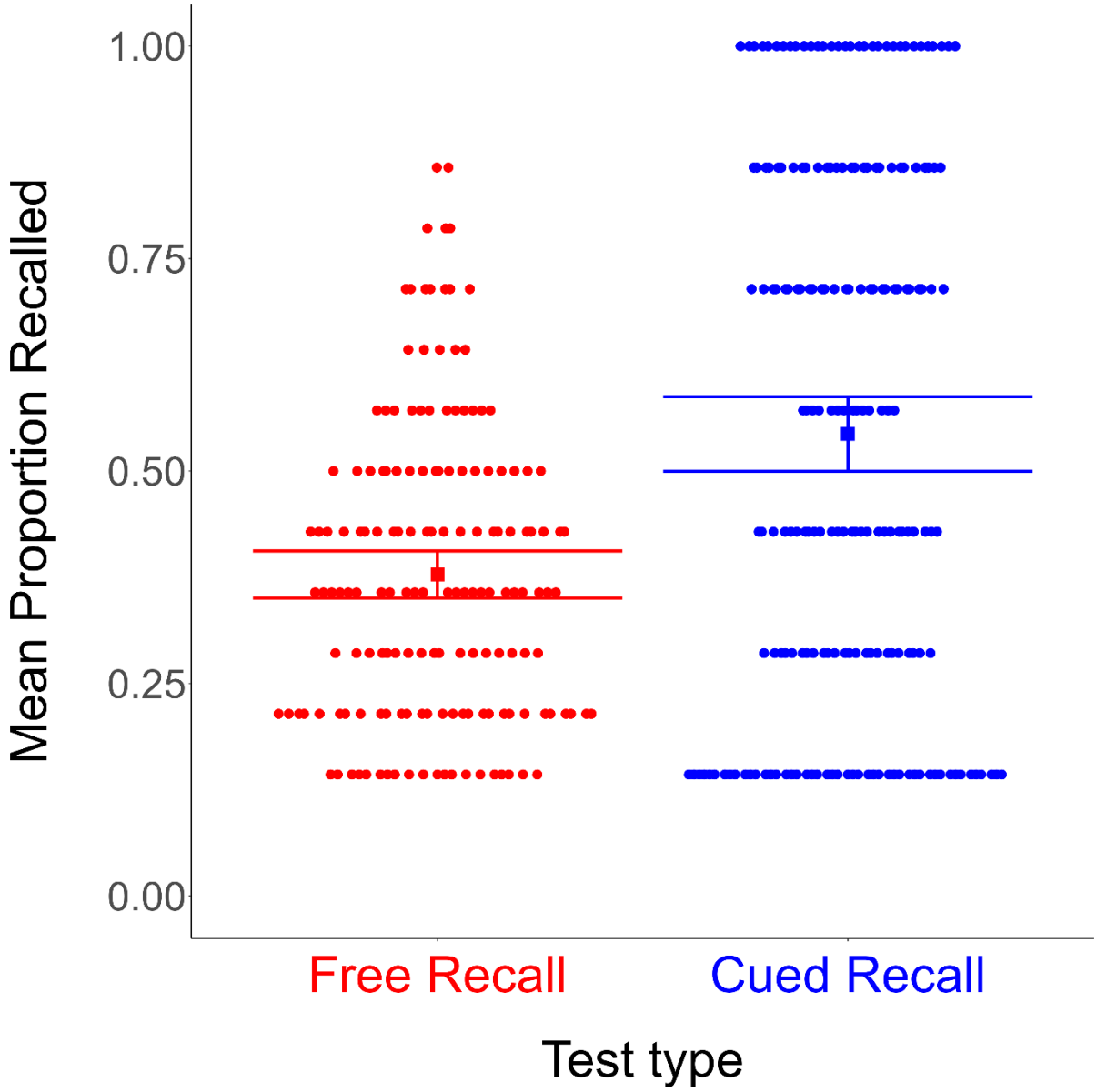
distraction, and 1 who reported cheating. Our final sample included 360 participants aged 18-72 (M = 36.64, SD = 12.06).

Results & Discussion

First, a visualization of CR and FR performance:

Figure 19

Experiment 4: Memory performance as a function of recall test type



Note. Error bars = 95% CIs (between-subjects). Points jittered horizontally based on relative frequency.

Although the variability difference is immediately apparent, one must account for the fact that the number of *tested* words differed for CR and FR. Specifically, accuracy on the final test was out of 14 for FR and 7 for CR. Via simulations (see the wiki page at https://osf.io/tnspg/wiki/home/), we determined that given equal underlying variance, traditional variability analyses (Levene's test, calculation of the variance ratio) can falsely attribute greater variability to the measure computed from fewer items (in this case, CR).

To account for this "variance ratio inflation", we preregistered and conducted two main binomial logistic GLMM analyses (that did not show this inflation in simulations) at the item-level. In the first, to account for potential inflation of the CR:FR variance ratio due to lower FR performance we simulated new data with the same means but equal underlying variances to obtain a "null" CR:FR variance ratio to compare against. Fortuitously, this "null" ratio was close to 1 (.98), and did not show evidence of variance inflation. The observed CR:FR variance ratio was much higher: 2.38 [bootstrapped 95% CI: 1.95, 2.89]. In the second analysis, we compared participant-level variability estimates (i.e., random-intercepts variance) using profiled 90% CIs[14]. This analysis showed that CR variability (SD_{logit} = 1.43, 90% CI [1.24, 1.65]) was reliably greater than FR variability (SD_{logit} = .54, 90% CI [.43, .65]). Finally, as a rough exploratory comparison to previous experiments, we computed the CR:FR variance ratio traditionally (i.e., on proportion correct), and compared to a simulated "null" variance ratio accounting for inflation due to differing *n*-items. The null ratio was 1.27, with an observed ratio of 1.79 [bootstrapped 95% CI: 1.61, 2.01]. Adjusting for the null, this ratio was 1.41—slightly larger than that observed in prior experiments. Together, these analyses provide evidence that the differing number of studied words does not account for the CR:FR variability difference.

[14] 90% because our hypothesis of CR > FR variability was one-sided.

5. Experiment 5: "Self-paced study"

The second additional methodological possibility that we considered was study time. In all prior experiments, participants were given 5s to study FR singletons and CR pairs. It is possible, for instance, that 5s is sufficient for most participants to read and encode single words, but insufficient for some participants to read/encode CR pairs. Thus, the CR:FR variability difference observed thus far could be due to a factor(s) unrelated to memory that participants varied on, such as reading speed. To address this possibility, we conducted an experiment in which the study phases were self-paced. Theoretically, self-paced study phases should also reduce the influence of individual differences in study strategies, as all participants would ideally study each word/pair to a threshold of perceived memorability (although differences in this threshold could still contribute to variability in performance). We predicted that the CR:FR variability effect would persist even when participants could study at their own pace, and preregistered our materials, hypotheses, and analyses: https://osf.io/my53w

Methods

Materials. We used the wordset (83 nouns) from Experiments 2A and 2B (see https://osf.io/xdsk8/.

Procedure. We replicated the design of Experiment 1, with each participant completed one CR and one FR study-test cycle (order counterbalanced, 15 words/pairs per list), except that the study phase was self-paced. That is, each participant had up to 30s to study each word/pair, and could proceed to the next word/pair by pressing the space bar. After 30s, the current word/pair disappeared, displaying a blank screen until the participant

pressed space to continue[15]. The experiment was programmed in PsychoPy (see https://osf.io/xdsk8) and administered online via Pavlovia.

Sample. We adopted the same target sample size as in Experiment 1, $N = 120$. We collected data from a total of $N = 182$ undergraduate participants and based on preregistered criteria excluded: 20 participants who did not get at least one correct on both lists, 27 participants who reported understanding less than 75% of the words, 13 who reported a major distraction, 11 who reported a substantial technical difficulty, and 4 who reported cheating. Our final sample included 124 participants aged 17-48 ($M = 20.27$, $SD = 3.40$)[16], with 68 completing CR first and 56 completing FR first.

Results & Discussion

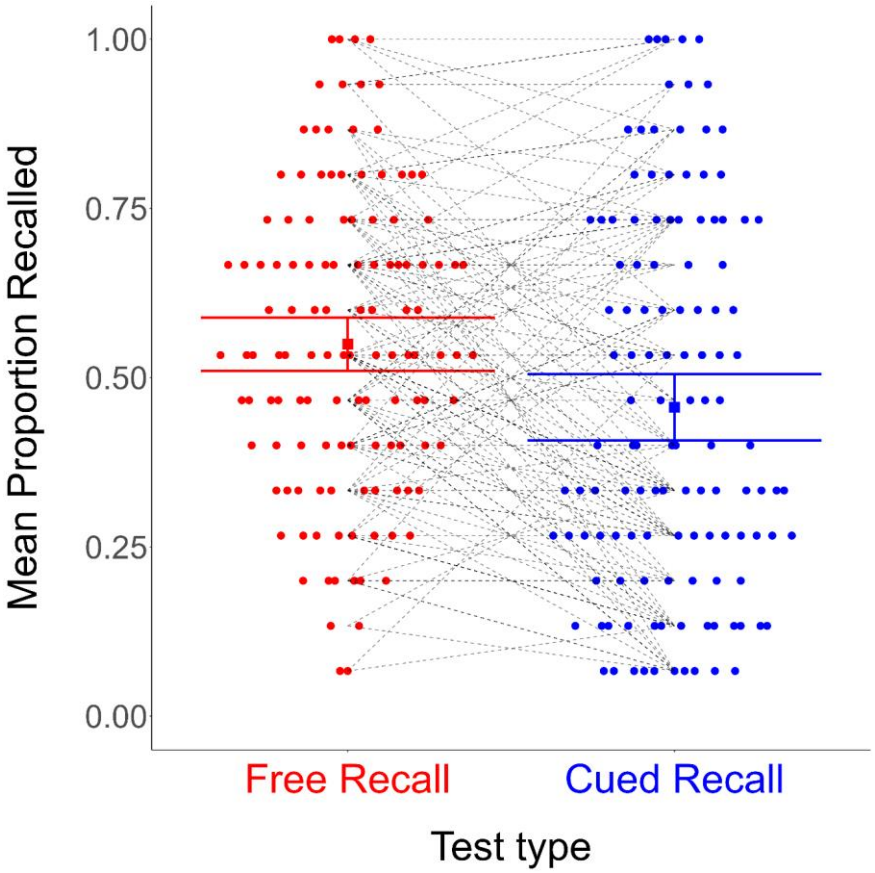
Recall accuracy. Figure 20 shows proportion correct for FR and CR when the study phases were self-paced.

[15] We intended for the experiment program to auto-advance to the next word/pair after 30s, but did not detect the programming error leading to the design reported in-text until after data had been collected. Study trials >30s were rare (~4.5% of all trials), and excluding these trials *or* participants who had more than one such trial ($n = 20$) *or* participants who had any such trials ($n = 33$) did not change the results of our primary analysis. So, for subsequent analyses we did not exclude any trials/participants on this basis.

[16] The inclusion/exclusion of the additional four participants above our preregistered target N did not change the results of our primary confirmatory analysis, so they were included for all subsequent analyses.

Figure 20

Experiment 5: Memory performance as a function of recall test type



Note. Error bars = 95% CIs (between-subjects). Points jittered horizontally based on relative frequency. Dashed lines connect FR and CR performance for individual participants.

The corresponding Pitman-Morgan test was significant, $t(122) = 2.66$, $p = .009$ (when restricting to $N = 120$, $p = .01$; when excluding participants with more than one study trial longer than 30s, $p < .001$; when excluding participants with *any* study trials longer than 30s, $p < .001$). The bootstrapped CR:FR variance ratio was 1.26 [95% CI: 1.09, 1.42], slightly lower than the ratio observed in Experiment 1 (1.35). Results were similar for those who completed CR first, but in the group that completed FR first the variance difference was non-significant (see SOM 7A). As the variance difference was directionally in favour of CR, we suggest that the lack of significance in the FR-first group is due to a combination of a) reduced power and

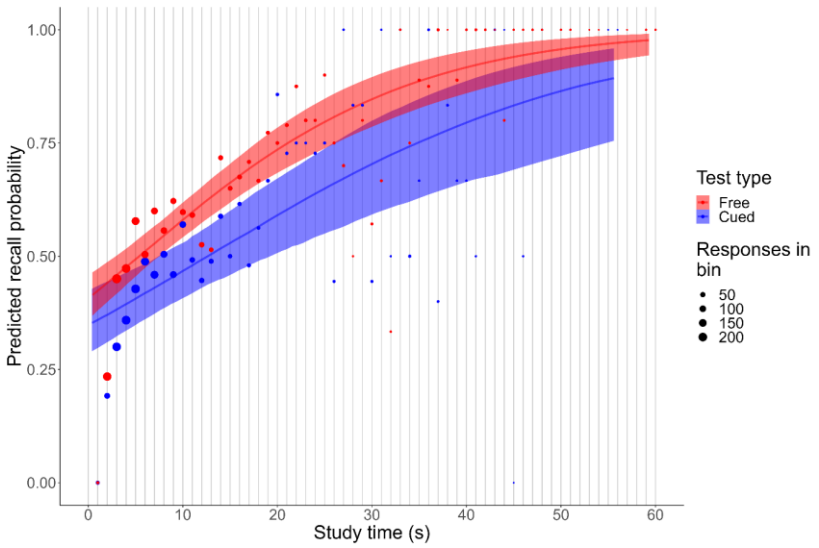
b) lower CR performance restricting CR variance. Thus, the CR:FR variability difference persisted even when participants could study at their own pace.

Study time. We were also interested in differences in study time and the relationship between study time and accuracy. Study time was similar for FR singletons. *M* = 9.56, *Median* = 6.05, *SD* = 12.78, and CR pairs, *M* = 9.2, *Median* = 6.29, *SD* = 12.18 (see SOM 7B for a density plot). Importantly, an exploratory Pitman-Morgan test of variance in study time was not significant, $t(122) = 1.62$, $p = .11$, bootstrapped CR:FR variance ratio = .92 [95% CI: .61, 1.32]. That mean/median study times were longer than the fixed 5s used in previous experiments suggests that the fixed time may have constrained performance somewhat.

Did study time predict accuracy, and if so, was this relationship different for FR and CR? Figure 21 shows accuracy (empirical, model-predicted) as a function of study time.

Figure 21

Experiment 5: Memory performance as a function of recall test type



Note. Points = accuracy averaged for each 1s bin (size = relative frequency of responses in each bin). Lines and ribbons = Bayesian[17] GLMM posterior means and 95% credible intervals. 17 trials with study time > 60s were removed.

[17] Bayesian model used due to difficulties in obtaining predictions from the NHST model.

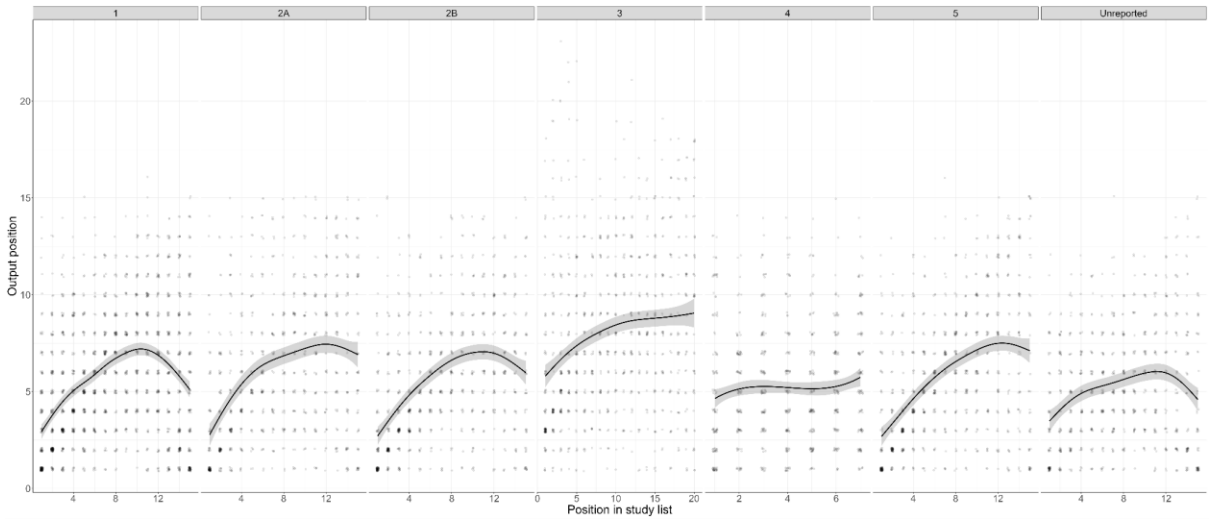
A GLMM predicting item-level binomial accuracy (0/1) from study time and test type (with random intercepts and test type effects by participant) revealed a significant effect of study time, $\chi^2(1) = 82.65$, $p < .001$, significantly lower CR than FR accuracy, $\chi^2(1) = 13.57$, $p < .001$, but no interaction between study time and test type, $\chi^2(1) = 2.55$, $p = .11$. These results did not change when excluding study times longer than 30s (ps < .001, < .001, and .54, respectively). The positive (and similar) relationship between study time and accuracy for both test types is not particularly surprising, though it may be noteworthy that at least descriptively, the CR variability advantage was present at all study times.

6. Experiment 6: "Serial cued recall"

One additional possible explanation for the CR:FR variability effect is differences in output order. That is, in all previous experiments, FR allowed participants to recall targets in any order they desired. In an exploratory analysis of FR output order (as a function of study list position) in all previous experiments:

Figure 22

Previous experiments: FR output order as a function of study order

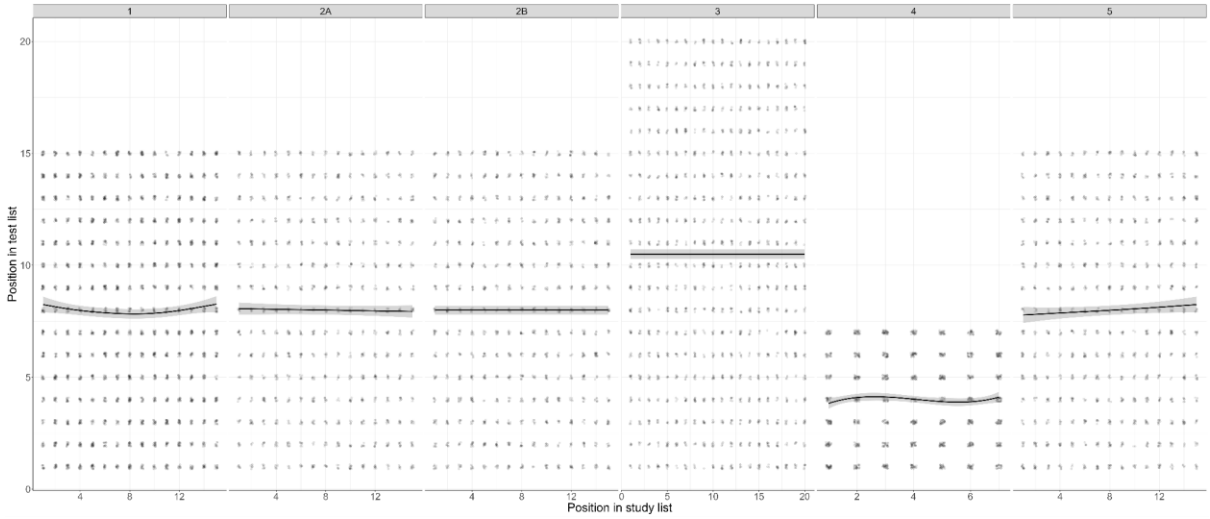


Note. Points represent individual items (added jitter). Lines and ribbons = multilevel spline regressions & 95% CIs.

In almost all experiments, output patterns followed the classic ‘serial position curve’ (Deese, 1957), with early-list items recalled first, late-list items recalled next, and middle-list items recalled afterwards. On the contrary, CR 'output order' (i.e., the order in which cues were presented) was by the nature of our experimental designs random:

Figure 23

Previous experiments: CR output order as a function of study order



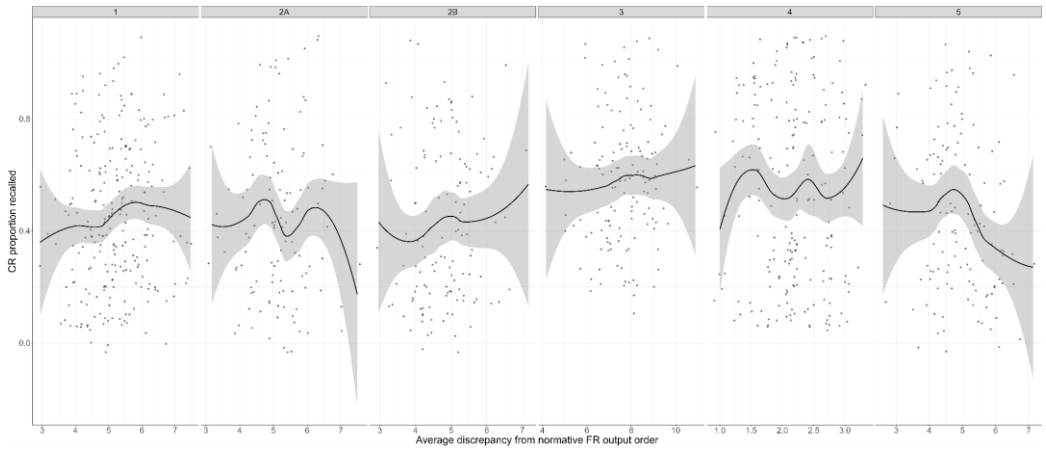
Note. Points represent individual items (added jitter). Lines and ribbons = multilevel spline regressions & 95% CIs.

During FR, participants were free to choose their output order (and often adopted a relatively homogenous pattern), whereas for CR they were forced to adopt whatever random order the experimental program assigned them. For some participants, this order could be close to their preferred output order (e.g., the serial position curve). But for other participants, random assignment could have left them with an order maximally different from their preferred order. In other words, for FR the variability in ‘discrepancy from preferred recall order’ would be zero–all participants should recall targets in their preferred order. But for CR, there would be variability in this discrepancy. This potential source of variability–present for CR but not for FR–might explain the greater variability in CR performance.

In an exploratory analysis of the six experiments in which participants completed both FR and CR, we obtained a "normative" FR recall order (i.e., the most common recall position as a function of test position). For each subject, we then computed the average discrepancy between this order and the random CR test order they received. Participants with a low discrepancy had (by chance) CR test orders that were closer to what one might consider an 'optimal' or 'preferred' FR serial recall order. We then predicted overall CR accuracy from this discrepancy, reasoning that if recall order played a role in accuracy (and variability thereof), discrepancy should show a negative relationship with accuracy. We did not find this to be the case:

Figure 24

CR accuracy as a function of discrepancy from normative output order



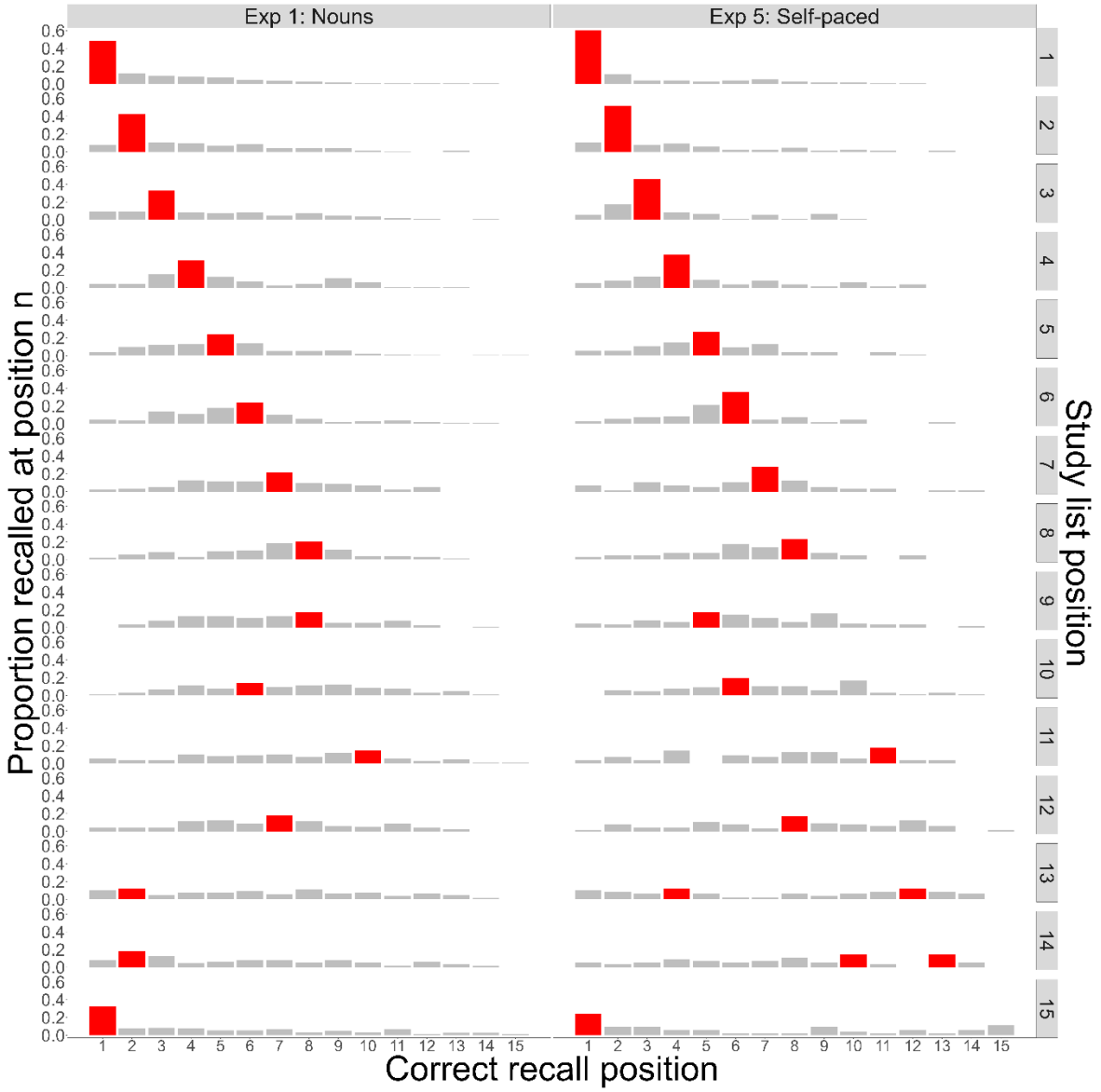
Note. Points represent individual participants. Lines and ribbons = multilevel spline regressions & 95% CIs.

We did not observe a consistent relationship between discrepancy from a normative FR output order and CR accuracy. In other words, participants who by chance ended up with a CR test list that approximated the serial position curve did not appear to perform better than participants who by chance ended up with a CR test list quite different from the optimal serial position curve.

A stronger and more conclusive test would be to experimentally manipulate CR output order. Thus, we looked at the "normative" FR output order in two of our experiments with the most typical FR tasks:

Figure 25

Experiments 1 & 5: Normative FR output positions for words at each study position



Note. This figure shows the proportion of times words at each position in the study list (y-axis) were correctly recalled at each position at test (x-axis), with the red bars denoting the most frequent recall position(s), computed as the position with the highest proportion of correct recall for a given study position.

The most common positions followed a serial recall order, with typical recency effects. Recall was typically initiated with either the first or last couple words studied, then proceeded serially. Based on this, we designed Experiment 6, in which the order of CR cues at test exactly matched the order of CR pairs at study[18]. Although we considered the possibility that output order might explain the CR:FR variability effect, we predicted that this manipulation would *not* eliminate the effect.

Methods

Materials. Based on the Experiment 5 data, we pruned the wordset from 83 to 60 words, cutting words for which FR performance was low, and words we judged subjectively that participants might be unlikely to know (see https://osf.io/6gbh2 for the reduced wordlist).

Procedure. Participants completed one FR *or* one CR study-test cycle consisting of 15 words (FR) or 15 word-pairs (CR). Like the previous experiment, this experiment included self-paced study (up to 30s per word/pair). Unlike the previous experiment, we introduced a brief 'familiarization phase' where participants completed a sample FR or CR test cycle of 5 words (pairs) before proceeding to the main study-test cycle. Our reasoning was that some of the increased CR variability might be due to participants being less familiar with the CR task, and that a familiarization phase might reduce some of this nuisance variability. We also introduced a battery of more fine-grained and quantitative questions about study strategies (adapted from Morrison et al., 2016) in an attempt to revisit the possibility that variability in strategies lay behind the CR:FR variability effect. Specifically, participants were asked to rate the frequency with which they used six strategies (Rehearsal, Semantic, Grouping, Imagery, Sentence/Story, Other) on a 5-point Likert scale. Each strategy

[18] Of course, we could have adopted a more "optimal" order, i.e., first present the first pair studied, then the last pair studied, then the second pair studied, so on, so forth. However, it was simpler to both implement (and explain) a serial recall order to participants. And in any case, a *consistent* recall order (regardless of optimality) should still eliminate the variability effect if the explanation for the effect lies with the randomly generated CR test orders.

was described to participants, and strategies were either queried after study ("How often did you use each of the following strategies when *studying* the words/pairs") or after test ("How often did you use each of the following strategies when *trying to remember* the words/pairs"). Because we queried strategies (and provided explicit examples), we chose to keep test type (FR/CR) between-subjects, to avoid priming participants toward particular strategies. The timing of the strategy questions (after study vs. after test) was also counterbalanced between subjects in case exposure to the different strategies prior to test influenced participant behaviour. After the strategy rating questions, participants were also asked to summarize their general strategy use, with the options: "I didn't really have any strategy that I know of, I just tried my best to remember.", "I tried various strategies as I went along, shifting from one to another", "I had a pretty clear strategy and I used it pretty consistently", "I pretty much just went through it without too much thought to get done ASAP". We did not have explicit predictions about study/test strategies (e.g., whether the timing of the questions affects performance, or whether strategies predict accuracy, or whether variability in strategies differs for free and cued recall). Still, our aim was to explore the possibility that self-reported strategies at study and/or test vary more from one subject to another when the task is CR than when it is FR. Finally, we asked the same demographics and exclusion questions as in the previous experiment. The experiment was programmed in PsychoPy (see https://osf.io/gmebn) and administered to undergraduate participants in on-campus computer labs. The move to in-person testing was aimed at reducing some of the potential error variability introduced by online testing (e.g., different computers, test conditions, etc.).

Sample. Based on power simulations using data from the previous experiment (Experiment 5), we determined that we would need an N of 190 (split evenly between FR/CR) to achieve a power of .9 to detect CR > FR variability (one-sided). We aimed for a final post-exclusion sample size of N = 200. We collected data from a total of N = 242

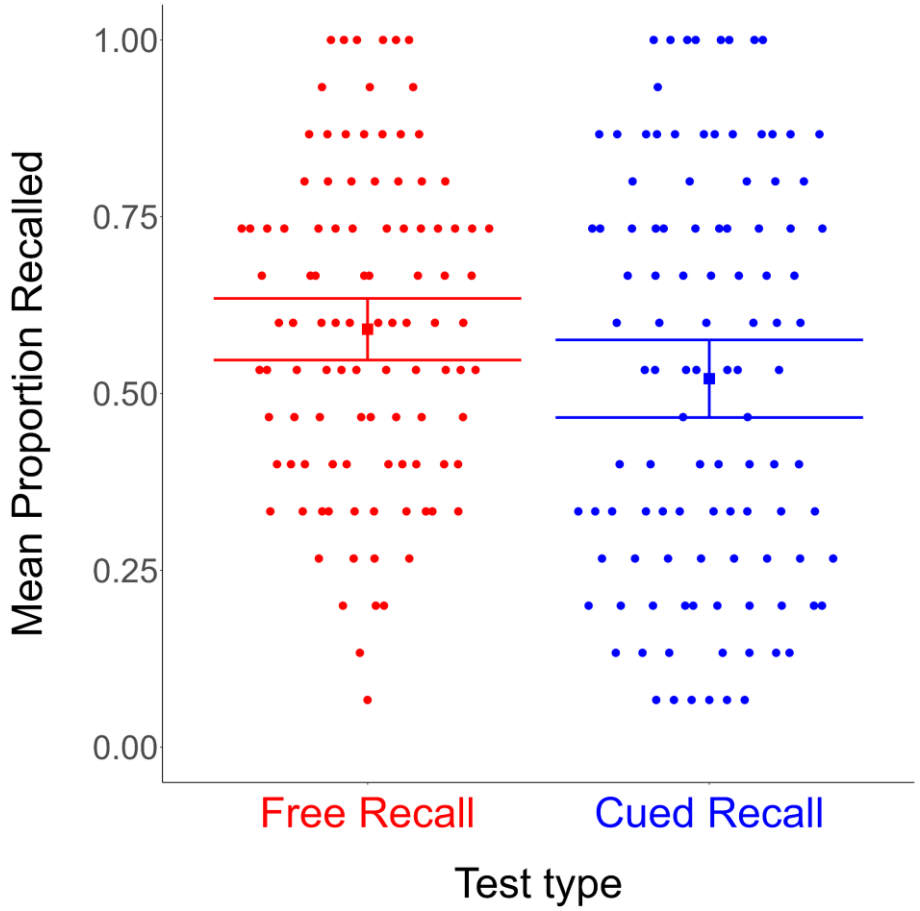
undergraduate participants and based on preregistered criteria excluded: 3 participants who did not get at least one correct at test, 29 participants who reported understanding less than 75% of the words, and 4 who reported a major distraction. Our final sample included 211 participants aged 17-53 (M = 21.30, SD = 5.62), with 108 completing CR, 103 completing FR, 101 participant completing qualitative strategy questions after study, and 110 completing these questions after test.

Results & Discussion

Recall accuracy. Figure 26 shows proportion correct for FR and CR:

Figure 26

Experiment 6: Memory performance as a function of recall test type

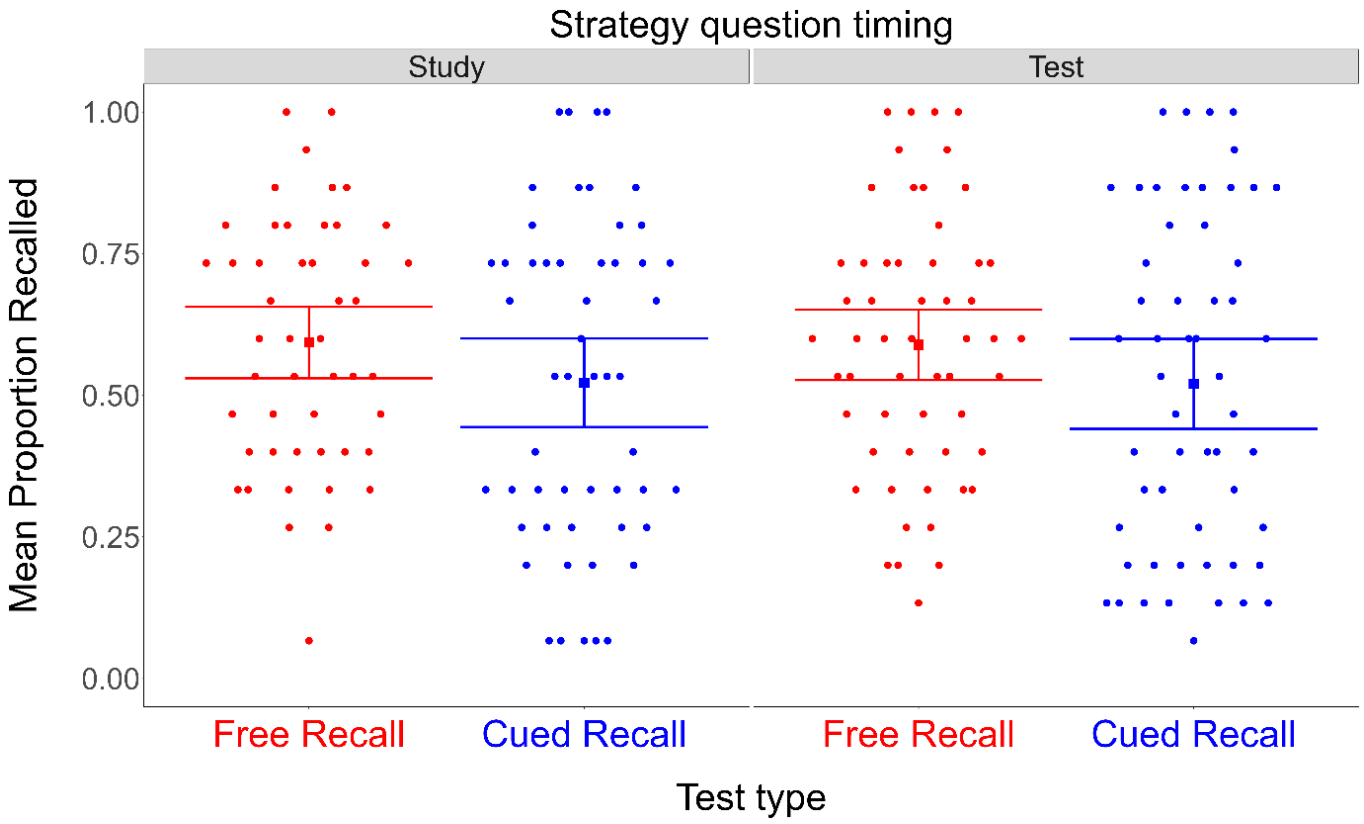


Note. Error bars = 95% CIs (between-subjects). Points jittered horizontally based on relative frequency.

Once again, CR variability was significantly higher than FR variability, via Levene's test, $F(107, 102) = 1.66$, $p = .01$. We supplemented the Levene's test with non-parametric bootstrapping. The bootstrapped ratio of CR:FR variance was 1.29 [95% CI: 1.13, 1.48]. Results were similar when splitting by strategy question timing (strategy questions after study, strategy questions after test):

Figure 27

Experiment 6: Memory performance as a function of recall test type and strategy question timing



Note. Error bars = 95% CIs (between-subjects). Points jittered horizontally based on relative frequency.

Perhaps due to lower power, the variability difference was non-significant for the *after-study group*, $F(52, 47) = 1.71$, $p = .06$. But descriptively, the CR:FR variance ratio was similar to the combined analysis: 1.32 [95% CI: 1.07, 1.61]. For the *after-test* group, the variability difference was again non-significant, $F(54, 54) = 1.63$, $p = .08$. But the bootstrapped ratio was robust and similar in magnitude, 1.28 [95% CI: 1.06, 1.55]. Thus, it's likely that the lack of significance is a power issue and not some substantive effect of strategy question timing on the variability effect.

Memory strategies. Participants were asked two kinds of questions about the strategies they used:

1. The *frequency* with which they used grouping, imagery, rehearsal, semantic, story-based, or other strategies (1 - 5)
2. The *consistency* of their strategy use (i.e., whether they used a consistent strategy, various strategies, no strategies, or just rushed through.

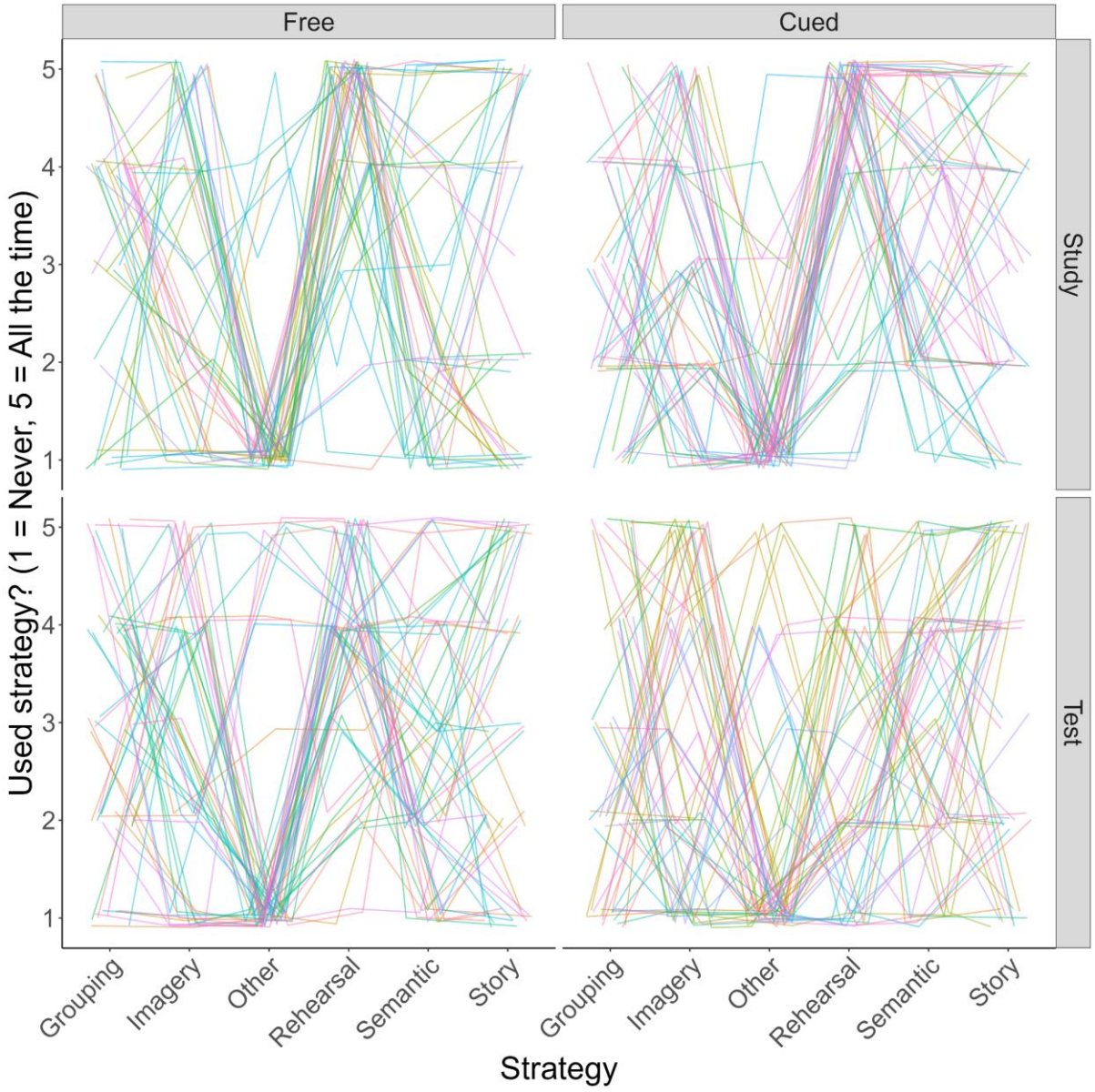
Participants either answered these questions after the study phase, or after the test phase. Although the study/test questions for 1. differed slightly in their wording, they assessed the same general strategies. Our aim with these questions was to determine if there were possible links between strategy use and performance variability. For instance, it could be that participants vary more in the strategies used for CR than for FR.

To investigate this possibility, I first examined what I'll refer to as participant *strategy profiles*. These are essentially the vector of responses to the 6 strategies participants could rate frequencies for. So a participant could have a 5 for rehearsal, a 1 for semantic, a 2 for imagery, etc. . .

Figure 28 plots these vectors as lines, by test type and strategy question timing:

Figure 28

Experiment 6: Strategy profiles as a function of test type and strategy question timing



Note. Lines = individual participants

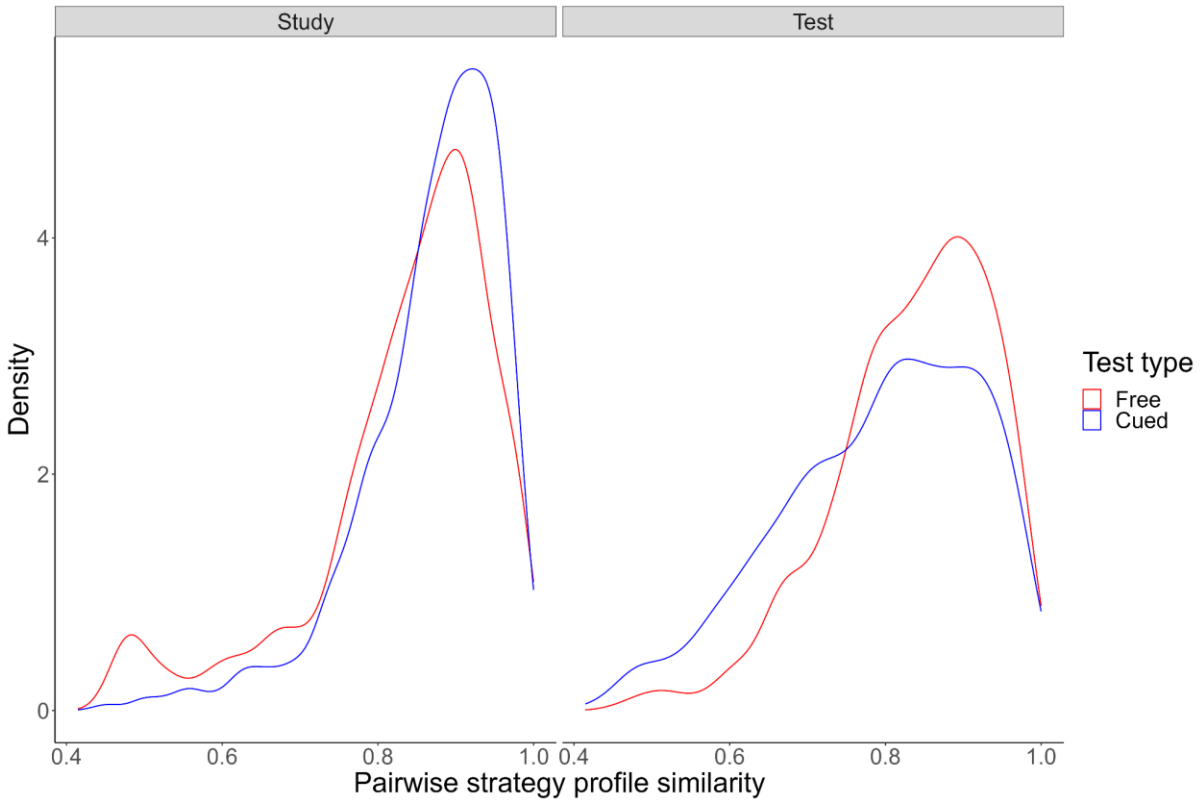
If participants are more similar in terms of the kinds of strategies they adopt, the plots should be 'cleaner', with more overlapping lines. Some consistency can be seen in responses to the questions after study for FR and CR, e.g., a lot of participants did not report using an 'other' strategy, and many reported relying on rehearsal. And at least via the interocular test, the

most variability in strategy profiles seems to be for CR, and when strategies were queried at test.

To formally examine variability in strategy profiles, we examined differences in these vectors across participants. For example, for CR participants who answered strategy questions at test, we computed the cosine similarity between each *CR-test* participant's vector and all other *CR-test* participants' vectors. The logic here was that if strategy use is generally similar, then pairwise similarity should be high. Figure 29 shows the distributions of strategy profile similarity by test type and strategy question timing.

Figure 29

Experiment 6: Strategy profile similarity by test type and strategy question timing



Note. Density plots constructed using all pairwise comparisons among participants in a given test type X strategy question timing cell.

Interestingly, participant *study* strategies appeared to be more consistent for CR than for FR ($p < .001$), but the reverse was true for *test* strategies ($p < .001$), with the magnitude of the differences similar. Of course, it is worth noting that the significance tests were based on some 5,000-odd pairwise comparisons, so it is perhaps not surprising that we detected significant differences. That said, the increased CR variability in test strategies might be playing a role in increased CR variability. Maybe CR performance is more variable because participants choose from a wider range of test strategies, some of which are effective and some of which are not.

Digging deeper, we examined strategy use for the top- and bottom-25% CR participants who answered strategy questions after test (*n*s = 14, 15 respectively):

Figure 30

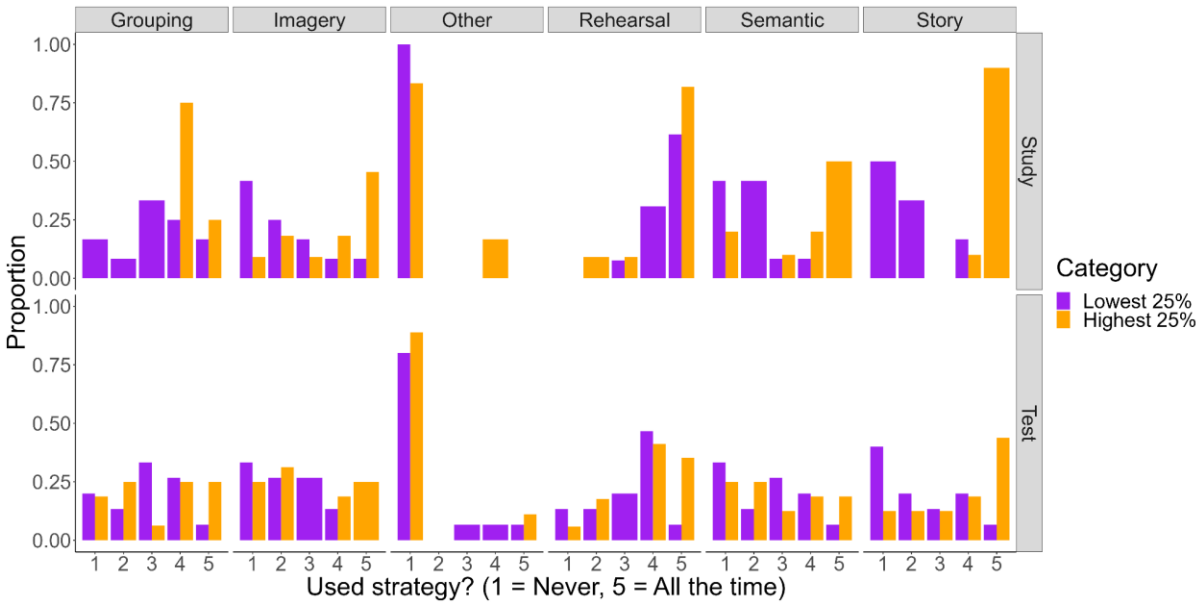
Experiment 6: Test strategies used by top- and bottom-25% performing CR participants

Note. Bars = proportion of participants in each category reporting that frequency of strategy use.

Pairwise comparisons by group revealed that the top 25% were significantly more likely to report using imagery- (p = .005) and story-based strategies (p < .001), and significantly less likely to report using other strategies (p = .008). These results are consistent with prior work showing that participants instructed to use interactive-imagery strategies show improved memory for word pairs (Sahadevan et al., 2021) and that more imageable word pairs tend to be more memorable (Caplan & Madan, 2016). Do we get the same pattern for high- and low-performing FR participants?

Figure 31

Experiment 6: Study and test strategies used by top- and bottom-25% performing FR participants



Note. Bars = proportion of participants in each category reporting that frequency of strategy use.

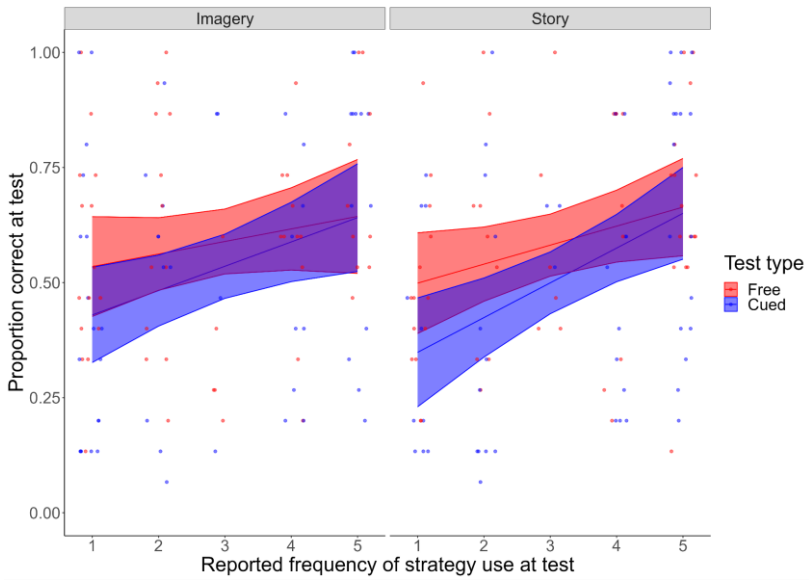
Results were less clear–although the high performers were significantly more likely to report using imagery at study (p = .02), the same was not true at test (p = .44). Additionally,

there were no differences in reports of using other strategies (all *p*s > .57). High performers at study and test were significantly more likely to report using story-based strategies ($p < .001$, $p = .01$). So, it could be that 'non-standard' strategy use is more prevalent in CR than FR, and 'non-standard' strategy users account for greater CR variability. Or it could be that although imagery- and story-based strategies similarly discriminate high- and low-performers on both CR and FR, these strategies matter more for CR.

If this were the case, one might expect that the relationship between reported frequency of use for these strategies and performance is *stronger* for CR than for FR. This can be tested–Figure 32 shows proportion correct (empirical, predicted) as a function of reported frequency of using imagery-based and story-based strategies *at test*:

Figure 32

Experiment 6: Proportion correctly recalled as a function of imagery- and story-based strategy usage at test



Note. Points = individual participants, lines and ribbons = best-fitting linear regression lines & 95% CIs.

Although the interaction between test type and strategy-use frequency was non-significant for both imagery ($F(1) = .63$, $p = .43$) and story ($F(1) = 1.24$, $p = .27$), at least directionally the relationship between these strategies and accuracy appears stronger for CR. Finally, we examined responses to the strategy consistency questions. We did not find any significant differences in self-reports of strategy consistency between FR and CR, at either study or test (all *p*s > .11). This suggests that increased CR variability is not due to, say, individual participants switching between strategies more for CR than for FR.

What to conclude? First, it seems like the CR:FR variability effect is *not* due to the random ordering of CR cues at test (i.e., differential disruption of preferred output order). Second, the strategy data seems to suggest that a) strategies used for CR tests may vary more across participants and/or b) variable use of imagery- and story-based strategies (vs. non-standard 'other' strategies) may explain the greater variability in CR relative to FR because c) although these strategies similarly discriminate high- and low-performers for both test types, imagery- and story-based strategies may matter more for CR than for FR. These three possibilities informed our seventh and final experiment.

7. Experiment 7: "Imageability"

In Experiment 6, we found some evidence that a) participants may have varied more in the strategies they adopted at test for CR, relative to FR, b) high-performers on CR and FR tended to report more use of *imagery* and *story*-based strategies, relative to low-performers, and c) use of these strategies may be more predictive of performance for CR than for FR. Taken together, these results suggest a potential role of imagery and imageability in the effects we have observed thus far. Previous research has found that both the imageability of word pairs (Madan et al., 2010) and imagery instructions (Hockley & Cristi, 1996) impact cued recall and associative recognition performance. This benefit seems to rely on *interactive*

imagery (imagining items as a combined/holistic/Gestalt image) rather than *separation imagery* (forming distinct images for pair members; Madan et al., 2010). Thus, the specific nature of the association between pair members (in this case, formation of a combined image) determines the success of cued recall. This idea is consistent with memory-model conceptualizations of cued recall as qualitatively different than free recall–particularly TODAM (Murdock, 1995) and CHARM (Metcalfe, 1990), which posit that cued recall pairs are encoded as a transformed, combined image. The difficulties in applying the SAM model (Raaijmakers & Shiffrin, 1981) to our initial data could have been in part due to the fact that the model does not explicitly parameterize this qualitative difference between FR and CR.

Applying these findings to the current effect, it could be that participants vary in their tendency to adopt imagery-based strategies (either similarly for FR and CR, or more for CR). Because imagery/imageability primarily affect association memory, perhaps the variance in adoption of these strategies (either at study or at test) inflates performance variance more for CR than for FR. It might be the case that even when participants in prior studies reported using imagery strategies for CR, some of them meant *separation imagery* and some meant *interactive imagery*–a heretofore ignored distinction that seems critical for CR but less so for FR. Related, it could also be the case that most participants are generally aware of (and adopt) effective encoding/retrieval strategies for FR (e.g., rehearsal, imagery), but only some participants are aware of or make use of effective strategies like imagery for CR (e.g., because paired association might be less familiar). Finally, it could be that the varying imageability of the words in the pairs inflated variance in the memorability of CR pairs (more so than the memorability of items in FR lists). Although we attempted to restrict imageability in our primary wordset to a middle/average range, we did not specifically control for the relative imageability of cues and targets in pairs, and did not restrict imageability in any of the secondary wordsets (i.e., object words, DRM word pairs).

Madan et al. (2010) tested cued recall of pairs of varying imageability (e.g., high-high, high-low), holding other important word characteristics (e.g., frequency, neighbors) constant. They found that high-high pairs were significantly more memorable than low-low and mixed pairs–suggesting that relative cue-target imageability matters. The authors did not examine variability, but visual inspection of error bars (Fig. 3a) did not seem to indicate a difference in variability, i.e., it was not obvious that performance on mixed pairs was more or less variable than pure pairs. So maybe imageability impacts mean performance but not variability. Of course, Madan et al. (2010) did not directly analyze variability, did not test FR of the words, and did not specifically encourage participants to adopt an imagery-based strategy. Our final experiment was designed to test the hypothesis that imageability and imagery-based strategies explain differences in CR and FR variability. In this experiment, we used words that were highly imageable, instructions encouraging participants to adopt an interactive-imagery strategy, and measures of mental imagery ability. We hypothesized that inter-individual variability in cued recall accuracy would *not* be greater than inter-individual variability in free recall accuracy when participants were explicitly instructed to use an imagery-based recall strategy and the studied words were high in imageability. We also hypothesized that imagery ability (behavioural using the *Image Comparisons Task*; Suggate & Lenhard, 2022, & self-report using the *Object and Spatial Imagery Questionnaire*; Blajenkova et al., 2006) would predict recall accuracy, and that imagery ability would more strongly predict cued recall than free recall accuracy. This experiment was preregistered (https://osf.io/uq3jm).

Methods

Materials & Measures.

Word pool. The word pool for Experiment 7 consisted of the 87 "high imageability" words from Madan et al. (2010), who chose the stimuli based on average imageability ratings

from four separate databases and restricted word frequency in the set to an intermediate range (see https://osf.io/scghk for the wordlist).

Image Comparisons Task (ICT; Suggate & Lenhard, 2022). In this behavioural imagery task, participants are presented with a perceptual adjective (e.g., "shiny") and two nouns (e.g., "trumpet", "violin"), and must decide which of the two nouns best fits the adjective. This task predicts reading performance in adults, and other versions of the task (e.g., comparing size of stimuli, whether animals have a long tail compared to their body, whether the colour of two objects are similar) have been widely used in the literature to measure imagery (Pearson et al., 2013). Additionally, this task is closer to the imagery and memory tasks we want participants to engage in (i.e., imagining pairs of concrete nouns). Participants viewed 75 adjectives (13 unique) in a random order and had to choose between two options for both (141 unique), one of which was correct. The adjectives and attendant pairs were translated from Suggate & Lenhard's German set (2022, see https://osf.io/mv493 for the stimuli). For each trial, the adjective was presented for 1s, and then each alternative was presented for 2s each (to reduce the likelihood of reading-speed effects), before the adjective and both responses were presented together for a keypress response. Accuracy and reaction time were recorded, and both were examined as hypothesized predictors of recall accuracy.

Object and Spatial Imagery Questionnaire (OSIQ; Blajenkova et al., 2006). In this self-report imagery measure, participants answered 30 five-point Agree-Disagree Likert-type questions assessing object imagery (15 items, e.g., “My mental images are colourful and bright”) and spatial imagery (15 items, e.g., “I was very good in 3-D geometry as a student”). This measure has been widely used to measure imagery ability and correlates with performance on object/spatial imagery tasks (Blajenkova et al., 2006), but has seen limited use in memory experiments (e.g., in the lone associative memory experiment, the OSIQ was

found to correlate with associative memory, albeit in aphantasics but not controls; Wittmann & Şatırer, 2022). Although object imagery is theoretically more related to the kinds of imagery we examined in this experiment, we had no firm predictions regarding differences between the two subscales (i.e., we hypothesized that both subscales would more strongly predict CR than FR).

Procedure. Participants first completed the ICT. Our reasoning in putting this task before the memory test was that the imagery exercise might serve as an additional imagery prime for the subsequent primary task. After this, participants were told about the upcoming memory task and were given the imagery instructions. These instructions were adapted from Thomas et al. (2023), who found that they improved memory performance relative to control instructions. The instructions for CR were:

"Studies have indicated that forming mental images of words significantly improves one's memory for them. Please try this technique for the pairs you are about to study. Form a mental image with both of the words interacting together when you are presented with a word pair. For example: For the word pair CAT-DOG, you could imagine the cat chasing the dog."

And for FR:

"Studies have indicated that forming mental images of words significantly improves one's memory for them. Please try this technique for the words you are about to study. Form a mental image of the words presented, and try to imagine them interacting. For example: If you study the words CAT, DOG, and HOUSE you could imagine the cat chasing the dog in front of the house."

Thomas et al. (2023) did not include FR, so the latter instructions were modified from the CR instructions.

Participants then completed a brief 'familiarization phase' consisting of one sample FR or CR test cycle of 5 words (pairs) before proceeding to the main study-test cycle. The main study-test cycle consisted of one FR or one CR study-test cycle consisting of 21 randomly selected words (FR) or word-pairs (CR), with one primacy and one recency buffer. Study phases were again self-paced (up to 30s per word/pair). After study, participants were reminded of the imagery strategy instructions before proceeding to the test phase. After test, participants were asked how much they used the interactive imagery strategy when studying and recalling the words, and whether they had used any other memory strategies. Participants then completed the OSIQ and the same debriefing questions used in prior experiments (age, proportion of words understood, distractions, technical difficulties, cheating). The experiment was programmed in jsPsych (see https://osf.io/7exsy) and administered online to participants on Prolific.co who reported English as a first & fluent language, had at least a 95% Prolific approval rating, and had completed at least three prior submissions on the site.

Sample. Our planned sample size was determined via two power simulations: one to detect a significant CR:FR variability effect, and to detect a null CR:FR variability effect via TOST equivalence tests (Lakens, 2017). The simulations suggested an $N = 200$ to achieve $\geq$ .8 power for the first analysis and an $N = 300$ to achieve $\geq .8$ power for the second. We planned to conduct both analyses first at $N = 200$, terminating data collection if we observed a significant variability effect *or* a non-significant effect and a significant TOST, otherwise continuing to $N = 300$. We met criteria for termination with a total sample of $N = 246$, from which we excluded participants based on preregistered criteria: 23 participants who got < 4 correct on the main study/test cycle, 9 who reported "Never" using the assigned imagery strategy, 5 CR participants with 6+ CR "fast skips" (skipping a test item in < 1s), 3 who

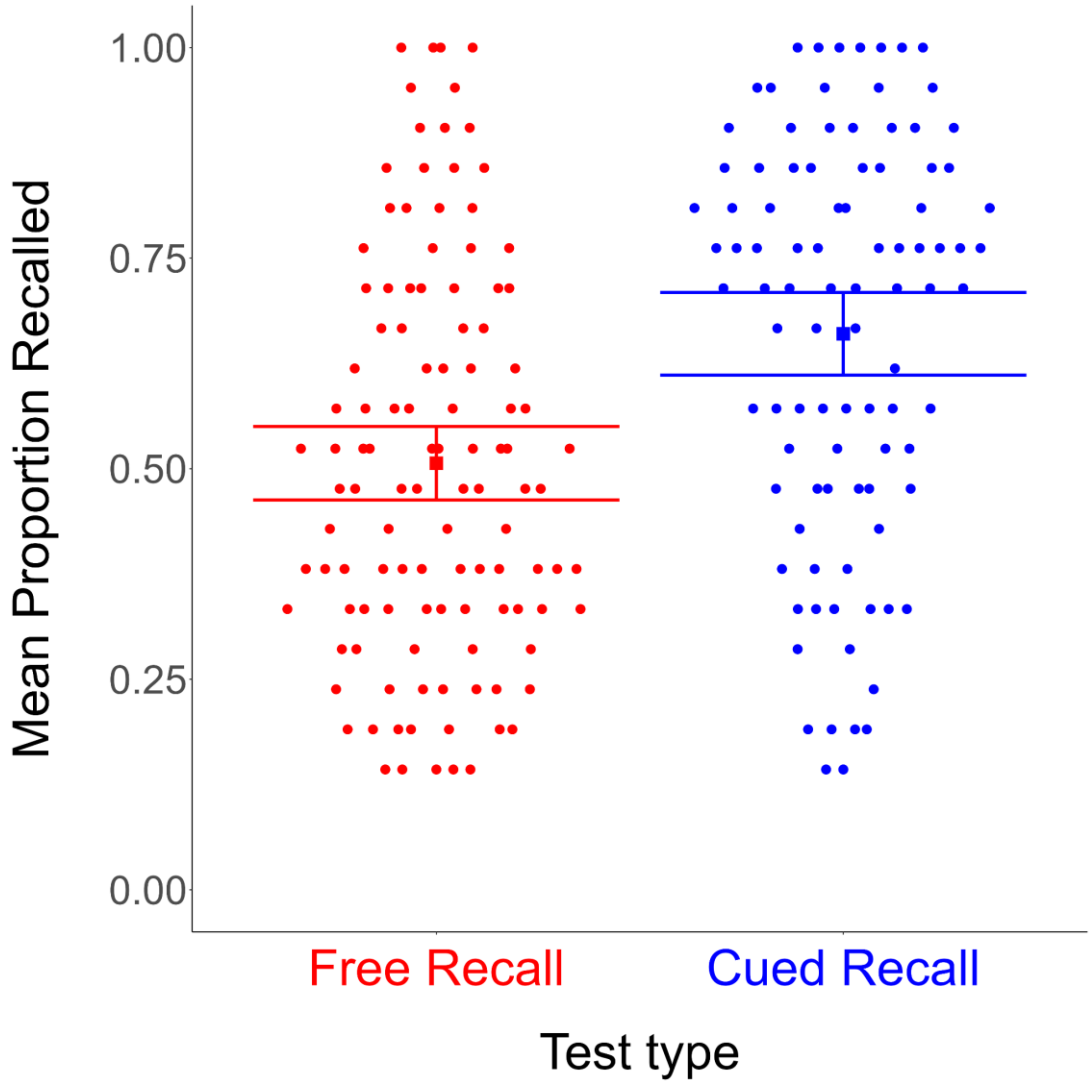
reported cheating, and 2 reporting technical difficulties. Our final sample included 208 participants aged 18-74 (M = 38.29, SD = 12.16), with 95 completing CR and 113 FR.

Results

Recall accuracy. Figure 33 shows recall as a function of test type:

Figure 33

Experiment 7: Memory performance as a function of recall test type



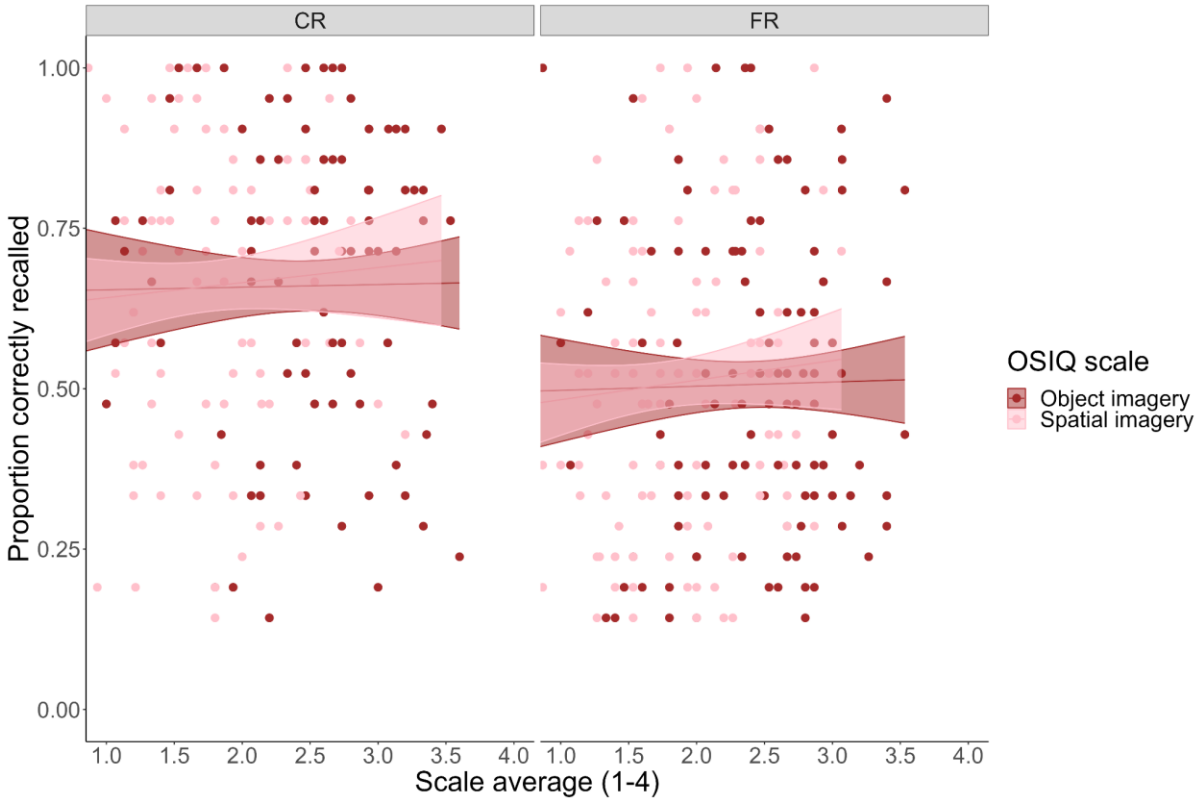
Note. Error bars = 95% CIs (between-subjects). Points jittered horizontally based on relative frequency.

The bootstrapped CR:FR variance ratio was 1.03 [95% CI: .87, 1.19], and the Levene's test of CR and FR variances was non-significant, $F(94) = 1.06$, $p = .38$. Crucially, the TOST equivalence test of the bootstrapped variance ratio was significant, $t(999) = 27.64$, $p < .001$, providing evidence that the observed variance ratio did not exceed our prespecified equivalence bounds (.9 and 1.1). Thus, for the first time across our experiments, we did not observe greater CR than FR variance. This was also the second time we obtained higher CR than FR performance, providing further evidence that the variability effect is not confounded with recall performance. Because we obtained compelling evidence for equivalence at our first prespecified stopping point, we chose not to continue data collection to our final prespecified stopping point ($N = 300$).

Imagery and recall. We found support for our first hypothesis–that imageability and/or imagery-based strategy use would eliminate the CR:FR variability effect. What of our second hypothesis–that imagery ability predicts recall (and more strongly predicts CR than FR)? We first predicted proportion correctly recalled from our self-report imagery measure–specifically, both subscales of the OSIQ:

Figure 34

Experiment 7: Memory performance as a function of test type and OSIQ subscales

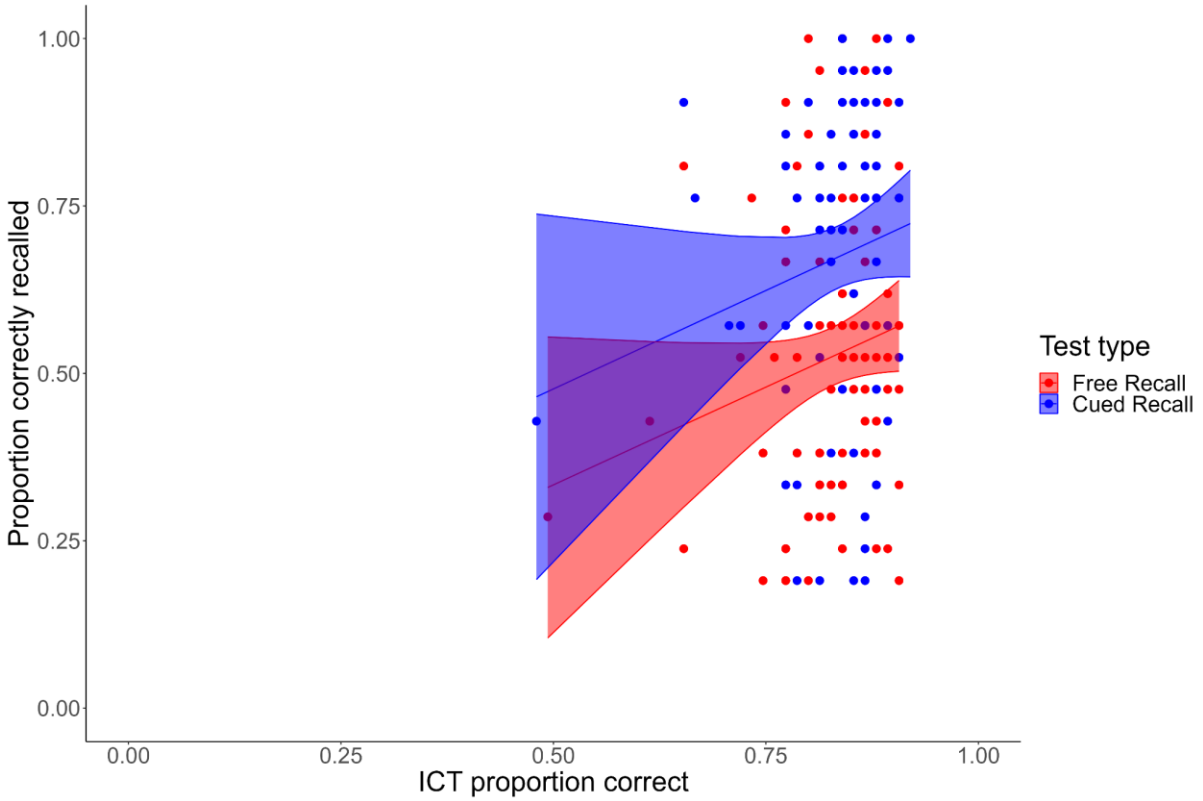


Note. Points = individual participants. Lines & ribbons = linear regressions & 95% CIs.

There was no effect of OSIQ average on memory test performance, $\chi^2(1) = 1.17$, $p = .28$, and no interactions between subscale, test type, and OSIQ score (all *p*s > .39). This is unlikely to be due to any restrictions of range, as we had a good amount of variability in both performance and OSIQ scores. Next, we examined the relationship between ICT accuracy and proportion correctly recalled:

Figure 35

Experiment 7: Memory performance as a function of ICT accuracy

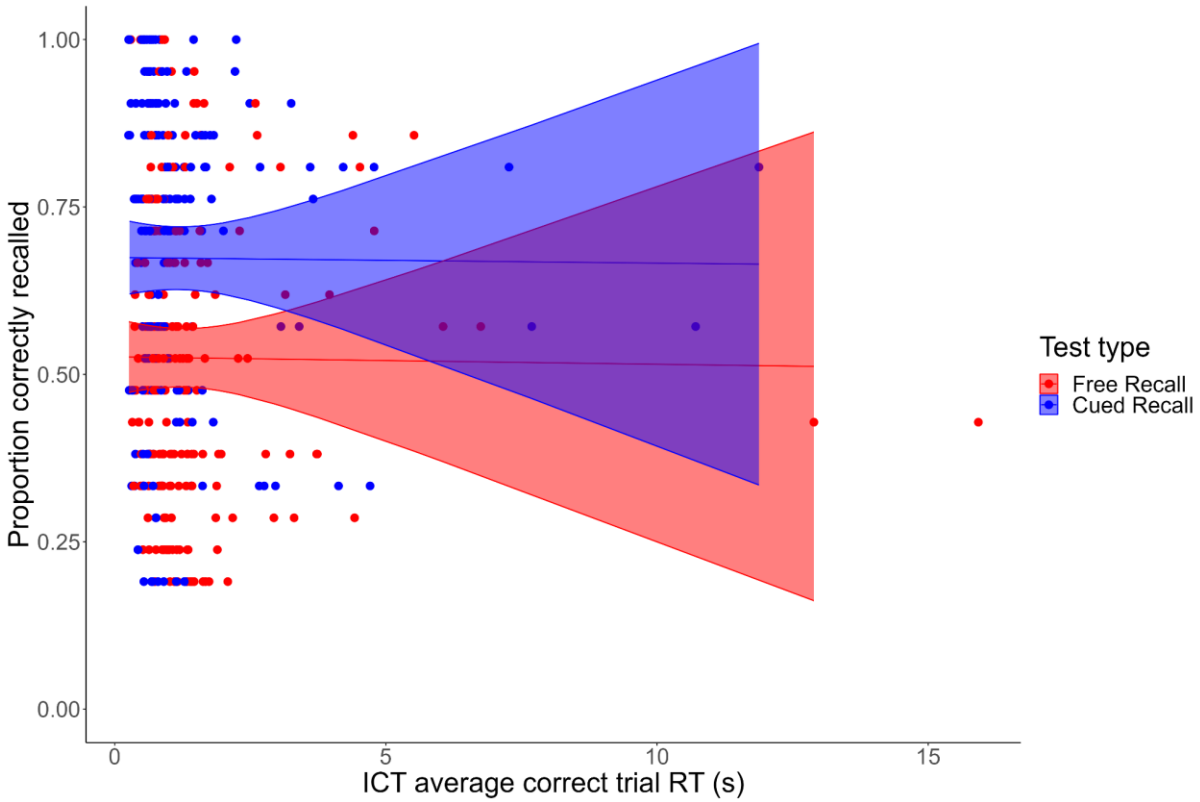


Note. Points = individual participants. Lines & ribbons = linear regressions & 95% CIs.

Performance on the ICT significantly predicted performance on the memory test, $F(1) = 5.32$, $p = .02$, but this appears to be largely driven by a few low-performing outliers. Accuracy on the ICT was generally quite high, with little variability across participants. What about reaction time–a potentially more sensitive measure? When examining the relationship between participant-level average RT on correct ICT trials:

Figure 36

Experiment 7: Memory performance as a function of ICT RT on correct trials

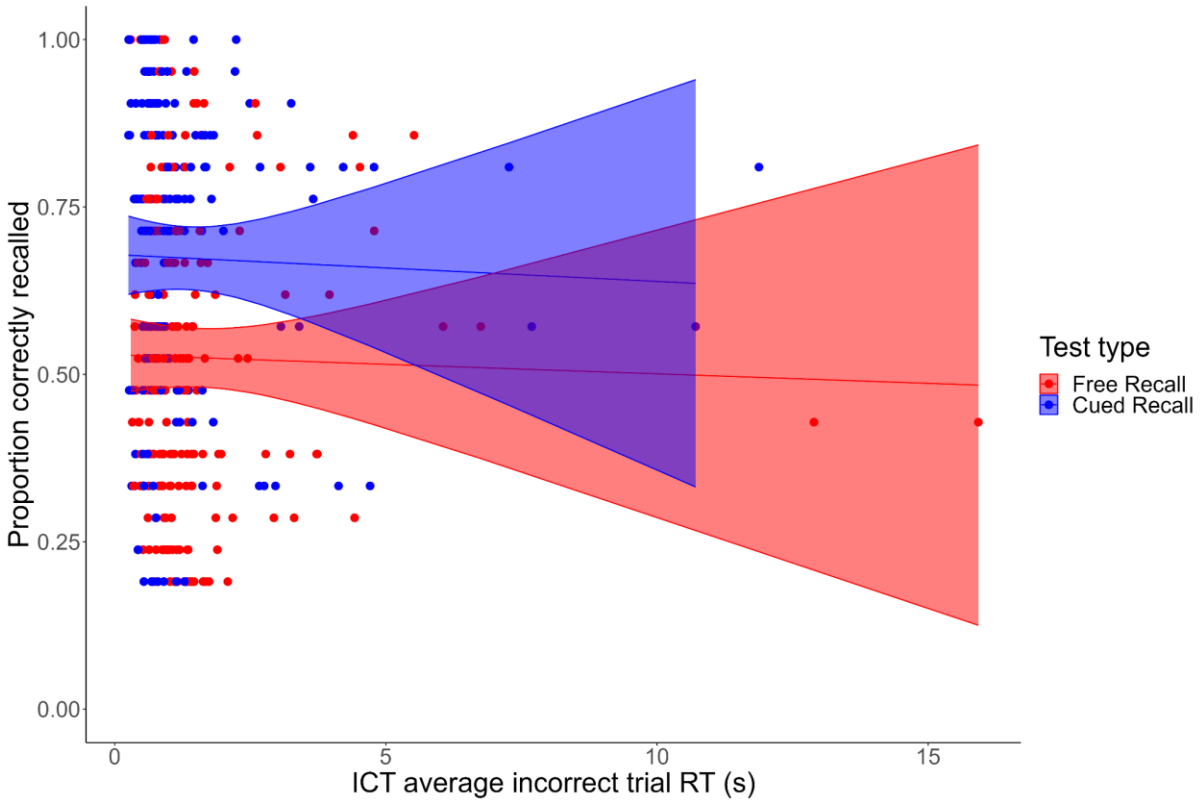


Note. Points = individual participants. Lines & ribbons = linear regressions & 95% CIs.

There was no significant relationship between average RT and memory test performance, $F(1) = .01$, $p = .93$. Again there were some outliers here, but even when conducting an exploratory analysis restricted to participants with an avg. RT < 5s ($n = 188$), the relationship was not significant, $F(1) = .67$, $p = .42$, and similarly when restricting to participants with an avg. RT < 2s ($n = 173$, $F(1) = .31$, $p = .58$). Results were similar when looking at average RT on *incorrect* ICT trials:

Figure 37

Experiment 7: Memory performance as a function of ICT RT on incorrect trials



Note. Points = individual participants. Lines & ribbons = linear regressions & 95% CIs.

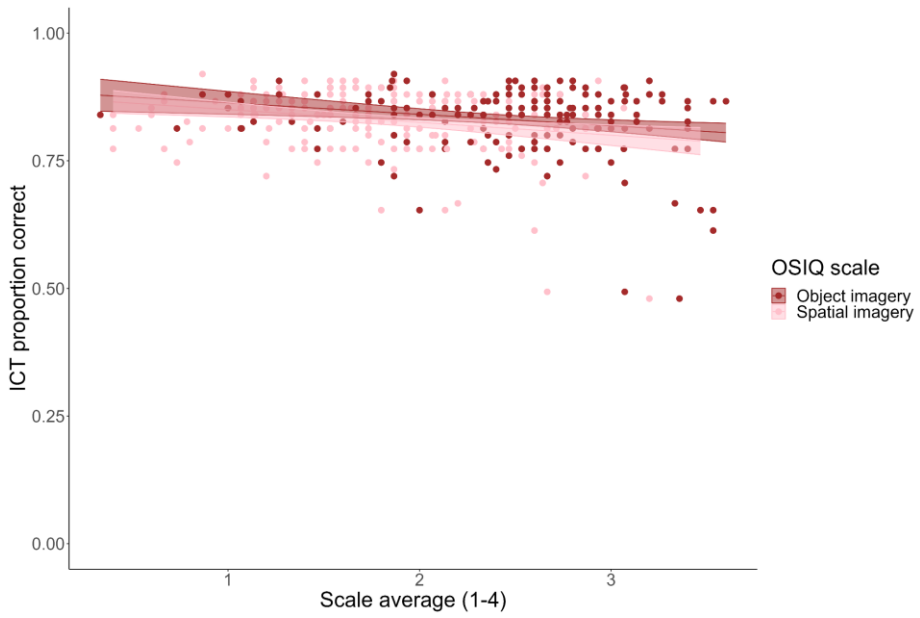
...with a non-significant effect of RT, $F(1) = .11$, $p = .74$ ($ps = .89, .06$ when restricting to participants with < 5s avg. RT and < 2s avg. RT respectively). Thus, we found little evidence that our imagery measures meaningfully predicted recall performance. It could be that because all the words in this experiment were highly imageable, even participants lower in imagery ability were able to generate images conducive to recall. Additionally, other experiments using similar self-report and behavioural imagery measures have failed to find significant relationships between these measures and recall performance (e.g., Kluger et al., 2022; Thomas et al., 2023). Of course, it is still possible that we could have detected a

relationship between our imagery tasks and recall performance with other measures of mental imagery, such as the *Vividness of Visual Imagery Questionnaire* (McKelvie, 1995) or the *Spontaneous Use of Imagery Scale* (Reisberg et al., 2002). Our rationale for choosing the OSIQ and ICT was that these tasks seemed more closely related specifically to word memory (e.g., a clear distinction between *object* and *spatial* mental imagery for the OSIQ, with the former being more relevant, and the explicit word-visualization aspect of the ICT). But again, our primary focus in Experiment 7 was to examine the impact of imagery strategy instructions, independent of imagery ability.

Were the behavioural and self-report imagery measures related? If so, this would at least provide some evidence for the face validity of the measures. We conducted exploratory analyses investigating the relationships between the measures. First, the OSIQ subscales and accuracy on the ICT:

Figure 38

Experiment 7: Relationship between OSIQ and ICT accuracy

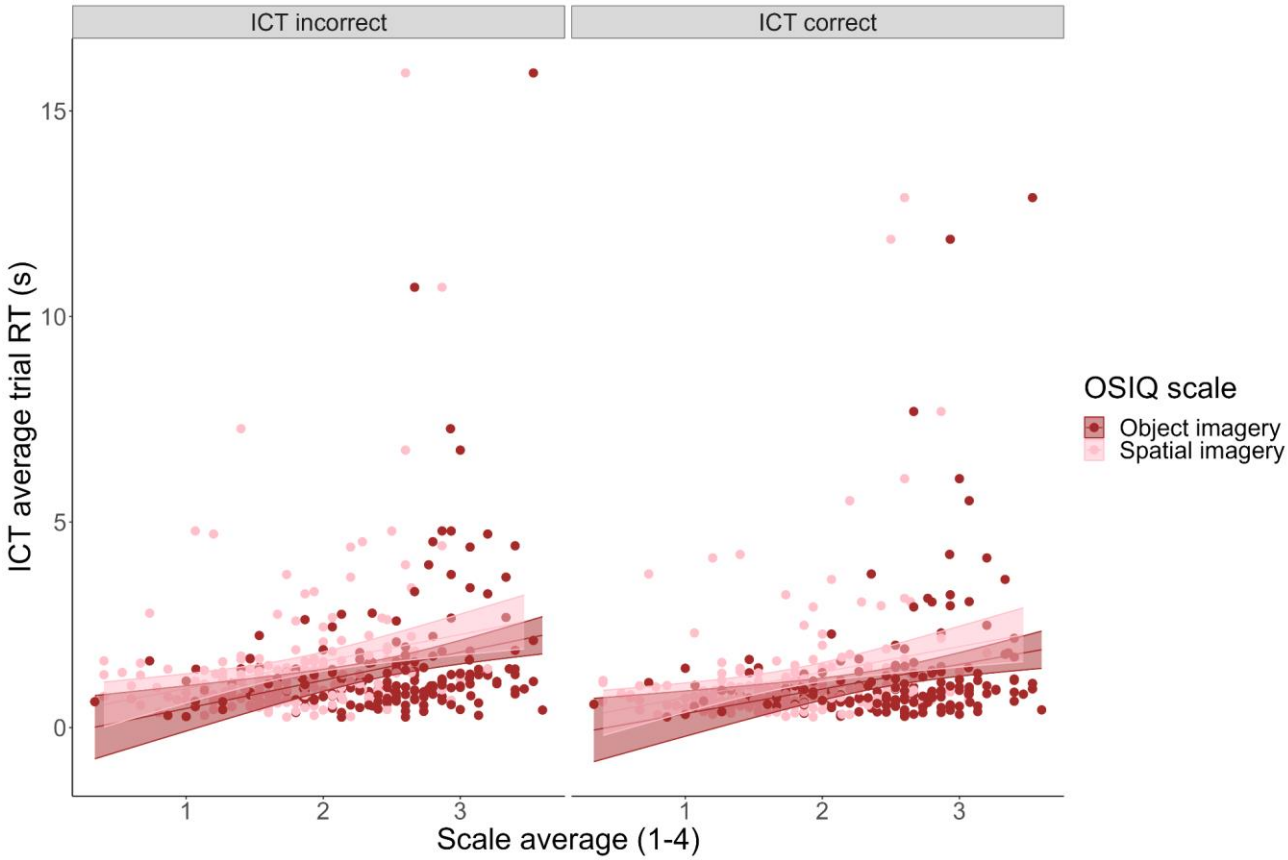


Note. Points = individual participants. Lines & ribbons = linear regressions & 95% CIs.

Surprisingly, there was a (modest) negative relationship between OSIQ score and ICT performance, $F(1) = 4.91$, $p = .03$ (but no interaction between subscale and score, F(1) = .06, $p = .80$). For RT on the ICT task:

Figure 39

Experiment 7: Relationship between OSIQ and ICT RT



Note. Points = individual participants. Lines & ribbons = linear regressions & 95% CIs.

Interesting again was the significant relationship between ICT RT and OSIQ scores, $F(1) = 49.52$, $p < .001$, with a similar relationship for both subscales and for correct and incorrect ICT trials (all interaction *p*s > .73). These results held when restricting to

participants with average RT < 5s ($p < .001$), but not when restricting to participants with average RT < 2s ($p = .09$). Thus, it appears that better self-reported imagery ability (as indexed by the OSIQ) predicted *poorer* performance on the ICT, both in terms of (slightly) lower performance and longer reaction times. One possibility is that participants higher in imagery spent longer imagining and comparing the words in the pairs. High-imagery participants may also have been (slightly) more likely to choose the normatively incorrect imagery-comparison alternative because they relied on idiosyncratic imagery when making the comparisons (vs. low-imagery participants who may have relied on semantic information). This relationship was not of primary interest for our experiment, but at least suggests that the lack of relationship between imagery and recall performance we observed was not due to insensitivity or invalidity of the measures.

Discussion

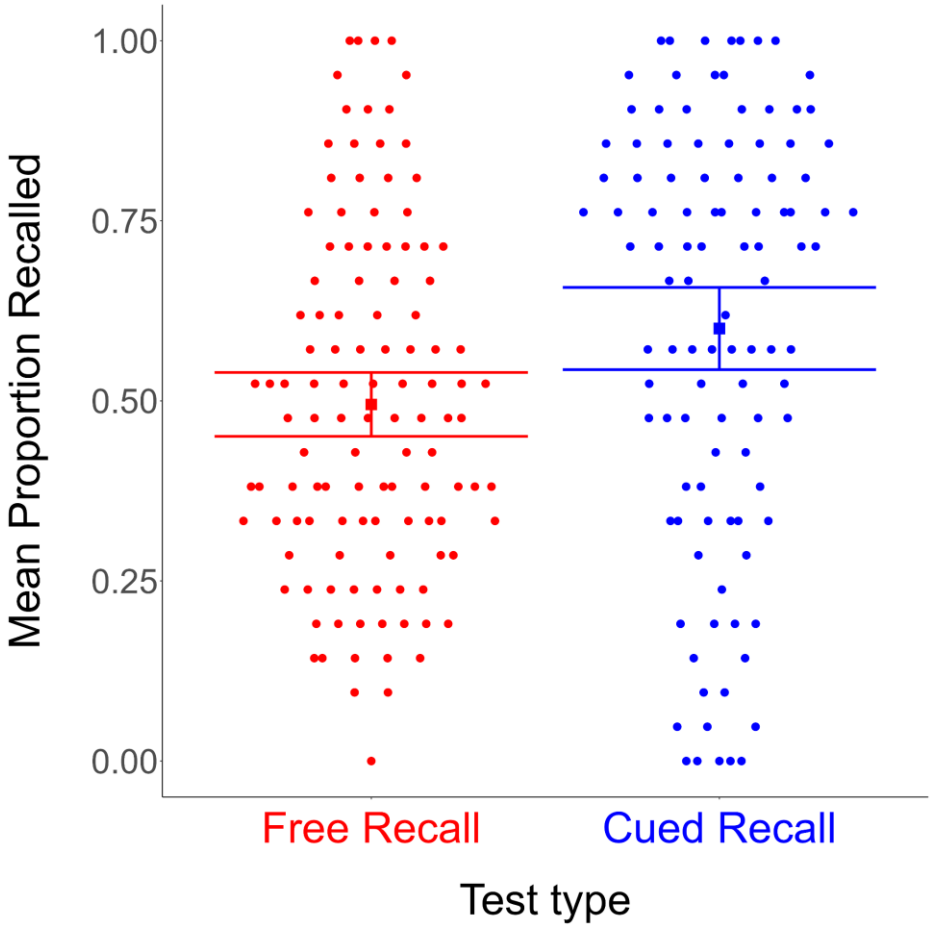
Our imagery manipulation in Experiment 7 eliminated the variability effect. However, there are a number of possible explanations that could explain the results we obtained. We considered each of these explanations and (where possible) tested them with additional exploratory analyses:

1. The lack of a variability difference can be explained by CR ceiling/FR floor effects. In all of our experiments, there was an ever-present threat of variability-restricting ceiling and floor effects. Typically, the threat was to the maximum amount of FR variability we could observe, but in this experiment, one could argue that a potential CR ceiling effect might have artificially reduced CR variability. I would argue against this–in the final sample only 7% of CR responses were at ceiling, and there were the same total number of participants at ceiling or (task) floor in CR and FR ($n = 9$, and similar proportions: 9% and 8%). Qualitatively, in Figure 33 the distributions appear almost mirrored (whereas in previous experiments the CR distribution often had a noticeably different shape).

At the lower end of the distribution, there was a higher rate of CR participants excluded for low performance (nearly double that of FR). Does including these participants change the overall pattern of results?

Figure 40

Experiment 7: Memory performance as a function of recall test type, including low-performers



Note. Error bars = 95% CIs (between-subjects). Points jittered horizontally based on relative frequency.

With the addition of ten CR and three FR previously-excluded participants, the bootstrapped CR:FR variance ratio jumps to 1.23 [95% CI: 1.06, 1.40], with a significant Levene's test, $F(104) = 1.49$, $p = .02$, and a non-significant equivalence test, $t(999) = 44.20$, p

= 1. Putting aside for a moment the question of whether these exclusions are justified (e.g., these participants blew off the task vs. these participants made a bona fide attempt but were low CR-performers), what do these differing results imply for this (and prior) experiments?

First, it is merely a statistical fact that extreme scores exert greater influence on measures of central tendency and variability, so not surprising that 10 outliers substantially shift the standard deviation. Second, in all prior experiments, the CR:FR variability difference was observed even when excluding low performers on criteria similar to this experiment -- not so this time. Still, the fact that there were more low-performance exclusions for CR than for FR is potentially noteworthy. One possibility there is that participants may have been more easily overwhelmed or fatigued by the CR task (i.e., double the words to study), and checked out when it came time to test.

Although we thought it unlikely that these excluded participants cast doubt on our conclusions, we conducted several exploratory analyses of these “low CR performers”. These participants spent directionally but not significantly less average time studying each pair (3.9s, $z = 1.52$, $p = .43$), and significantly more average time per-pair at test (2.8s, $z = 2.66$, $p = .04$) than the rest of the CR sample. Neither was it the case that the low performers left more blank responses ($\chi^2(1) = .66$, $p = .42$), or were any faster/less accurate on the ICT (ps > .18). So, we can't say with confidence that the low-performers were merely blazing through the tasks. What kind of responses did low-performers give at test (e.g., were they responding to cues with the wrong targets, other cues, or non-studied words)? The vast majority of incorrect responses for low performers were *non-studied* words (i.e., neither cues, targets, nor words from the ICT task)–for all low-performing CR participants, 89-100% of their commission errors were non-studied words.

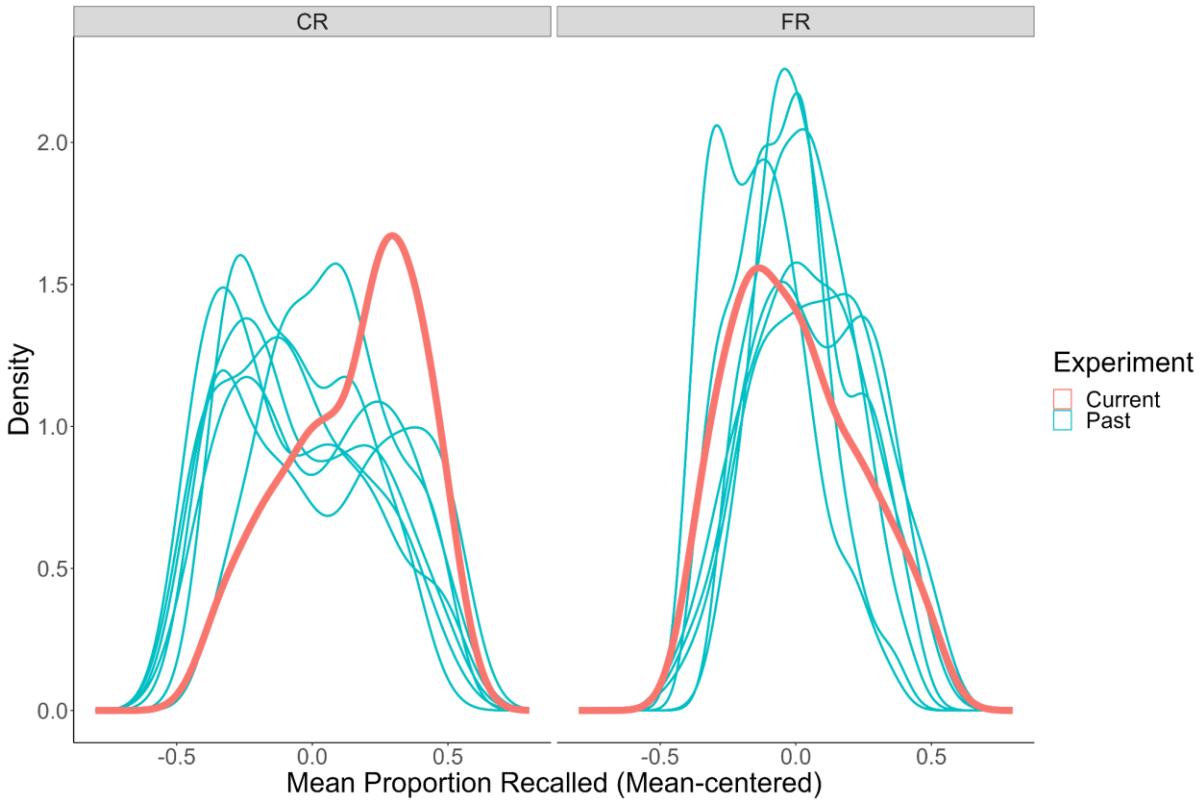
What to conclude about the CR low performers whose inclusion potentially changes

the fundamental conclusions? The high proportion of non-studied words given as responses and the slightly faster study RTs contrasted with good performance on the ICT suggests that these participants, while they were probably not outright shirking the tasks, may have not made a serious legitimate attempt on the CR test phase. Thus, we argue that their exclusion is justified, although it remains a point of interest that there were more such exclusions for CR than for FR.

2. The imagery manipulation increased FR variability instead of decreasing CR variability. Our prediction was that the imagery strategy manipulation would eliminate the CR:FR variability effect by *reducing* variability in CR performance. However, our results could just as likely be due to the manipulation *increasing* variability in FR performance. The instructions (adapted from Thomas et al., 2022) were somewhat more ambiguous/complex for FR than for CR–participants were asked to imagine items in an increasingly long list interacting. This (or some other related factor) could have contributed to an increase in FR variability. Although we could not definitively rule out this possibility without a non-imagery-instructions condition in this experimental design, past experimental data could provide some insight into this potential explanation. First, we plotted the CR and FR distributions for the current and past experiments, mean-centering accuracy to make cross-experiment comparisons in variability easier:

Figure 41

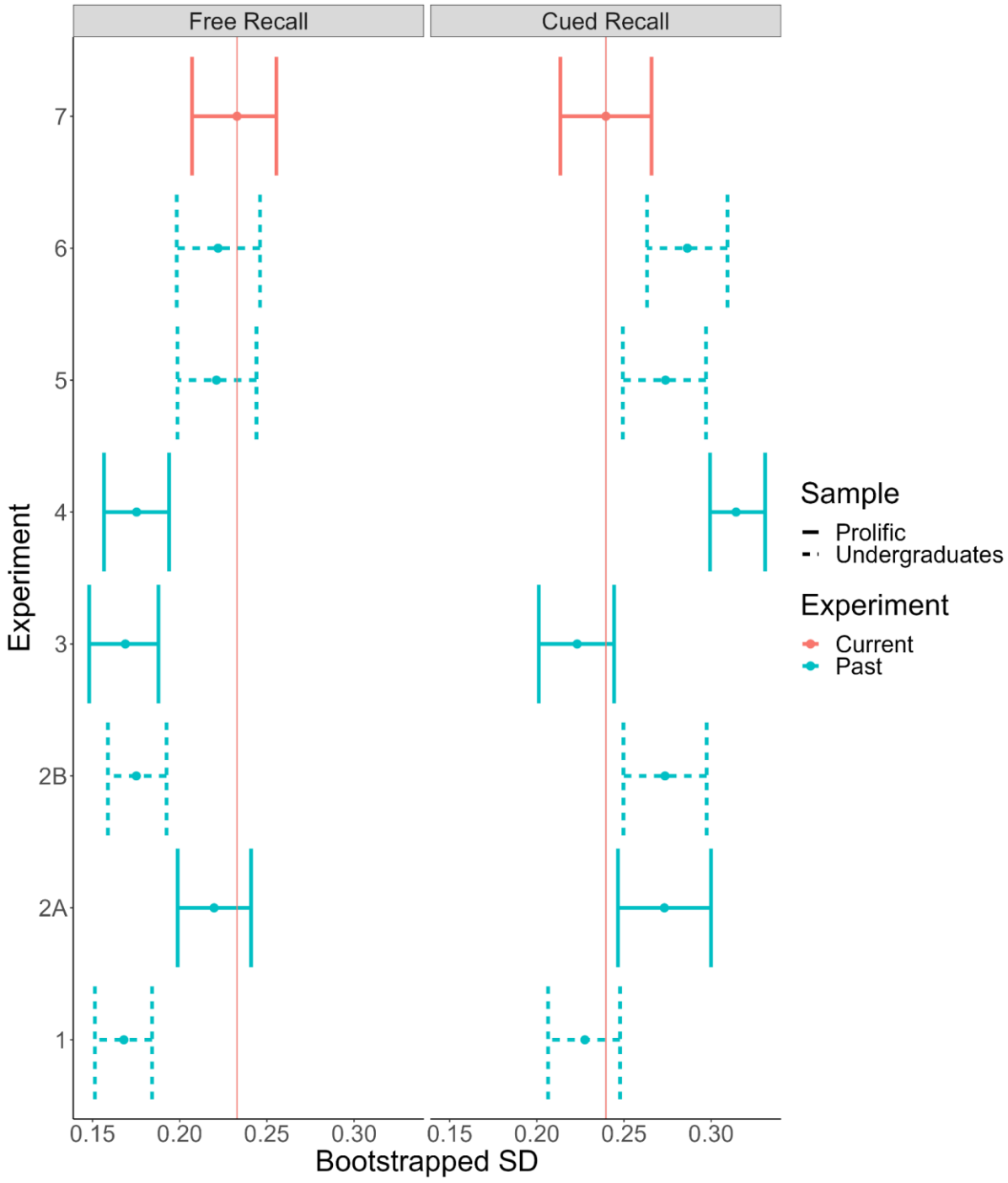
All experiments: Mean-centred distributions of FR and CR performance



At least from an interocular test, FR variability seems similar to that observed in prior experiments, with CR variability potentially decreased. This comparison also reveals that CR performance in the current experiment lacks the apparent bi-modality characteristic of most of the other experiments. As a more quantitative test, we compared the bootstrapped FR and CR standard deviations across experiments:

Figure 42

All experiments: Bootstrapped FR and CR standard deviations



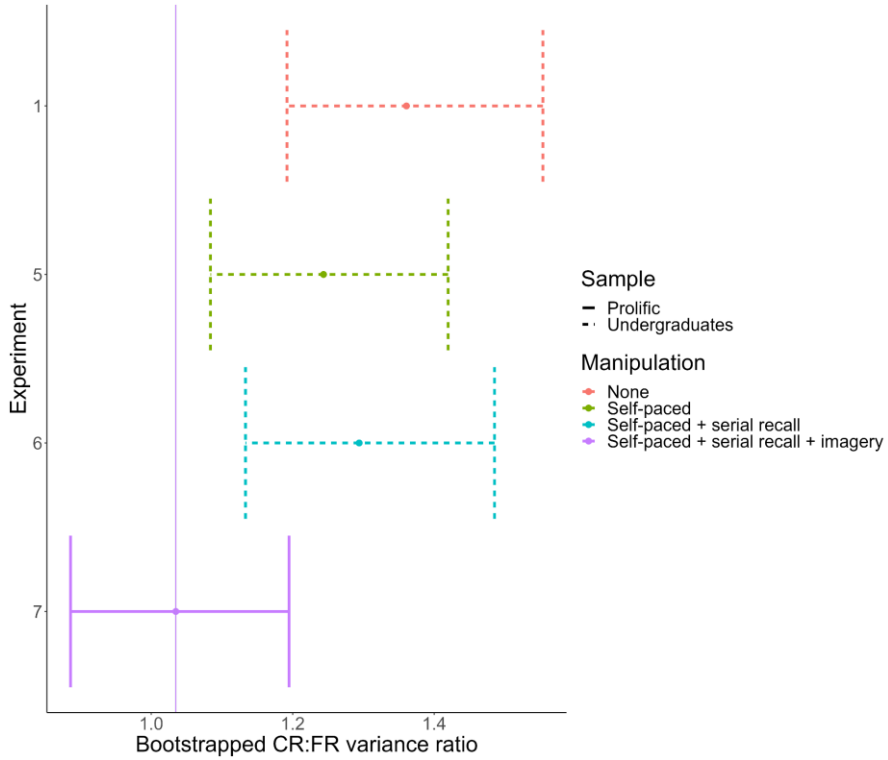
Note. Error bars = bootstrapped 95% CIs.

Analyzing the bootstrap samples, CR variability in this experiment was significantly lower than 5/7 prior experiments, and significantly higher than 2/7 prior experiments (*p*s < .001). FR variability was significantly higher in this experiment than in all prior experiments (*p*s < .001). Although these cross-experiment comparisons of bootstrap samples should be taken with a grain of salt, it does seem that the imagery manipulation exerted its effects by both increasing FR variability and decreasing CR variability. Indeed, there was an overall significant interaction between test type (FR, CR) and experiment (Current, Past) for standard deviation $\chi^2(1) = 6{,}219.3$, $p < .001$, with similar magnitudes for the FR increase (+.04) and the CR decrease (-.03). Although we cannot eliminate the possibility that our manipulation operated on both test types, we can be fairly certain that it did reduce CR variability. And this possibility does not invalidate the underlying central thesis of this experiment that strategy use–in particular imagery strategy use–can modulate the CR:FR variability effect.

3. It was not our imagery manipulation that specifically eliminated the effect, but the additive result of the manipulation and the incremental methodological changes that we made. Another possibility is that it was not the imagery manipulation per se, but the combined effects of the manipulation with prior manipulations–in particular the introduction of self-paced study and serial cue presentation at test–that produced the pattern of results in the current experiment. Although neither of these prior manipulations *individually* eliminated the CR variance effect and it appears that the imagery manipulation did, it could be that our latest manipulation was simply the proverbial straw that broke the camel's back. Although the current data do not permit a direct test of this possibility, we can compare the size of the CR:FR variability effect across incremental experiments. The results of this exploratory comparison:

Figure 43

Selected experiments: Manipulations and bootstrapped CR:FR variance ratios and manipulations



Note. Error bars = bootstrapped 95% CIs.

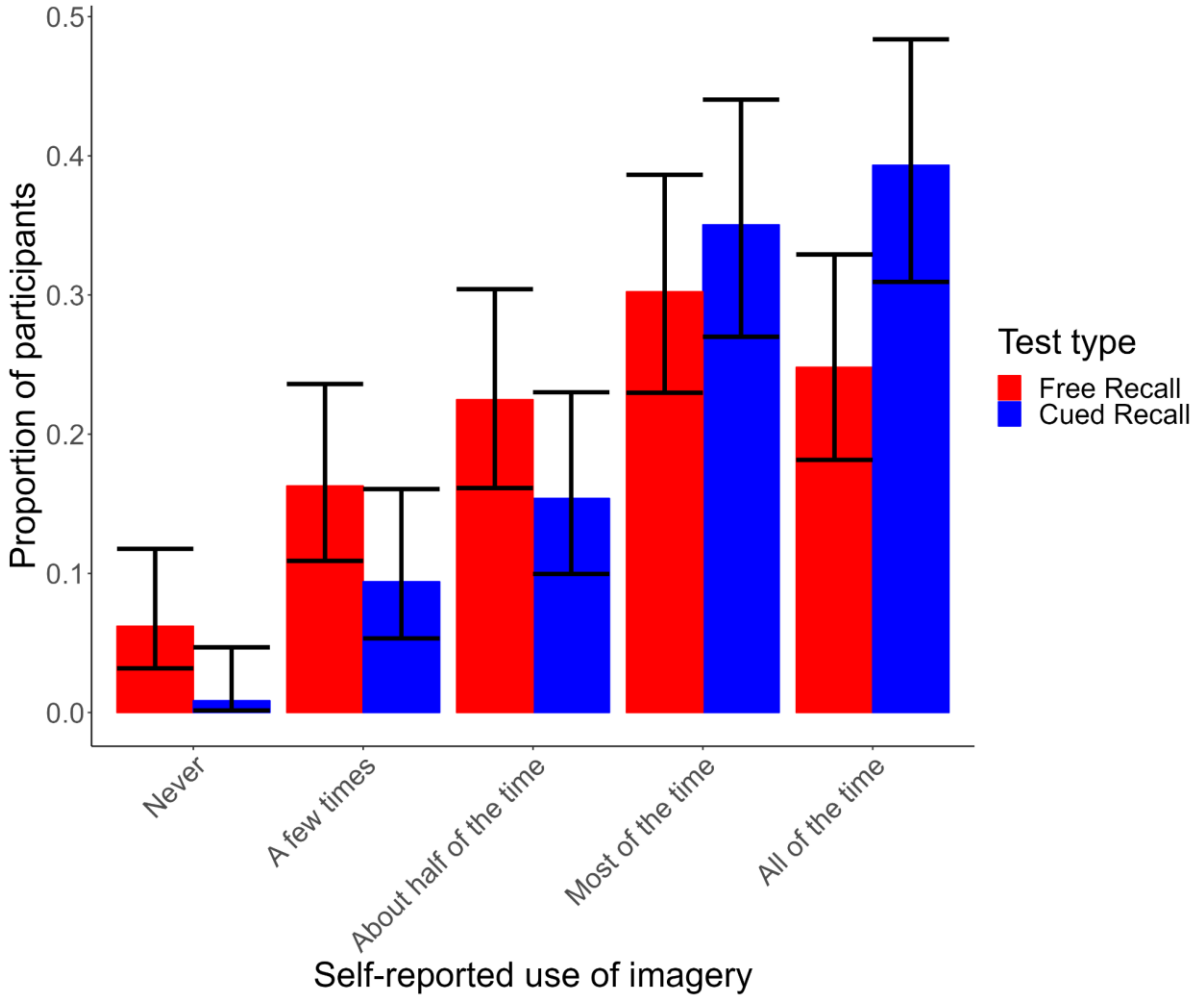
Using the bootstrap samples, we examined whether there were differences in the magnitude of the effect across experiments. Although all differences were significant (*p*s < .001), crucially, the addition of serial CR to self-paced study phases *increased* the CR:FR variance ratio. Even if we assume that self-paced study phases reduces the variance ratio relative to a no-manipulation control, the relative effect of the imagery manipulation (i.e., the difference in the size of the effect between Experiments 1 and 7: .32) was nearly triple that of the relative effect of self-paced study (i.e., the difference in the size of the effect between

Experiments 1 and 5: .11). So, even if the other manipulations served to reduce the effect, it seems that the imagery manipulation had a large and unique effect on the CR:FR variance ratio (cross-experiment bootstrap analyses notwithstanding).

4. It is not imagery per se, but the homogenization of strategy use that reduced CR variability. This experiment was the first in which we specifically instructed participants to use a particular strategy–in this case an imagery-based strategy shown to be effective for paired-associates learning (Thomas et al., 2023). Our supplementary analyses of self-reported strategy use in prior experiments showed hints of increased variability in the strategies participants adopted for CR relative to FR (e.g., Figure 5, SOM 4E, 5E). Thus, it is possible that instructing participants to use *any* particular strategy would have had the same effect. Further, it could be that the imagery instructions in some way disrupted participants' preferred FR strategies, increasing variability in FR performance. In prior experiments querying strategy use, rehearsal was a dominant strategy (e.g., Figure 5, Figure 17, SOM 4E, 5E). When instructed to use imagery on the FR task, it is possible that some participants still relied mostly on rehearsal, some used a combination of imagery and rehearsal, and others mainly used imagery. We did ask post-test self-report questions about strategy use, specifically a) the degree to which participants adopted the assigned imagery strategy (specifically, "When studying and recalling the words, how often would you say you used the strategy of generating mental images of the words interacting?") and b) what, if any, other strategies that participants used. The results for a) were as follows:

Figure 44

Experiment 7: Self-reported frequency of imagery strategy use by test type



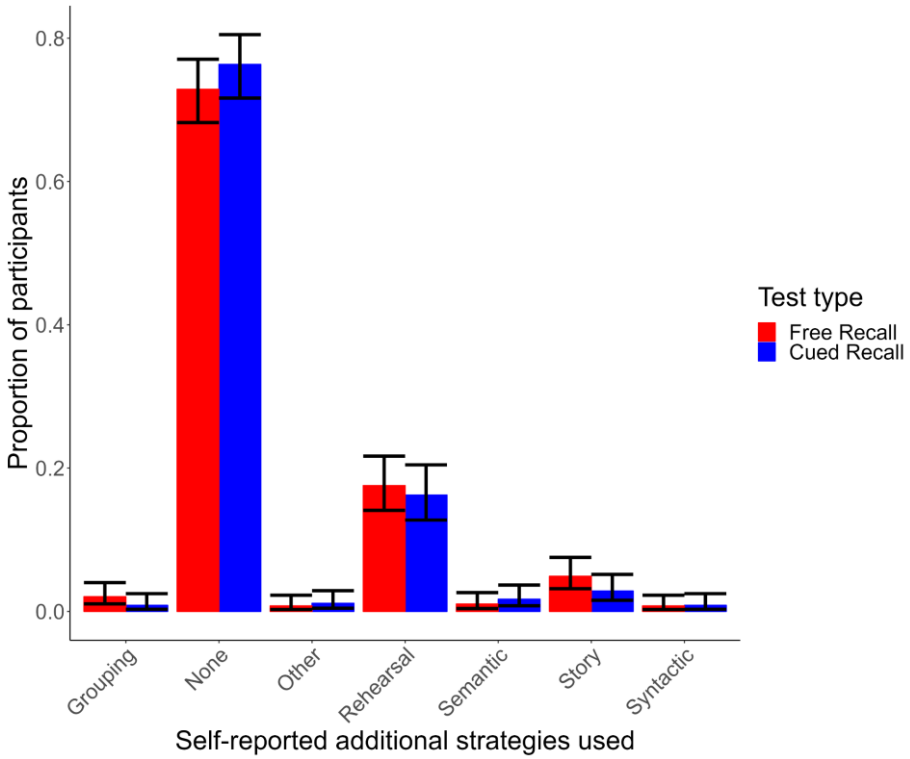
Note. Error bars = proportion 95% CIs (using Wilson's (1927) method).

Indeed, CR participants reported significantly more use of the imagery strategy, $t(242.26) = 3.64$, $p < .001$, with the majority of CR participants reporting using the imagery-based strategy "all of the time", most FR participants reported only using the imagery-based strategy "most of the time". Did that mean that FR participants were more likely to report using *other* strategies? We examined qualitative self-reports of other strategy use and the first author coded them using the same guidelines used in prior experiments. The proportions for

the different self-reported strategies were as follows:

Figure 45

Experiment 7: Self-reported other strategies used, by test type



Note. Error bars = proportion 95% CIs (using Wilson's (1927) method).

Importantly, the vast majority of participants for both FR and CR did not report using any other strategies. There were no significant differences in use of other strategies by test type (*p*s > .21), which suggests that FR participants were not necessarily using a greater variety of strategies. In light of the finding that FR participants *did* report using imagery less frequently, it could be that FR participants were not adopting different strategies, they merely were not able to apply the imagery strategy all the time (e.g., as lists became longer and words in the list less related). This might partially explain the lower average performance for FR than CR in this experiment. Still, we cannot rule out the possibility that the elimination of the CR:FR

variability effect may be due to a homogenization of CR strategies (or heterogenization of FR strategies). But this possibility does not conflict with our core argument that FR and CR variability differences are driven by differences in strategy use. This argument aligns with prior findings that differences in strategy use explain

5. Participants are more familiar with FR and generally adopt similar strategies, but are less familiar with CR and only some participants adopt effective strategies (like imagery). Or, adopted strategy matters more for CR than for FR. By making all participants aware of effective CR strategies, we created a similar situation for FR and CR. In other words, it may be that for FR most participants adopt a rehearsal-based strategy, and variation in performance is due to variation in effective use of that strategy. Perhaps the task of memorizing lists of single items is also more familiar to participants (e.g., grocery lists) than memorizing pairs of unrelated words. For CR, it could be that for whatever reason, some participants adopt effective imagery-based strategies while others do not. Prior findings of substantial boosts to CR performance with imagery instructions (relative to control or repetition instructions) suggests that most participants may not be aware of the efficacy of imagery for associative learning (e.g., Bower & Winzenz, 1970). The apparent bimodality of CR distributions in prior experiments supports such an explanation. What might be different about those participants who did well on the CR task, and in prior experiments adopted imagery-based strategies without instruction? When queried about effective memorization strategies, ‘skilled memorizers’ recognize the importance of mental imagery (Thomas et al., 2023). It is possible that some proportion of the ‘upper mode’ in our CR distributions represent these skilled memorizers. These participants may have had more experience with CR-like tasks, or have had prior instruction on the use of imagery-based strategies. Interestingly, the results of Experiment 7 suggest that it is not differences in mental imagery ability that differentiate memorizers, merely use of an imagery-based strategy (although this

could be due to the fact that all the words studied were highly imageable, and other studies have failed to find a link between imagery measures and recall performance; Thomas et al., 2023). So it is not necessarily that high-performing CR participants have better mental imagery—perhaps they are more aware of the effectiveness of imagery-based memorization, or more motivated to use imagery on memory tasks. Future research could test whether and why higher-performing CR participants might spontaneously adopt effective strategies like imagery (e.g., prior experience, knowledge of strategies, etc.), or whether other individual differences (e.g., motivation) explain their improved memory.

Imagery instructions have been shown to have similar benefits for FR in terms of output and clustering (Begg, 1978). Thus, one might ask why the same pattern of bimodality is not observed for FR. If adoption of imagery-based strategies (e.g., by 'skilled memorizers') explains patterns of CR variability, why is the same not true of FR variability? One possibility is that although imagery aids free recall, it may be less important than other aspects such as *order information.* Participants often rely on serial order information to guide output in FR, particularly when items in the list are unrelated (Nairne et al., 1991), as was the case in all of our experiments. The hegemony of the serial position curve in free recall across our experiments (Fig. 11) is consistent with the idea that participants overwhelmingly relied on order information when attempting free recall. With the exception of our two serial CR experiments, participants could not use order information to aid cued recall (perhaps the 'lower mode' of CR participants attempted to use order-based rather than imagery-based strategies). Thus, FR variability may be modulated by more quantitative differences in ability to effectively encode order information, whereas CR variability may be modulated by more qualitative differences in encoding strategies adopted *and* quantitative differences in ability to use the various strategies adopted. Again, although we cannot decisively adjudicate between this possibility and possibility 4.–that *any* instructed CR strategy would have done the trick in

reducing variability–it seems clear that the novel variability effect stems from individual differences in memory strategy use for free and cued recall, and perhaps in the relative importance of differential strategy use for the two tasks. Perhaps this is not too surprising given the powerful role encoding strategies play in determining recall performance (e.g., 27% of variance in recall explained by encoding strategy, beating out search efficiency, monitoring, study time allocation, and other factors; Unsworth, 2016).

General Discussion

In an initial experiment, we observed a surprising incidental finding–greater inter-individual variability in cued recall than in free recall. This finding had not been reported on prior, ran counter to many intuitions about the memory tasks, and did not seem to neatly fit into existing formal theories of memory. In a series of experiments, we systematically varied primarily *methodological* factors (e.g., nature of the words, task constraints, encoding time, list length) and primarily *theoretical factors* (e.g., test presentation order, encoding strategy instructions). Across a variety of manipulations, wordsets, and samples, we observed a robust CR:FR variability effect, both within- and between-subjects. The sole manipulation that eliminated the effect was our imagery-strategy manipulation, which we argue points to a strategy-based explanation for variations in individual differences across tasks. Specifically, we suggest that the variability difference is most likely due to participants adopting more similar encoding and retrieval *rehearsal*- or *order-based* strategies for free recall, and more varied strategies for cued recall.

Consider again the predictions by the various memory researchers (emphasis mine):

- Henry L Roediger wrote, “If you had asked me to guess beforehand which procedure was more variable, **I would have guessed free recall. That task is … prone to various strategies**, from forming a story with the words (depending on presentation rate) to rote rehearsal (and many others). Using Craik’s logic, paired-associate learning provides more retrieval support (the stimulus or cue at test) than does free recall (a blank computer screen). I would have thought that the additional retrieval support would have constrained variance.”

- Colleen Kelley replied “I am actually surprised to think that there would be more strategies available for paired associates than free recall, **wouldn't you think there are more constraints in paired associates, so less room for variation**?”

- John Dunlosky responded, “If I hadn’t read your note first, I’m pretty sure I would have predicted that **individual differences in strategy use would contribute to larger individual differences in free recall** than paired associate recall.”

Most of these predictions are based on the assumption that participants have more strategies available for free recall. Was this truly the case in our experiments? There is a well-documented dissociation between *item* memory and *order* memory that is potentially relevant here–e.g., item memory benefits and order memory suffers when items are meaningfully semantically related (Nairne et al., 1991). That is:

“With categorized lists, the inherent category structure provides an alternative organizational scheme (other than seriation): Subjects can simply generate category instances as output candidates and then check their legitimacy as episodic targets. With . . .unrelated lists . . ., organizational structure could only be established at the point of encoding (e.g., linking the items together in a serial chain). . .” (Nairne et al., 1991, *p.* 706)

All this to say that in our experiments–where lists were not categorized–participants may have *only* been able to rely on serial information (e.g., rote rehearsal) for encoding and retrieval. Paradoxically, a lack of structure in the lists may have imposed more structure on participants’ encoding and retrieval strategies. With related lists, participants might have been more able to adopt a variety of strategies.

The story is different for CR–in all but two of our experiments seriation-based strategies would not have been effective. And in the two final experiments (where cues were presented in the same order at study and test), serial order still seems to be of limited use (versus associations formed between pairs). How then might participants approach this task? Consider a hypothetical 'low-performing' CR participant who adopts a purely seriation-based rehearsal strategy for both tasks. This participant might do fine on FR, but poorly on CR. Alternatively, a 'high-performing' CR participant (i.e., a 'skilled memorizer'; Thomas et al, 2023) might adopt some combination of a seriation-based rehearsal strategy and a more effective strategy like imagery. Because the nature of our FR task (unrelated words) may have privileged seriation strategies, the high-performer's accuracy might be primarily determined by their effective use of rehearsal–a strategy that at least intuitively seems to be less prone to individual differences than something like imagery. So they might do a bit better than the low-performing participant. But for CR, the scales flip, and the potentially more idiosyncratic strategies like imagery carry more weight, resulting in much higher performance than the low-performing participant. By introducing otherwise low-performing participants to the effective imagery strategy, we changed the situation from one in which CR variability was being driven primarily by large qualitative differences in *strategy choice* to one in which CR variability is driven by perhaps smaller quantitative individual differences in *strategy use effectiveness*. The apparent increase in FR variability that we observed with the imagery manipulation and highly-imageable words is consistent with this explanation–we provided opportunities for participants to use more than serial information, and in doing so allowed for more variability in both strategy choice and strategy use effectiveness.

What are the potential implications of this conclusion? First, it suggests that individual differences in paired-associates learning are primarily driven by individual differences in adopted memory strategies (and that individual differences can be reduced by

instructing participants to use an effective strategy). It also suggests that differential strategy adoption may have more of an effect on paired-associates than free recall tasks. An interesting upshot of this is that CR-type tasks might be better suited for discriminating among users of different strategies than FR-type tasks. Another is that some tasks and contexts (like FR) may mask individual differences, while others (like CR) may better allow them to proliferate.

We did not directly test whether it is strategies at *test* or at *encoding* that primarily explain the effect. In Experiment 6, we found some evidence that absent any strategy instructions, participants' "strategy profiles" might be more variable for FR than CR at study, but vice versa at test. This might suggest that variability in strategy use at test explains the CR:FR variability effect. In Experiment 7, although we instructed participants to use the imagery strategy prior to study, we also reminded them to use mental imagery prior to test. Thus, the imagery strategy manipulation could have acted at either (or both) points. Certainly one might expect that participants adopt a consistent strategy at study and test, but even within a particular strategy participants likely vary in how they apply it at test. For example, participants who used a rehearsal-based strategy for FR may initiate serial recall of words in which they are highly confident, then use those words to cue subsequent searches (i.e., something like SAM posits). And participants who used an imagery-based strategy for CR might conjure up the images they generated at study whilst also vocalizing cues. Overall, it is highly likely that there are substantial individual differences *within* particular strategies. We do not make any strong claims about whether strategies modulate our effect primarily at study, at test, or both. Our argument is that individual differences in CR performance are likely due to variability in *general* memory strategy use.

In most of our experiments, average CR performance was lower than FR performance, and in our within-subjects experiments, participants differed more from

themselves on CR than FR. We took this as potential evidence that participants may generally be less familiar with the task of memorizing *pairs* of unrelated words than memorizing *lists* of unrelated words. Indeed, it seems easier to think of everyday tasks resembling the latter (e.g., shopping lists, locations along a route) that might contribute to more general experience with free recall than cued recall. However, this might also imply that the 'high-performing' CR participants may have done well not only due to their adoption of effective strategies but potentially also due to more experience with paired-associates-like tasks. One intriguing avenue for future research would be to further examine high- and low-performing CR participants (e.g., with qualitative questions) to better understand the factors that differentiate them. Is it that high-performers have more experience with memory tasks in general, that they are better able to flexibly adopt different strategies, etc.?

These results also have implications for theories of memory. When we fit the basic Search of Associative Memory (SAM; Raaijmakers & Shiffrin, 1981) model to our initial data, we found that although it could fit the data, it did not offer a theoretically satisfying account of the CR:FR variability difference. Part of this is almost certainly due to the limited, exploratory nature of our modelling exploits, but it also seems clear that a hierarchical/individual-differences version of the model is necessary to explain the effects we observed here. For instance, a version of SAM in which separate item-item association parameters are tuned separately for FR and CR for each participant seems like it could capture our intuitions about individual differences in strategy use. One might expect that in FR, such an association parameter might vary less across participants (reflecting reliance on rehearsal/serial information) than it would for CR (reflecting individual differences in imagery).

Implicit in our argument about strategy use is the notion that the associations formed in unrelated FR lists and within unrelated CR pairs are qualitatively different. This idea is not

directly captured by SAM, but other models do formally instantiate this idea in the form of pair associations stored as convolutions or combinative transformations of individual pair member representations (e.g., TODAM, Pike's *matrix model*, CHARM; Osth & Dennis, 2020). Perhaps in these models, individual difference parameters representing the *degree* to which pairs are convolved or transformed (i.e., individual differences in reliance on effective imagery or less-effective rehearsal strategies) might allow these models to readily explain the CR:FR variability effect. Because our initial modelling efforts here focused solely on SAM, further tests with these other models are necessary to fully explore the implications of our results for theories of memory more broadly. Regardless of the veracity of our effect and its specific implications for models of memory, we do not think it controversial in joining others that have argued for the importance of incorporating individual differences in modern computational models of memory (Haile et al., 2024; Lee & Webb, 2005).

Constraints on Generality

The general claims and hypotheses about memory that we draw from our results must also be considered in the context of the experimental designs, stimuli, and samples that we used. Specifically, the particular choices we made in these regards place some *constraints on generality* (Simons et al., 2017) on our conclusions. Although we used a wide variety of wordsets (e.g., nouns, animals, objects, highly-imageable words) and two different participant pools (i.e., undergraduates, online community participants), the free and cued recall tasks we used in all experiments were quite artificial (i.e., words with certain characteristics presented singly or in word-pairs one at a time in random order with instructions to remember them for a subsequent test, followed by a brief retention interval and then tests of the sorts we have reported). The results we observed might be different from those obtained from experiments using more ecologically valid memory tasks. For example, researchers have measured free recall performance 'in the wild' using shopping-list and local-landmark paradigms that more

closely resemble the kinds of episodic memory tasks individuals encounter in daily life (Barnett et al., 2023; Ross et al., 2004). Similarly, others have examined cued recall using face-name paradigms (Crumley et al., 2014). It may be that for tasks like face-name or cued object-location recall–common everyday cued memory tasks–we would not observe greater variability for CR than for FR. And with FR tasks like shopping-list recall–a task in which lists have meaningful structure outside of serial order–we might even observe greater variability for FR than for CR. And we make no strong claims about whether the effects observed here extend to other stimulus modalities like audio or images (e.g., Kazanas et al., 2020), or to other cultures or age groups. Our point here is not to claim that we have a theory that predicts different patterns of variability for free and cued recall as a function of the materials, procedures, or participants. Would that we did. We are merely acknowledging potential constraints on the generality of our findings (Simons et al., 2017).

Conclusion

Given these constraints, what can we ultimately conclude about variability in cued and free recall? First, these experiments speak to the influence that individual differences play in memory–and the importance of examining those differences both in experiments and in our formal models of memory. Indeed, some of the strikingly variable distributions we observed in our cued recall data stand as a warning to researchers examining these tasks: ignore variability at your own peril! A historical focus only on mean performance and intervals around the mean violates Newell's (1973) 'Second Injunction of Psychological Experimentation': "Never average over methods", a practice that "conceals rather than reveals" (p. 13). Our research demonstrates the value of examining both the performance of individuals and the overall distribution.

Our results align with prior work showing that the strategies participants adopt at encoding and retrieval are powerful determinants of later recall (Unsworth, 2016). We also

argue that task constraints can influence the variety of strategies available to participants (e.g., random FR may limit participants to a smaller set of strategies), and that there may be more individual differences *within* some strategies than others (e.g., more individual differences in effective use of imagery vs. rehearsal). This follows Newell's (1973) 'First Injunction of Psychological Experimentation': "Know the method your subject is using to perform the experimental task" (p. 12). Understanding how each participants interacts with the particular constraints of a given task is vital for interpreting results derived from that task.

Third, these results suggest some potential avenues for future research—examining individual differences to better determine *why* particular participants fall where they do on the CR distribution, asking participants qualitative questions (perhaps during encoding and recall), probing for differences in relevant experience (e.g., more experience with CR-like tasks or prior exposure to effective memory strategies), testing the effect of different kinds of strategy instructions on individual differences (e.g., increasing or decreasing individual differences using different strategies), extending this work to more ecologically valid recall tasks and other cognitive models of memory, etc. The current work–and these possible avenues–have the potential to provide greater insight into human memory.

References

Anderson, J. R. (1981). Interference: The relationship between response latency and response accuracy. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(5), 326–343. https://doi.org/10.1037/0278-7393.7.5.326

Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (2013). *The Architecture of Cognition* (0 ed.). Psychology Press. https://doi.org/10.4324/9781315799438

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An Integrated Theory of List Memory. *Journal of Memory and Language*, *38*(4), 341–380. https://doi.org/10.1006/jmla.1997.2553

Anderson, J. R., Lebiere, C., Lovett, M., & Reder, L. (1998). ACT-R: A higher-level account of processing capacity. *Behavioral and Brain Sciences*, *21*(6), 831–832. https://doi.org/10.1017/S0140525X98221765

Anderson, J. R., & Matessa, M. (1997). A production system theory of serial memory. *Psychological Review*, *104*(4), 728–748. https://doi.org/10.1037/0033-295X.104.4.728

Anderson, J. R., & Reder, L. M. (1999). The fan effect: New results and new theories. *Journal of Experimental Psychology: General*, *128*(2), 186–197. https://doi.org/10.1037/0096-3445.128.2.186

Atkins, P. W. B. (2001). What happens when we relearn part of what we previously knew? Predictions and constraints for models of long-term memory. *Psychological Research*, *65*(3), 202–215. https://doi.org/10.1007/s004269900015

Barnett, M. D., Hardesty, D. R., Griffin, R. A., & Parsons, T. D. (2023). Performance on a virtual environment shopping task and adaptive functioning among older adults.

Journal of Clinical and Experimental Neuropsychology, *45*(5), 464–472. https://doi.org/10.1080/13803395.2023.2249175

Begg, I. (1978). Imagery and organization in memory: Instructional effects. *Memory & Cognition*, *6*(2), 174–183. https://doi.org/10.3758/BF03197443

Blajenkova, O., Kozhevnikov, M., & Motes, M. A. (2006). Object-spatial imagery: A new self-report imagery questionnaire. *Applied Cognitive Psychology*, *20*(2), 239–263. https://doi.org/10.1002/acp.1182

Bower, G. H., & Winzenz, D. (1970). Comparison of associative learning strategies. *Psychonomic Science*, *20*(2), 119–120. https://doi.org/10.3758/BF03335632

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

Caplan, J. B., & Madan, C. R. (2016). Word Imageability Enhances Association-memory by Increasing Hippocampal Engagement. *Journal of Cognitive Neuroscience*, *28*(10), 1522–1538. https://doi.org/10.1162/jocn_a_00992

Christensen, H., Mackinnon, A. J., Korten, A. E., Jorm, A. F., Henderson, A. S., Jacomb, P., & Rodgers, B. (1999). An analysis of diversity in the cognitive performance of elderly community dwellers: Individual differences in change scores as a function of age. *Psychology and Aging*, *14*(3), 365–379. https://doi.org/10.1037/0882-7974.14.3.365

Cleary, A.M. (2018). Dependent measures in memory research: From free recall to recognition. In H. Otani & B.L. Schwartz (Eds.), *Handbook of Research Methods in Human Memory*. New York: Routledge.

Cowan, N., Saults, J. S., Elliott, E. M., & Moreno, M. V. (2002). Deconfounding Serial Recall. *Journal of Memory and Language*, *46*(1), 153–177. https://doi.org/10.1006/jmla.2001.2805

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

Crumley, J. J., Stetler, C. A., & Horhota, M. (2014). Examining the relationship between subjective and objective memory performance in older adults: A meta-analysis. *Psychology and Aging*, *29*(2), 250–263. https://doi.org/10.1037/a0035908

de Jonge, M., Tabbers, H. K., Pecher, D., & Zeelenberg, R. (2012). The effect of study time distribution on learning and retention: A Goldilocks principle for presentation rate. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(2), 405–412. https://doi.org/10.1037/a0025897

Deese, J. (1957). SERIAL ORGANIZATION IN THE RECALL OF DISCONNECTED ITEMS. *Psychological Reports*, *3*(7), 577. https://doi.org/10.2466/PR0.3.7.577-582

Delaney, P. F., Godbole, N. R., Holden, L. R., & Chang, Y. (2018). Working memory capacity and the spacing effect in cued recall. *Memory*, *26*(6), 784–797. https://doi.org/10.1080/09658211.2017.1408841

Ensor, T. M., Guitard, D., Bireta, T. J., Hockley, W. E., & Surprenant, A. M. (2020). The list-length effect occurs in cued recall with the retroactive design but not the proactive design. *Canadian Journal of Experimental Psychology = Revue Canadienne De Psychologie Experimentale*, *74*(1), 12–24. https://doi.org/10.1037/cep0000187

Eronen, M. I., & Bringmann, L. F. (2021). The Theory Crisis in Psychology: How to Move Forward. *Perspectives on Psychological Science*, *16*(4), 779–788. https://doi.org/10.1177/1745691620970586

García-Pérez, M.A. Statistical criteria for parallel tests: A comparison of accuracy and power. *Behav Res* **45**, 999–1010 (2013). https://doi.org/10.3758/s13428-013-0328-z

Goldsmith, M., & Koriat, A. (2007). The strategic regulation of memory accuracy and informativeness. In A. S. Benjamin & B. H. Ross (Eds.), *Skill and strategy in memory use.* (Vol. 48, pp. 1–60). Elsevier Academic Press.

Haile, T. M., Prat, C. S., & Stocco, A. (2024). One Size Does Not Fit All: Idiographic Computational Models Reveal Individual Differences in Learning and Meta-Learning Strategies. *Topics in Cognitive Science*, tops.12730. https://doi.org/10.1111/tops.12730

Handbook of Research Methods in Human Memory and Cognition. (1982). Elsevier. https://doi.org/10.1016/C2013-0-11333-5

Hartigan, J. A., & Hartigan, P. M. (1985). The Dip Test of Unimodality. *The Annals of Statistics*, *13*(1). https://doi.org/10.1214/aos/1176346577

Hatchavanich, D. (2014). A comparison of type I error and power of Bartlett's test, Levene's test and O'Brien's test for homogeneity of variance tests. *Southeast Asian Journal of Sciences, 3*(2), 181-194.

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

Hockley, W. E., & Cristi, C. (1996). Tests of encoding tradeoffs between item and associative information. *Memory & Cognition*, *24*(2), 202–216. https://doi.org/10.3758/BF03200881

Huber, D. E., Tomlinson, T. D., Jang, Y., & Hopper, W. J. (2015). The Search of Associative Memory with Recovery Interference (SAM-RI) memory model and its application to retrieval practice paradigms. In J. G. W. Raaijmakers, A. H. Criss, R. L. Goldstone, R. M. Nosofsky, & M. Steyvers (Eds.), *Cognitive modeling in perception and memory: A festschrift for Richard M. Shiffrin* (pp. 81–98). Psychology Press.

Jamieson, R. K., Mewhort, D. J. K., & Hockley, W. E. (2016). A computational account of the production effect: Still playing twenty questions with nature. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *70*(2), 154–164. https://doi.org/10.1037/cep0000081

Kahana, M. J. (2020). Computational Models of Memory Search. *Annual Review of Psychology*, *71*(1), 107–138. https://doi.org/10.1146/annurev-psych-010418-103358

Kazanas, S. A., Altarriba, J., & O'Brien, E. G. (2020). Paired-associate learning, animacy, and imageability effects in the survival advantage. *Memory & Cognition*, *48*(2), 244–255. https://doi.org/10.3758/s13421-019-01007-2

Kluger, F. E., Oladimeji, D. M., Tan, Y., Brown, N. R., & Caplan, J. B. (2022). Mnemonic scaffolds vary in effectiveness for serial recall. *Memory*, *30*(7), 869–894. https://doi.org/10.1080/09658211.2022.2052322

Lakens, D. (2017). Equivalence Tests: A Practical Primer for *t* Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177

LaPlume, A. A., Paterson, T. S. E., Gardner, S., Stokes, K. A., Freedman, M., Levine, B., Troyer, A. K., & Anderson, N. D. (2021). Interindividual and intraindividual variability in amnestic mild cognitive impairment (aMCI) measured with an online cognitive assessment. *Journal of Clinical and Experimental Neuropsychology*, 1–17. https://doi.org/10.1080/13803395.2021.1982867

Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review*, *12*(4), 605–621. https://doi.org/10.3758/BF03196751

Madan, C. R., Caplan, J. B., Lau, C. S. M., & Fujiwara, E. (2012). Emotional arousal does not enhance association-memory. *Journal of Memory and Language*, *66*(4), 695–716. https://doi.org/10.1016/j.jml.2012.04.001

Madan, C. R., Glaholt, M. G., & Caplan, J. B. (2010). The influence of item properties on association-memory. *Journal of Memory and Language*, *63*(1), 46–63. https://doi.org/10.1016/j.jml.2010.03.001

Madigan, S. (1980). The serial position curve in immediate serial recall. *Bulletin of the Psychonomic Society*, *15*(5), 335–338. https://doi.org/10.3758/BF03334550

Mah, E.Y., Campbell, A., Tamburri, C., Grannon, K., & Lindsay, D.S. (2023). A direct replication of Popp and Serra (2016, Experiment 1): Better free recall and worse cued recall of animal names than object names. *Frontiers in Psychological Science, 14.* doi:10.3389/fpsyg.2023.1146200

McKelvie, S. J. (1995). The VVIQ as a psychometric test of individual differences in visual imagery vividness: A critical quantitative review and plea for direction. *Journal of Mental Imagery*, *19*(3-4), 1–106.

Metcalfe, J. (1990). Composite Holographic Associative Recall Model (CHARM) and blended memories in eyewitness testimony. *Journal of Experimental Psychology: General*, *119*(2), 145–160. https://doi.org/10.1037/0096-3445.119.2.145

Morgan, W.A. (1939). A test for the significance of the difference between two variances in a sample from a normal bivariate distribution. *Biometrika*, *31*, 13-19.

Morrison, A. B., Rosenbaum, G. M., Fair, D., & Chein, J. M. (2016). Variation in strategy use across measures of verbal working memory. *Memory & Cognition*, *44*(6), 922–936. https://doi.org/10.3758/s13421-016-0608-9

Mudholkar, G. S., Wilding, G. E., & Mietlowski, W. L. (2003). Robustness Properties of the Pitman–Morgan Test. *Communications in Statistics - Theory and Methods*, *32*(9), 1801–1816. https://doi.org/10.1081/STA-120022710

Murdock, B. B. (1983). A distributed memory model for serial-order information. *Psychological Review*, *90*(4), 316–338. https://doi.org/10.1037/0033-295X.90.4.316

Murdock, B. B. (1995). Developing TODAM: Three models for serial-order information. *Memory & Cognition*, *23*(5), 631–645. https://doi.org/10.3758/BF03197264

Nairne, J. S., Riegler, G. L., & Serra, M. (1991). Dissociative effects of generation on item and order retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(4), 702–709. https://doi.org/10.1037/0278-7393.17.4.702

Nairne, J.S., VanArsdall, J.E., & Cogdill, M. (2017). Remembering the living: Episodic memory is tuned to animacy. *Current Directions in Psychological Science, 26*(1), 22-27. doi:10.1177/0963721416667711

Newall, A. (1974). You can't play 20 questions with nature and win. *Visual information processing*.

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*(5), 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

Osth, A. F., & Dennis, S. (2020). *Global matching models of recognition memory*. https://doi.org/10.31234/osf.io/mja6c

Othman, A. R., Keselman, H. J., Wilcox, R. R., Algina, J., Fradette, K., & Teh, S. Y. (2007). Robust Levene Tests for Variance Equality. *Bulletin of the International Statistical Institute 56th Session: Proceedings [CD-ROM]*.

Pastore, M., Alaimo Di Loro, P., Mingione, M., & Calcagni, A. (2022). overlapping: Estimation of Overlapping in Empirical Distributions. R package version 2.1. https://CRAN.R-project.org/package=overlapping

Pearson, D. G., Deeprose, C., Wallace-Hadrill, S. M. A., Heyes, S. B., & Holmes, E. A. (2013). Assessing mental imagery in clinical psychology: A review of imagery measures and a guiding framework. *Clinical Psychology Review*, *33*(1), 1–23. https://doi.org/10.1016/j.cpr.2012.09.001

Pike, R. (1984). Comparison of convolution and matrix distributed memory systems for associative recall and recognition. *Psychological Review*, *91*(3), 281–294. https://doi.org/10.1037/0033-295X.91.3.281

Pitman, E.J.G. (1939). A note on normal correlation. *Biometrika, 31*, 9-12.

Popp, E. Y., & Serra, M. J. (2016). Adaptive memory: Animacy enhances free recall but impairs cued recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(2), 186–201. https://doi.org/10.1037/xlm0000174

Postman, L., & Underwood, B. J. (1973). Critical issues in interference theory. *Memory & Cognition*, *1*(1), 19–40. https://doi.org/10.3758/BF03198064

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

Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM: A Theory of Probabilistic Search of Associative Memory. In *Psychology of Learning and Motivation* (Vol. 14, pp. 207–262). Elsevier. https://doi.org/10.1016/S0079-7421(08)60162-0

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. https://doi.org/10.1037/a0023810

Reisberg, D., Pearson, D. G., & Kosslyn, S. M. (2003). Intuitions and introspections about imagery: The role of imagery experience in shaping an investigator's theoretical views. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *17*(2), 147-160.

Roediger, H. L., & Payne, D. G. (1985). Recall criterion does not affect recall level or hypermnesia: A puzzle for generate/recognize theories. *Memory & Cognition*, *13*(1), 1–7. https://doi.org/10.3758/BF03198437

Roediger, H. L., Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review*, *8*(3), 385–407. https://doi.org/10.3758/BF03196177

Ross, M., Spencer, S. J., Linardatos, L., Lam, K. C. H., & Perunovic, M. (2004). Going shopping and identifying landmarks: Does collaboration improve older people's memory? *Applied Cognitive Psychology*, *18*(6), 683–696. https://doi.org/10.1002/acp.1023

Sahadevan, S. S., Chen, Y. Y., & Caplan, J. B. (2021). Imagery-based strategies for memory for associations. *Memory*, *29*(10), 1275–1295. https://doi.org/10.1080/09658211.2021.1978095

Schmidt, J. P., Tombaugh, T. N., & Faulkner, P. (1992). Free-recall, cued-recall and recognition procedures with three verbal memory tests: Normative data from age 20 to 79. *Clinical Neuropsychologist*. https://doi.org/10.1080/13854049208401855

Siedlecki, K. L. (2007). Investigating the structure and age invariance of episodic memory across the adult lifespan. *Psychology and Aging*, *22*(2), 251–268. https://doi.org/10.1037/0882-7974.22.2.251

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on Generality (COG): A Proposed Addition to All Empirical Papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630

Suggate, S., & Lenhard, W. (2022). Mental imagery skill predicts adults' reading performance. *Learning and Instruction*, *80*, 101633. https://doi.org/10.1016/j.learninstruc.2022.101633

Tan, L., Ward, G., Paulauskaite, L., & Markou, M. (2016). Beginning at the beginning: Recall order and the number of words to be recalled. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *42*(8), 1282–1292. https://doi.org/10.1037/xlm0000234

Thomas, J. J., Ayuno, K. C., Kluger, F. E., & Caplan, J. B. (2023). The relationship between

interactive-imagery instructions and association memory. *Memory & Cognition*, *51*(2), 371–390. https://doi.org/10.3758/s13421-022-01347-6

Unsworth, N. (2016). Working memory capacity and recall from long-term memory: Examining the influences of encoding strategies, study time allocation, search efficiency, and monitoring abilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(1), 50–61. https://doi.org/10.1037/xlm0000148

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2011). Inter- and intra-individual variation in immediate free recall: An examination of serial position functions and recall initiation strategies. *Memory*, *19*(1), 67–82. https://doi.org/10.1080/09658211.2010.535658

Unsworth, N., Miller, A. L., & Robison, M. K. (2019). Individual differences in encoding strategies and free recall dynamics. *Quarterly Journal of Experimental Psychology*, *72*(10), 2495–2508. https://doi.org/10.1177/1747021819847441

van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. Perspectives on Psychological Science, 16(4), 682–697. https://doi.org/10.1177/1745691620970604

Watkins, M. J. (1990). Mediationism and the obfuscation of memory. *American Psychologist*, *45*(3), 328–335. https://doi.org/10.1037/0003-066X.45.3.328

Wilcox, R. (2015). Comparing the variances of two dependent variables. *Journal of Statistical Distributions and Applications*, *2*(1), 7. https://doi.org/10.1186/s40488-015-0030-z

Wilson, M. (1988). MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, *20*(1), 6–10. https://doi.org/10.3758/BF03202594

Wilson, J. H., & Criss, A. H. (2017). The list strength effect in cued recall. *Journal of Memory and Language*, *95*, 78–88. https://doi.org/10.1016/j.jml.2017.01.006

Wittmann, B. C., & Şatırer, Y. (2022). Decreased associative processing and memory confidence in aphantasia. *Learning & Memory*, *29*(11), 412–420. https://doi.org/10.1101/lm.053610.122

Yao, C., Stawski, R.S., Hultsch, D.F., & McDonald, S.W.S. (2016). Selective attrition and intraindividual variability in response time moderate cognitive change. *Journal of Clinical and Experimental Neuropsychology*, *38*(2), 227-237. DOI: 10.1080/13803395.2015.1102869

Zerr, C. L., Berg, J. J., Nelson, S. M., Fishell, A. K., Savalia, N. K., & McDermott, K. B. (2018). Learning Efficiency: Identifying Individual Differences in Learning Rate and Retention in Healthy Adults. *Psychological Science*, *29*(9), 1436–1450. https://doi.org/10.1177/0956797618772540

Supplementary Material

1. Coding scheme and method for self-reported strategy data

Self-reported study strategies. After all study-test cycles, we asked participants to self-report the study strategies they used for FR and CR. We had no a priori predictions about these self-reports, but it is possible that participants use a greater variety of strategies for CR than FR (and that this may account for the greater variability in performance). Strategy data were coded by two coders using a scheme developed in Mah et al. (in preparation, see Supplementary Material X for the coding categories). For each answer to the separate strategy use questions for FR and CR, coders selected up to two strategies that best fit the participant's response.

Study strategy	**If the participant mentioned more than one strategy, code the first and second strategies mentioned (i.e., in "Strategy1Code" and "Strategy2Code"). If there were more than two strategies coded, only code the first two mentioned.**
Imagery (picture, visualize, method of loci)	If there is mention of an interaction between imagined things, use the "Story" category instead
Rehearsal/Repetition	"Saying aloud" counts for this one
Relate word to something in one's life	
Categorize	

Syntactic strategy	If any non-semantic features of words are used (e.g., letters, acronyms). Rhyming counts for this.
Tell story, narrative, song, movie linking words together	
Acting out	E.g., "pantomiming", "hand gestures"
General associative strategy	Use for non-specific associative strategies (e.g., "I tried to link the words", "I tried to associate the words")
No response or no strategy	Vague strategies (e.g., "I tried to remember") count for this one. **Don't use this for people who report one strategy but not two (i.e., leave the 2nd strategy empty instead of coding it as this).**
Other	

2. Bayesian computational models

For our Bayesian analyses, we fit and compared two computational models to CR and FR accuracy data–one model assuming that FR and CR performance came from normal distributions with differing means but the same variance, and one model assuming that FR and CR performance came from distributions with differing means and variances (models adapted from Gelman et al., 2013). If FR and CR differed in their variability, the latter model should show improved fit. Models were fit using cmdstanr (Gabry & Cesnovar, 2021), and used preregistered priors based on FR and CR means and standard deviations observed in Mah et al. (in preparation). Model fits were compared using Pareto-Smoothed Importance

Sampling Leave-one-out Cross-Validation (PSIS-LOO; Vehtari et al., 2017). PSIS-LOO is a Bayesian measure of a model's ability to predict hypothetical new data.

3. Experiment 1 Supplementary Results

A. **Bayesian computational model analysis.** The model comparison supported our NHST results–the model with *differing FR/CR variances* was superior to the model with the *equal FR/CR variances*, ΔLOO = 4.75, 95% CI[.37, 9.14][19].

B. **Generalized mixed-effects logistic regression results.** This exploratory analysis involved the computation of the following multilevel model:

Level 1: $\text{logit}(y_{ti}) = \beta_{0i} + \beta_{1i}(R_{ti}) + e_{ti}$

Level 2: $\beta_{0i} = \gamma_{00} + \mu_{0i}$

$\beta_{1i} = \gamma_{10} + \mu_{1i}$

Where *t* refers to trial/word, *i* refers to individual participants, and *R* refers to test type (Baseline = FR). Thus, we predicted item-level performance from test type, and estimated inter-individual variability in FR proportion recalled as well as the difference between FR proportion recalled and CR proportion recalled. By fitting this model twice (once with FR as the baseline reference category and once with CR as the baseline reference category), we were able to obtain estimates and profile 95% CIs on the inter-individual variability in proportion recalled for both memory types:

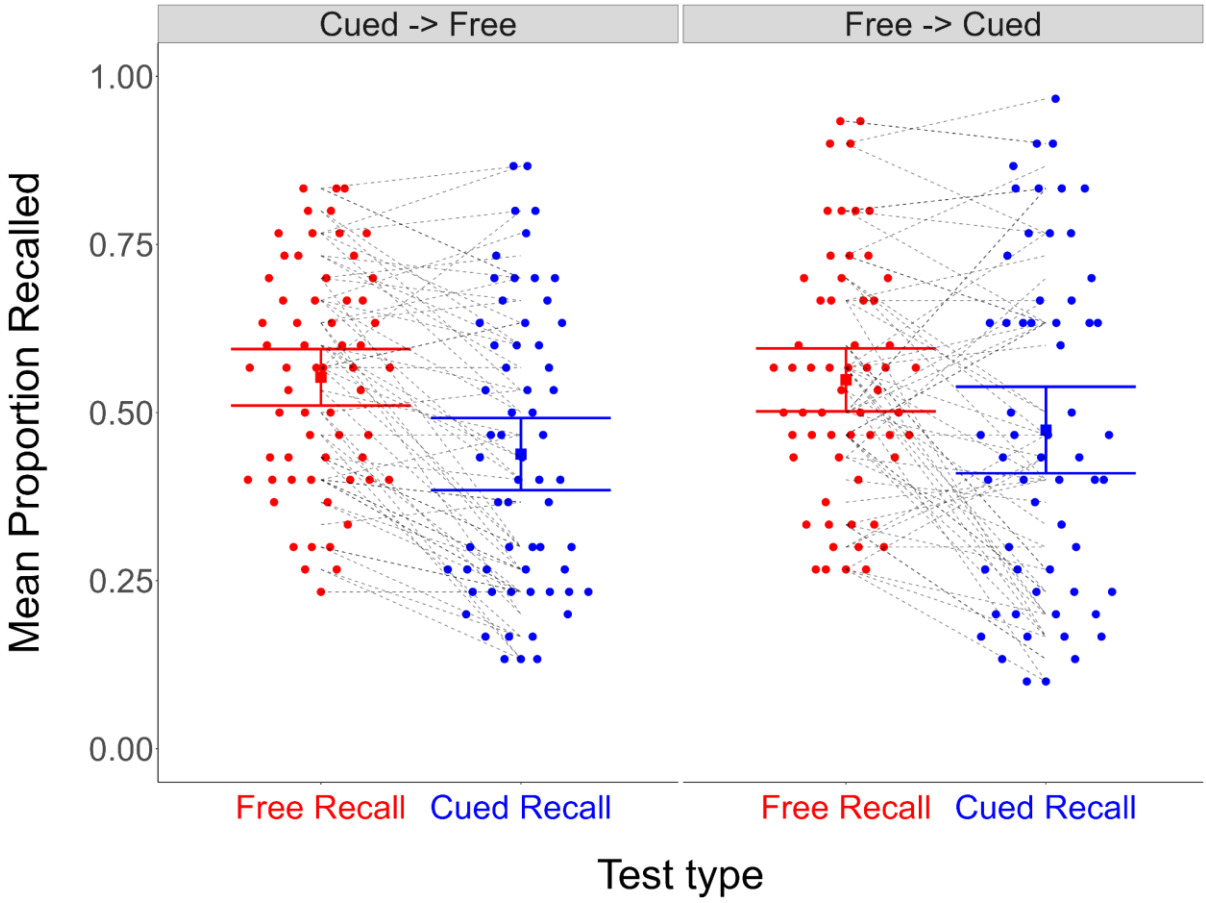
Test type	***SD* (Logit units)**	**95% CI lower**	**95% CI upper**

[19] 95% CI on the LOO difference was estimated by multiplying the SE of the difference by +/- 1.96.

FR	.65	.54	.78
CR	1.00	.85	1.17

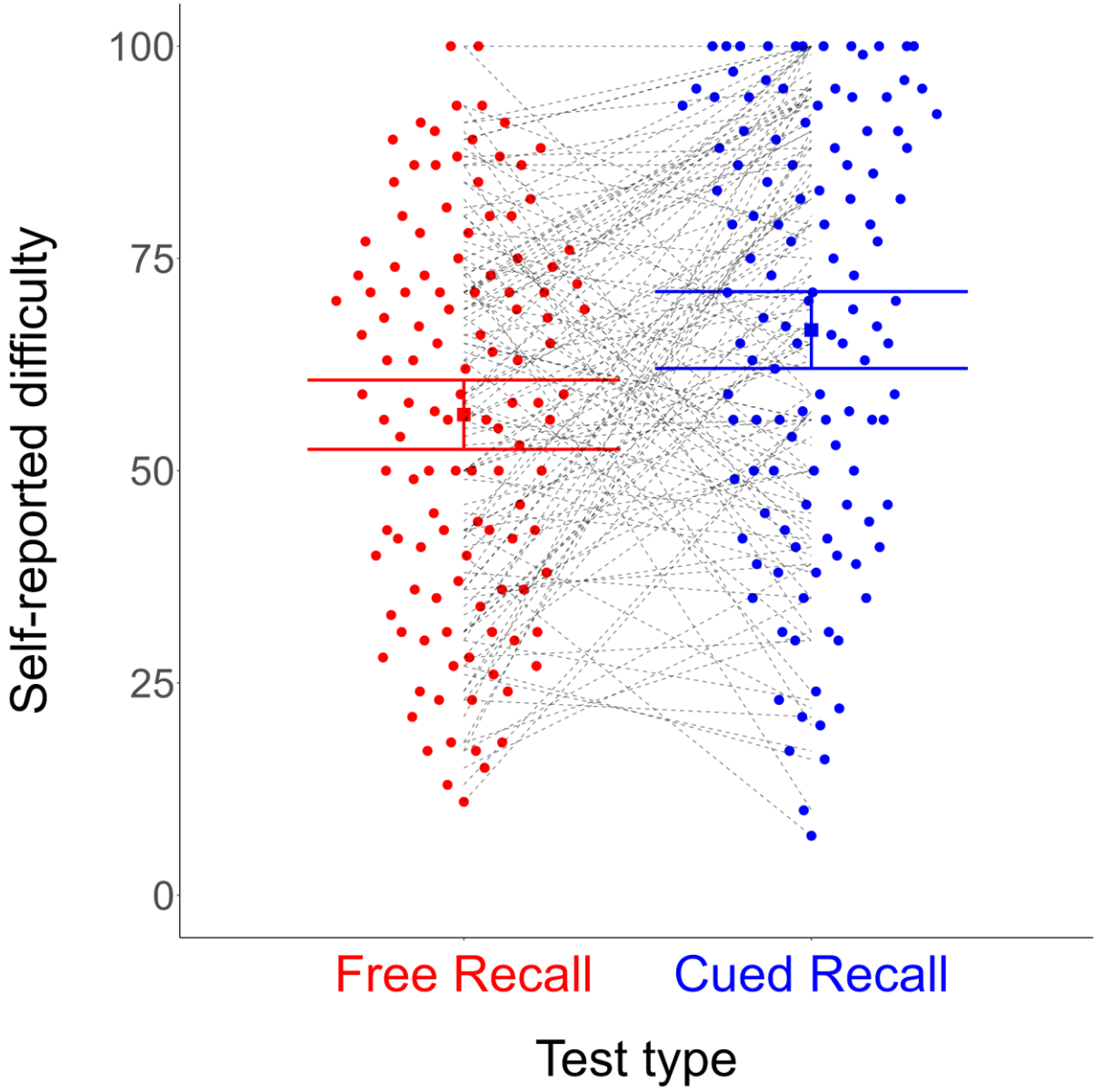
As the 95% CIs on the SD estimates did not overlap, this analysis provided evidence for differing FR/CR variances.

C. **Order effects.** 61 participants completed CR before FR, and 59 completed FR before CR. We examined the possible influence of task order on variability differences and overall accuracy. First, we conducted separate Pitman-Morgan tests in each order condition. The test was significant in the CR -> FR group, $t(59) = 2.84$, $p = .006$, and also in the FR -> CR group, $t(57) = 3.10$, $p = .003$. The bootstrapped CR:FR variance ratio in the CR -> FR group was 1.28 (95% percentile bootstrap CI [1.07, 1.53]), and in the FR -> CR group it was 1.38 (95% percentile bootstrap CI [1.14, 1.69]). Accuracy by test type and test order is shown in the figure below:



In an exploratory analysis of accuracy including both test type and test order factors, the interaction between test type and test order was not significant, $\chi^2(1) = 1.74$, $p = .19$.

D. Self-reported recall difficulty. Similar to the strategy questions, participants were asked to provide numerical self-reports of recall difficulty (0 = *Very Easy*, 100 = *Very Hard*). Although we had no a priori predictions about these self-reports, they provided us with the opportunity to test whether subjective impressions of recall were more variable for CR than for FR. Recall difficulty ratings are displayed in the figure below:



Note. Error bars = 95% CIs (within-subjects). Points jittered horizontally based on relative frequency. Dashed lines connect FR and CR difficulty ratings for individual participants.

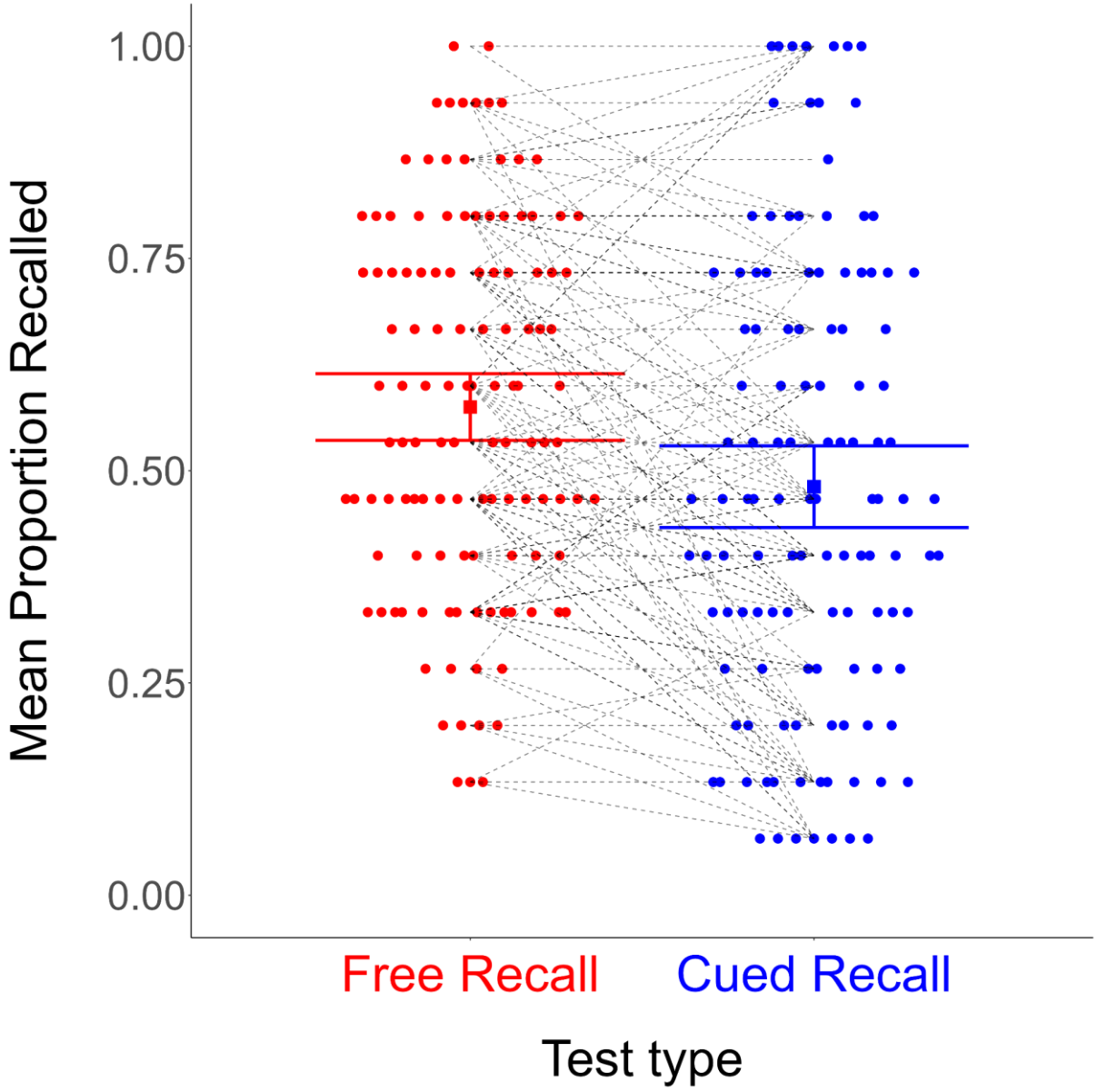
Other than a potential ceiling effect for CR difficulty ratings, there is no immediately obvious difference in variability ratings. Indeed, a Pitman-Morgan test failed to reject the null hypothesis of equal variances, $t(118) = 1.15$, $p = .25$, with a bootstrapped CR:FR variance ratio = 1.11 (95% percentile bootstrap CI [.97, 1.26]). The corresponding Bayesian model

comparison slightly favoured the model with the model with *equal FR/CR variances* over the model with *differing FR/CR variances*, but the difference was not statistically reliable, ΔLOO = .07, 95% CI [-1.40, 1.54]. Though this difference was slight, both models provided very similar predictions when FR/CR variances are close (making it difficult to observe a clear advantage for the equal-variances model). Thus, any instances where the equal-variances model is even slightly favoured may be reasonably interpreted as support for the equal-variances model on the grounds of parsimony.

4. Experiment 2A Supplementary Results

A. Bayesian computational modelling analysis. The model with *differing FR/CR variances* was only slightly favoured over the model with *equal FR/CR variances*, ΔLOO = 2.42 (*SE* = 1.76, 95% CI [-1.02, 5.88]), with the 95% CI on the difference containing 0.

B. Treating CR responses as correct as long as they were a target. We then conducted a version of our primary analysis where we treated CR responses as correct as long as they matched a target from the studied list (i.e., treating the CR test like an FR test). These results are shown in the figure below:



Note. Error bars = 95% CIs (within-subjects). Points jittered horizontally based on relative frequency. Dashed lines connect FR and CR performance for individual participants.

The results were quite similar, suggesting that recall of targets to incorrect cues was relatively uncommon (60/994 CR commission errors). FR and CR variances still significantly differed, Pitman-Morgan $t(118) = 2.75$, $p = .007$, with a bootstrapped CR:FR variance ratio of 1.23, (95% percentile bootstrap CI [1.07, 1.41]). The Bayesian model comparison results were also similar, with the model with *differing FR/CR variances* slightly favoured over the

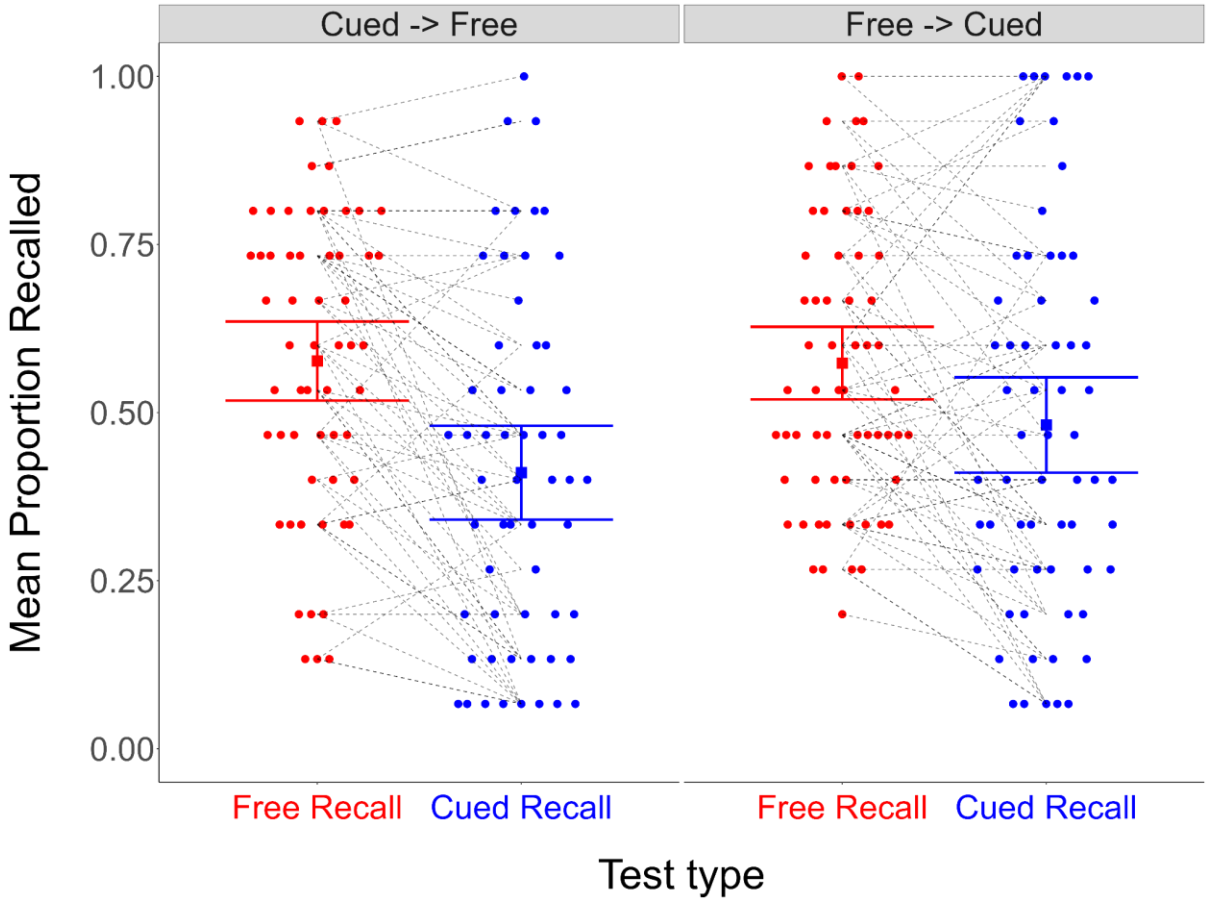
model with *equal FR/CR variances*, ΔLOO = 1.07 (*SE* = 1.27, 95% CI [-1.43, 3.57]), with the 95% CI containing 0.

C. Generalized mixed-effects logistic regression results. We used the same GLMM model as in Experiment 1 to compare FR and CR variability (for standard CR proportion recalled and proportion recalled when same-list commission errors were counted as correct):

Test type	***SD* (Logit units)**	**95% CI lower**	**95% CI upper**
FR	.84	.69	1.03
CR	1.31	1.10	1.57
CR (commission)	1.22	1.02	1.46

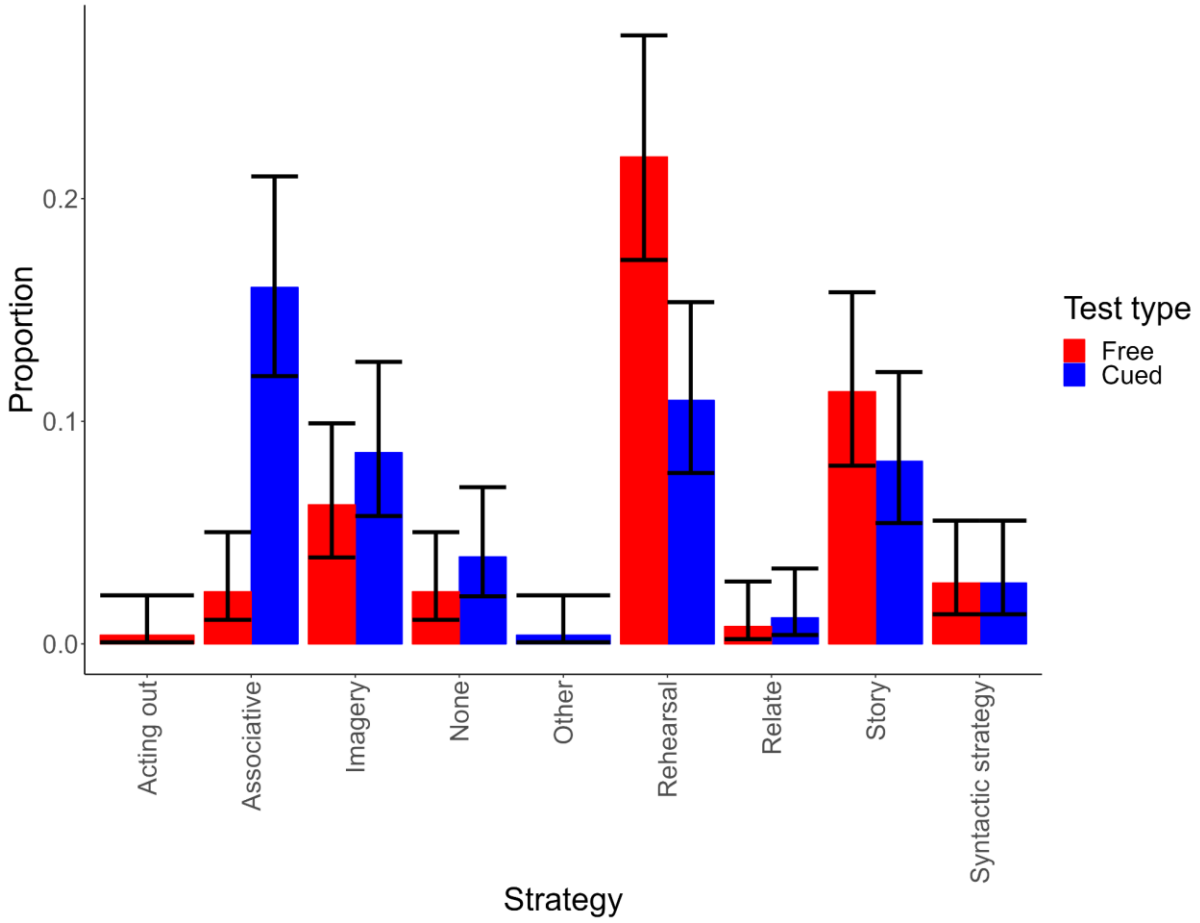
As the 95% CIs on the *SD* estimates did not overlap, this analysis provided evidence for differing FR/CR variances.

D. Order effects: 57 participants completed CR before FR, and 63 completed FR before CR. We examined the possible influence of task order on variability differences and overall accuracy. First, we conducted separate Pitman-Morgan tests in each order condition. The test was non-significant in the CR -> FR group, $t(55) = 1.55$, $p = .13$, but was significant in the FR -> CR group, $t(61) = 2.62$, $p = .01$. The bootstrapped CR:FR variance ratio in the CR -> FR group was 1.19 (95% percentile bootstrap CI [.97, 1.46]), and in the FR -> CR group it was 1.32 (95% percentile bootstrap CI [1.09, 1.58]). Accuracy by test type and test order is shown in the figure below:



In an exploratory analysis of accuracy including both test type and test order factors, the interaction between test type and test order was not significant, $\chi^2(1) = 3.59$, $p = .06$.

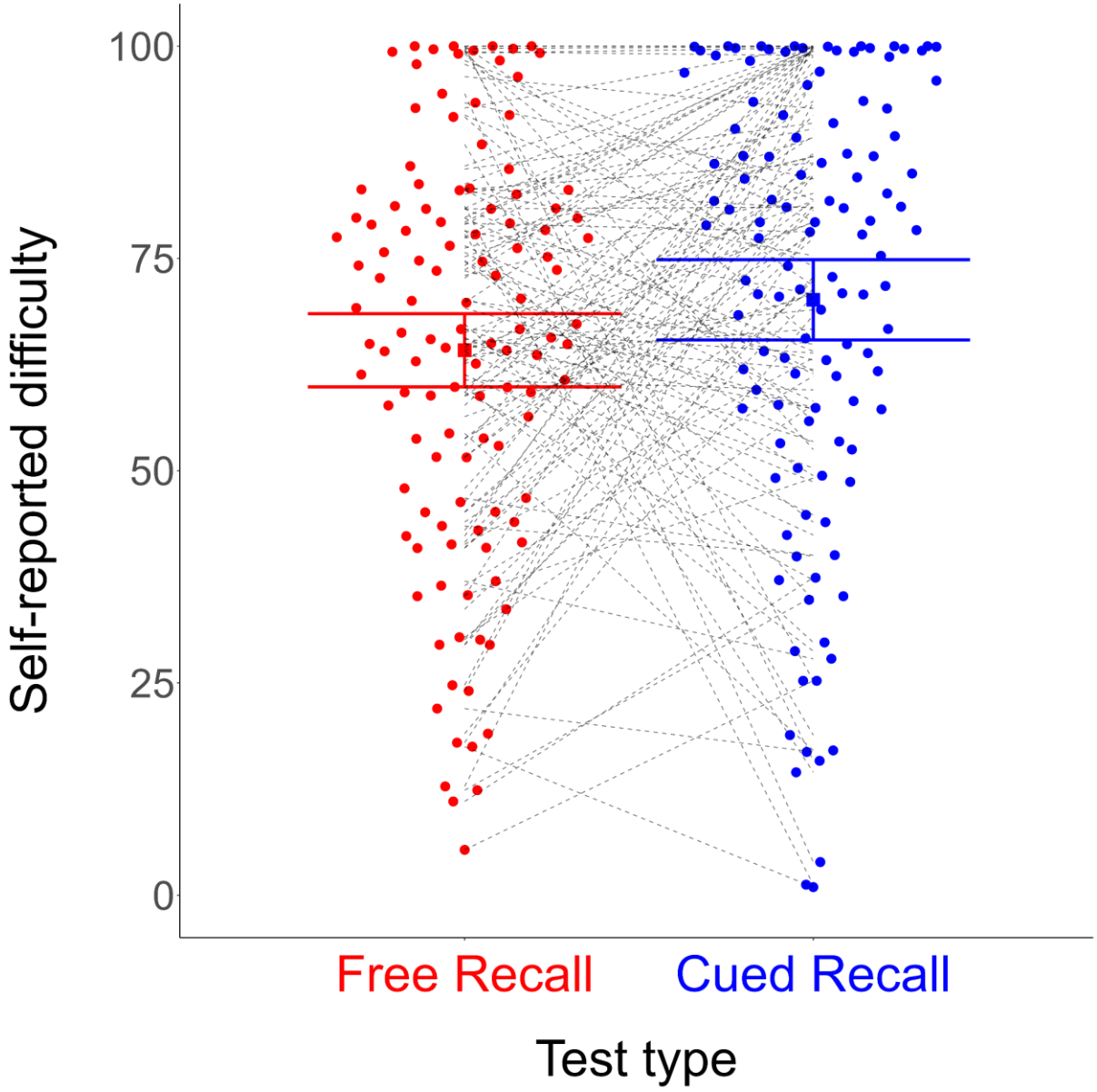
E. Self-reported study strategies. The re-introduction of the qualitative study strategy questions permitted analyses of potential differences in the variability (unalikeability) of strategies used in FR and CR. Two coders initially coded 300 reported responses, and agreed on 223. The remaining 77 responses were put to an independent third coder. The final strategy proportions reported in the figure below include strategies that for each participant were mentioned by at least two coders.



Note. Error bars = proportion 95% CIs (using Wilson's (1927) method).

Unalikeability was slightly higher for CR (.80, 95% percentile bootstrap CI [.77, .83) than FR (.71, 95% percentile bootstrap CI [.65, .76]). The CIs here overlap, but less so than in Experiment 1, and again the difference at least directionally favours CR.

F. Self-reported recall difficulty. Although we found no compelling evidence for variability differences in subjective impressions of FR and CR, we again analyzed self-reported recall difficulty:



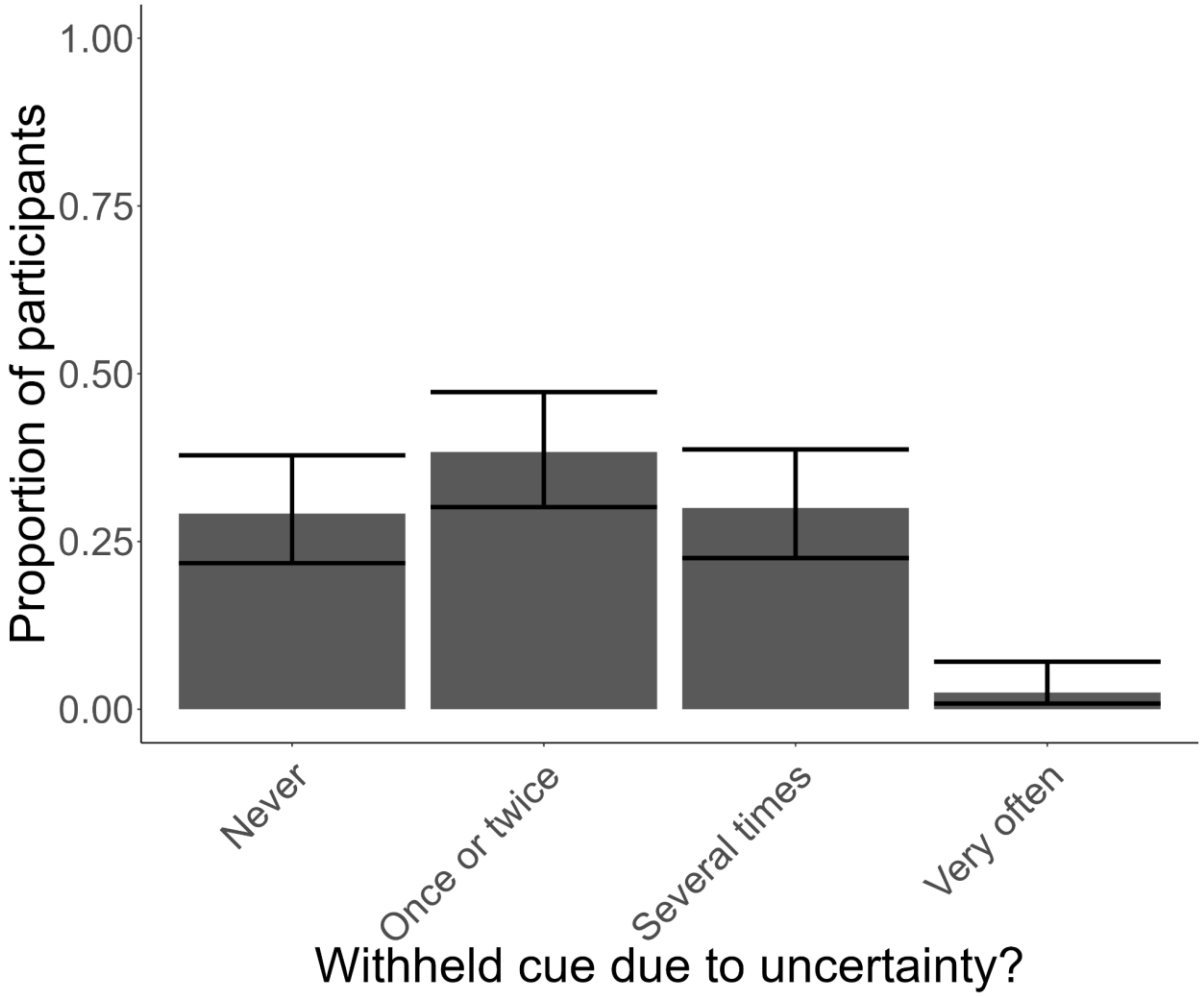
Note. Error bars = 95% CIs (within-subjects). Points jittered horizontally based on relative frequency. Dashed lines connect FR and CR difficulty ratings for individual participants.

FR and CR difficulty ratings did not significantly differ, Pitman-Morgan $t(117) = 1$, $p = .32$, bootstrapped CR:FR variance ratio = 1.10, 95% percentile bootstrap CI [.93, 1.29], with the Bayesian model comparison slightly favouring the *equal FR/CR variances* model over the *differing FR/CR variances* model, ΔLOO = .58 (*SE* = 1.10, 95% CI [-1.57, 2.72]), with the 95% CI on the difference containing 0.

G. Self-reported frequency of cued recall errors

a. Unsure of correct cue

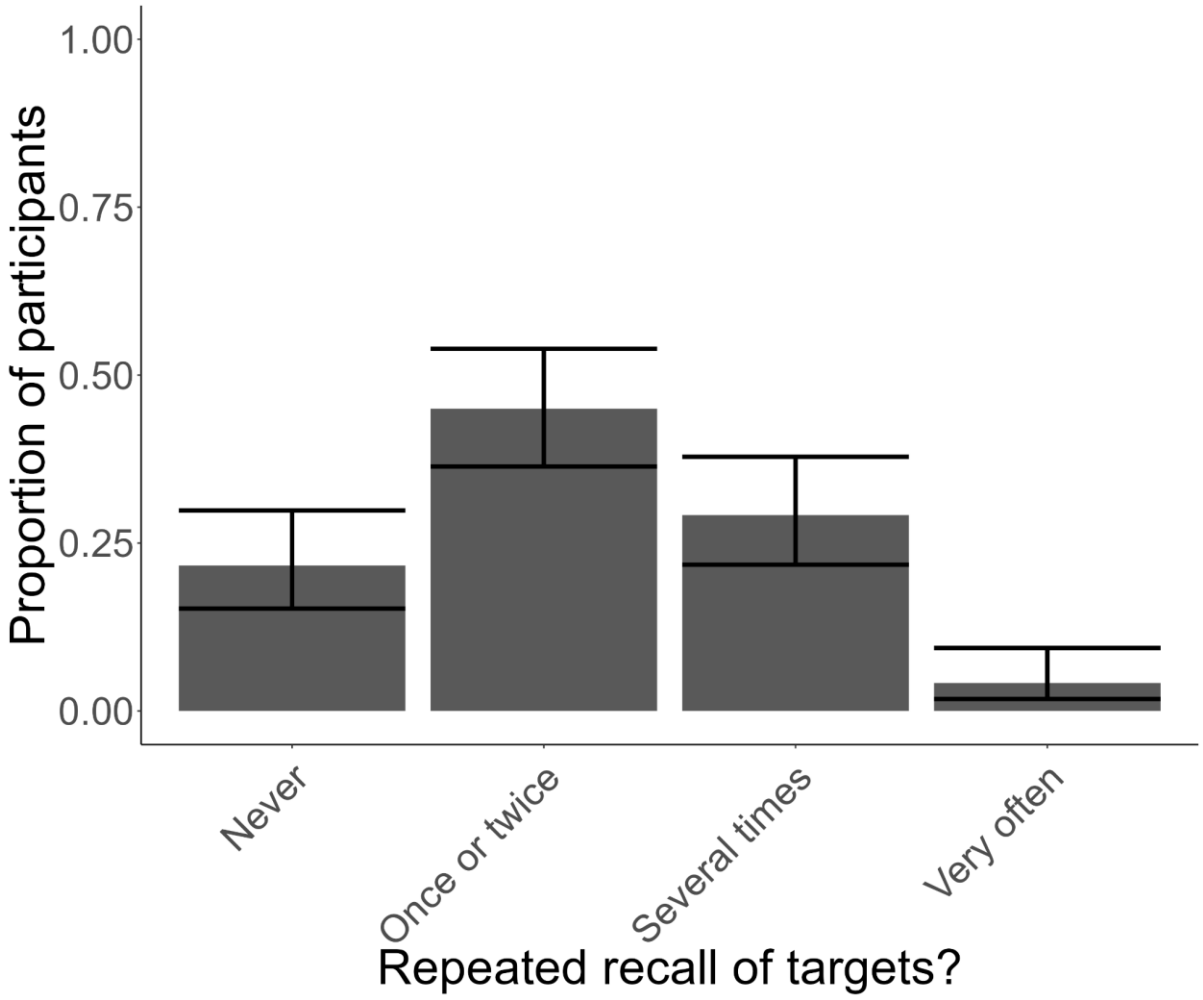
Participants were asked to self-report the general frequency with which they withheld a CR target that they thought of because they were unsure of whether it was paired with the current cue. Possible responses included: *Never*, *Once or twice*, *Several times*, and *Very often*. The figure below shows proportions of responses to this question:



Note. Error bars = 95% CIs on the proportions (Wilson method)

b. Repeated recall of targets

Participants were asked a similar question about the repeated recall of targets, i.e., whether they later recalled a target they had already given because they realized that the previous recall instance was to the incorrect cue. Proportions are shown below:

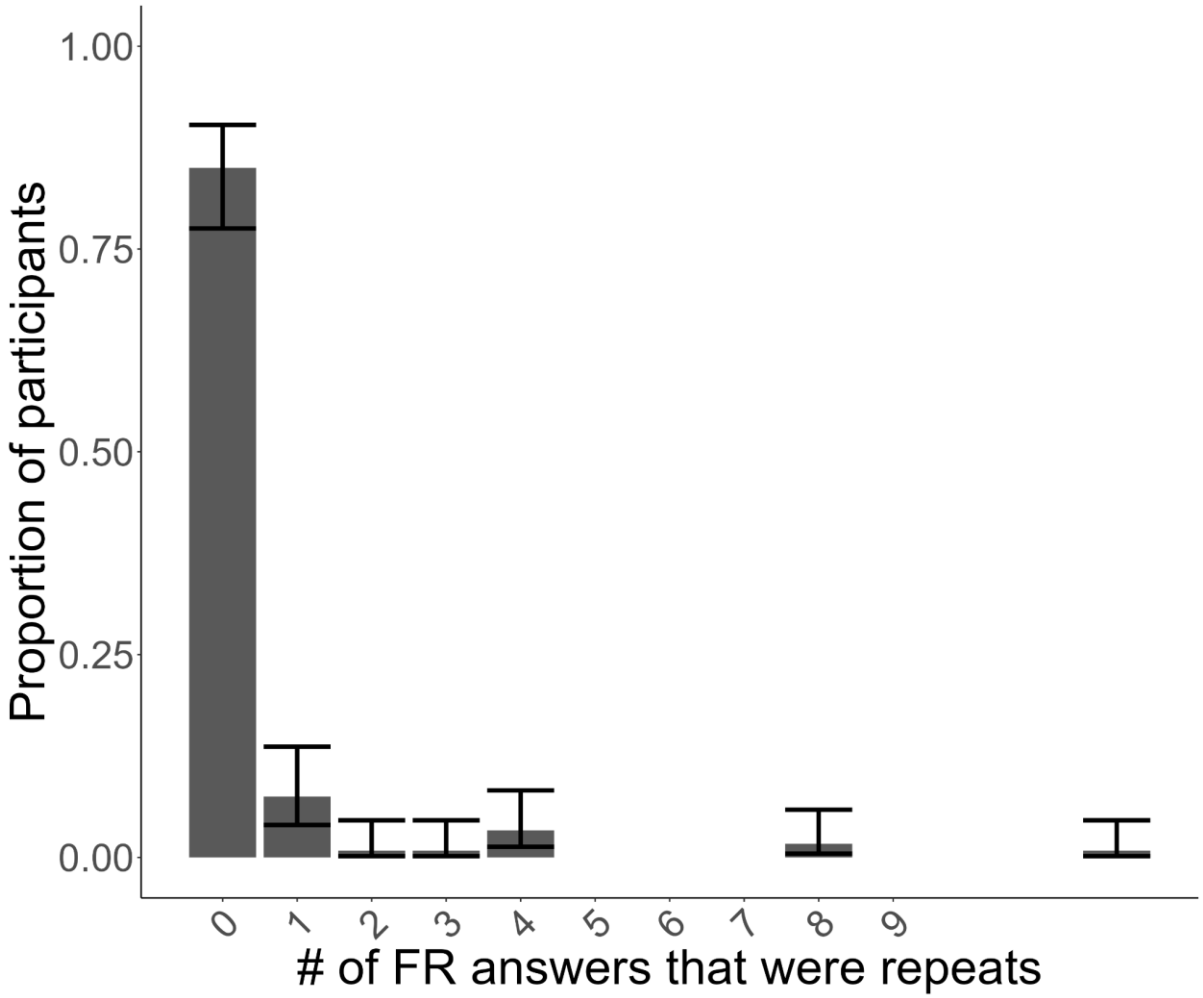


Note. Error bars = 95% CIs on the proportions (Wilson method)

H. Repeats in recall

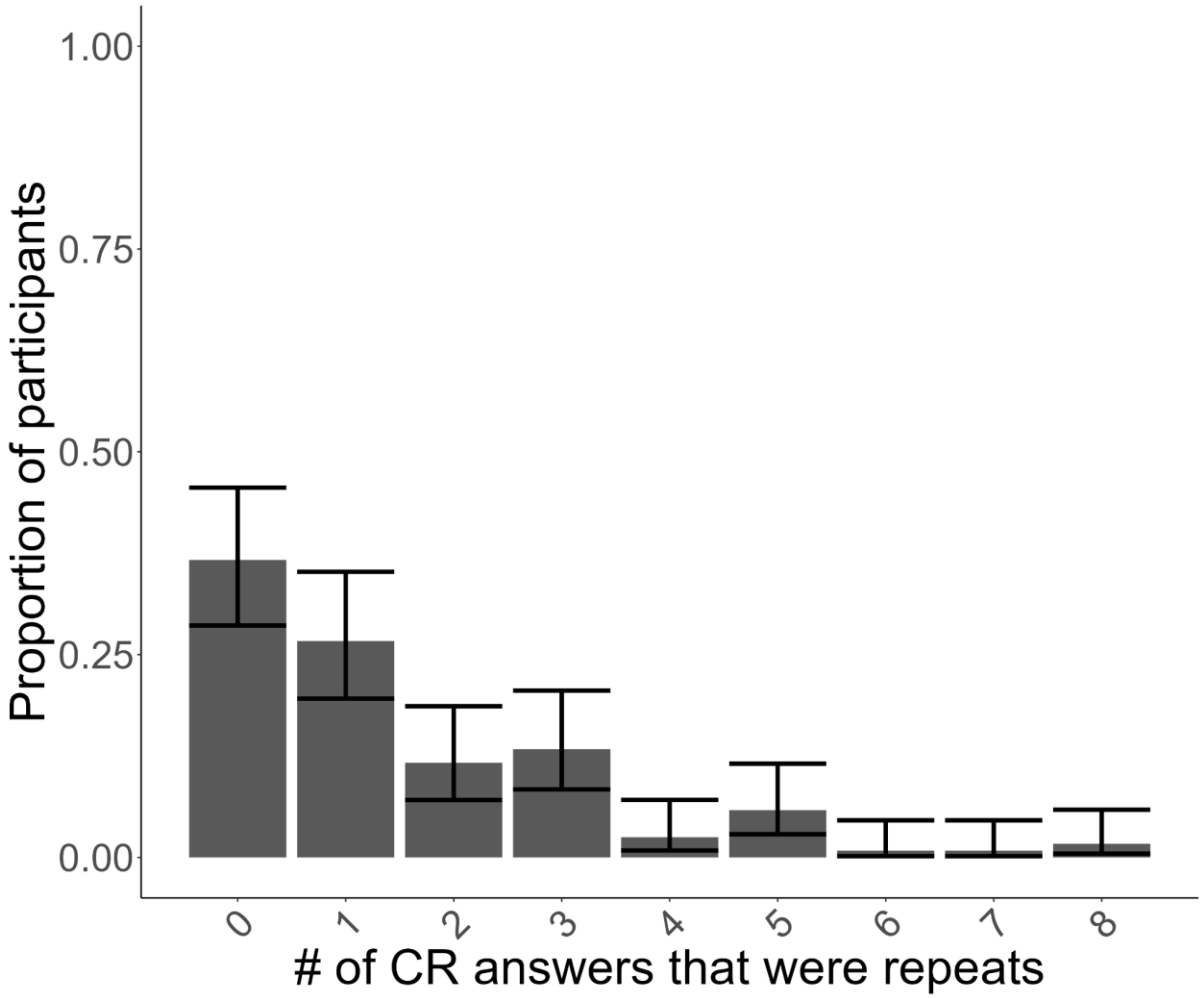
a. Free recall

We examined the number of free recall responses that were repeats. The vast majority of participants (85%, 95% CI [78%, 90%]) did not repeat any answers:



b. Cued recall

We examined the number of cued recall answers that were repeats. Though zero was the modal number of repeats (36.7%, 95% CI [28.6%, 45.6%]), most participants repeated one or more cued recall response:

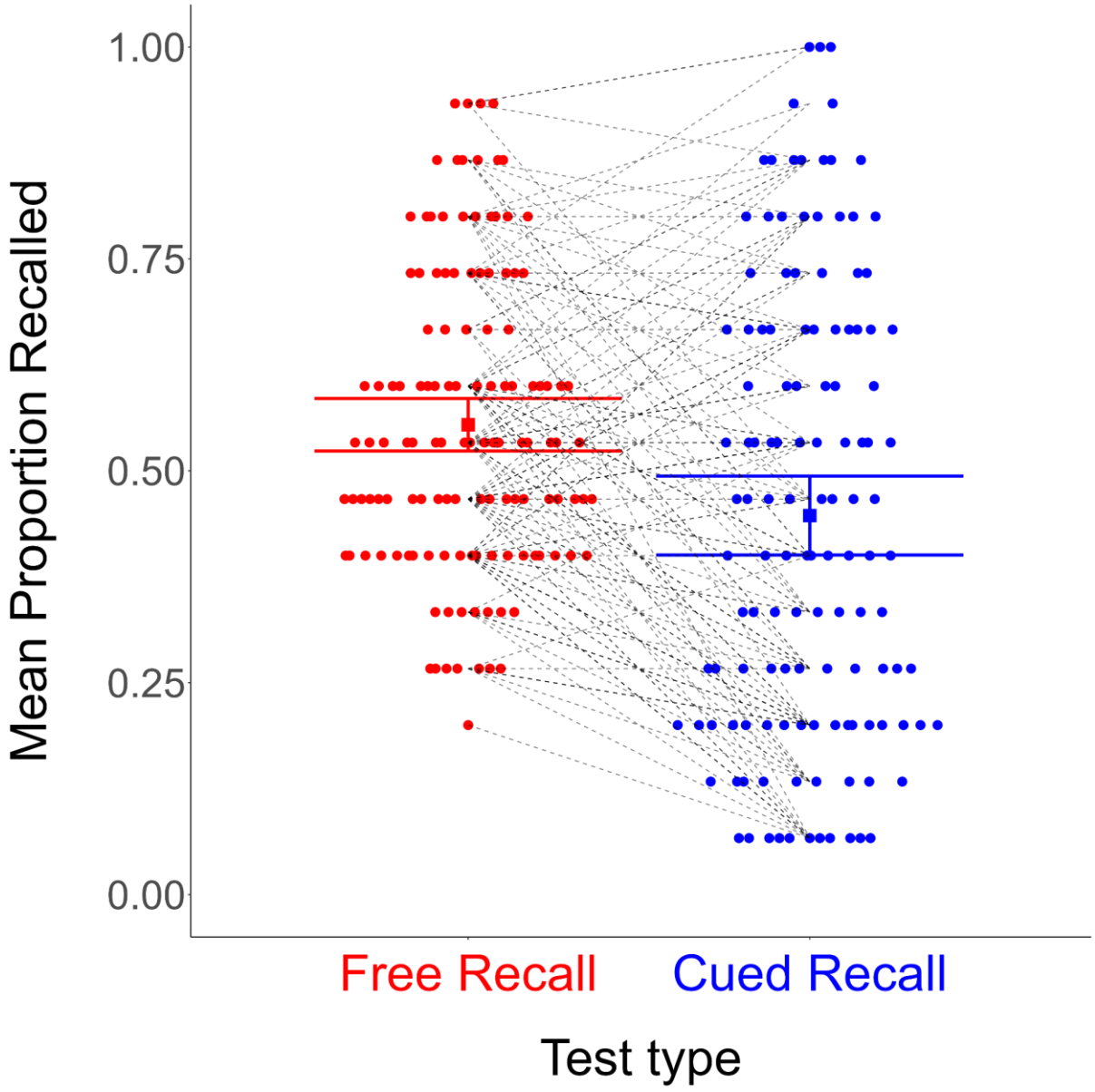


The majority of repeats (71.2%, 95% CI [64.8%, 77.7%]) were of studied targets.

5. Experiment 2b Supplementary Results

A. **Bayesian computational modelling analysis.** The Bayesian model comparison provided clear evidence favouring the model with *differing FR/CR variances* over the model with *equal FR/CR variances*, ΔLOO = 11.34 (*SE* = 3.40, 95% CI [4.68, 18.00]).
B. **Treating CR responses as correct as long as they were a target.** The results were nearly identical when conducting the analyses treating CR responses as correct as long as they came from the studied target list, perhaps due to the infrequency of

recalling a studied target in response to an studied but mismatched cue (56/1109 CR commission errors):



Note. Error bars = 95% CIs (within-subjects). Points jittered horizontally based on relative frequency. Dashed lines connect FR and CR performance for individual participants.

The variability difference was still significant, Pitman-Morgan $t(125) = 5.27$, $p < .001$, with a similar bootstrapped CR:FR variance ratio of 1.51 (95% percentile bootstrap CI [1.32, 1.72]), and similar Bayesian model comparison results clearly favouring the *differing FR/CR*

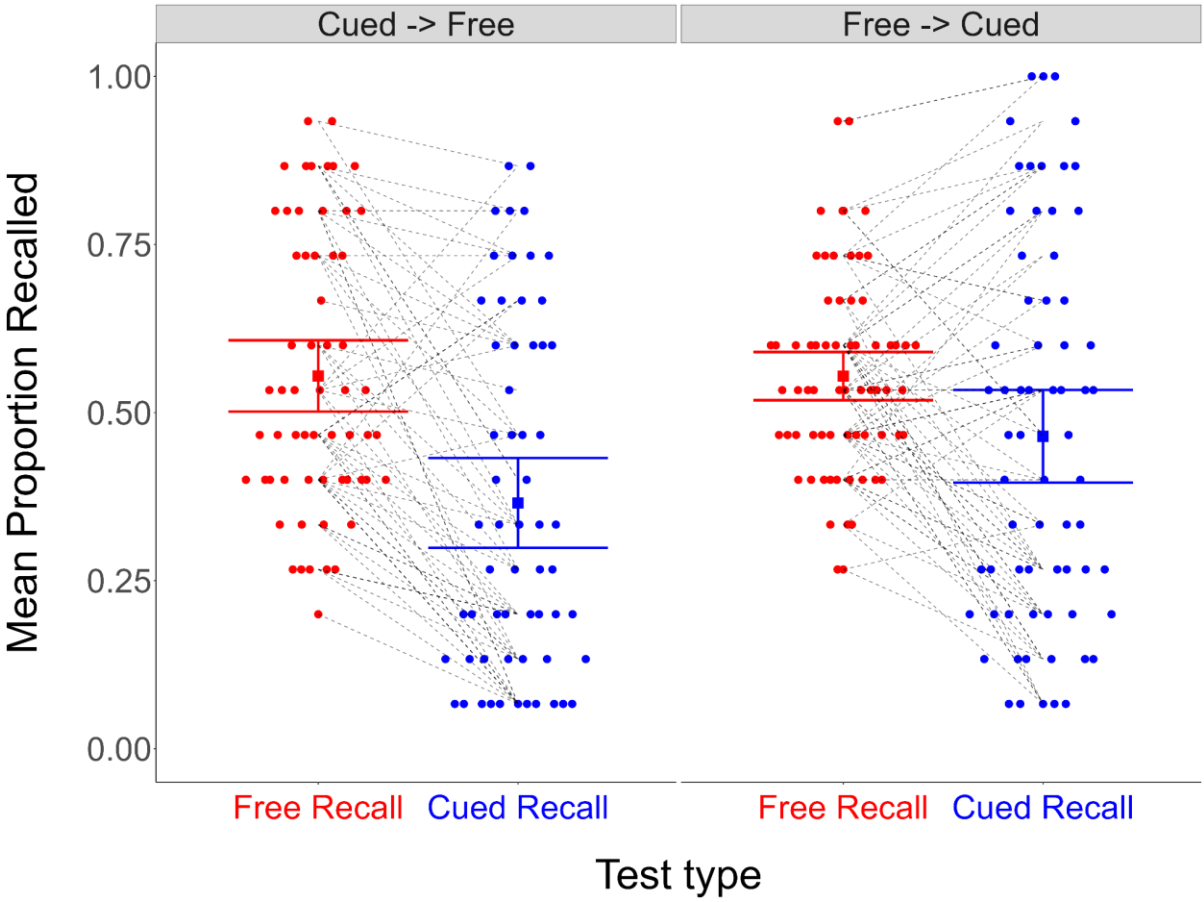
variances model over the *equal FR/CR variances* model, ΔLOO = 8.77 (*SE* = 3.07, 95% CI [2.74, 14.79]).

C. Generalized mixed-effects logistic regression results. We used the same GLMM model as in Experiment 1 to compare FR and CR variability (for standard CR proportion recalled and proportion recalled when same-list commission errors were counted as correct):

Test type	***SD* (Logit units)**	**95% CI lower**	**95% CI upper**
FR	.53	.39	.68
CR	1.29	1.09	1.53
CR (commission)	1.20	1.01	1.42

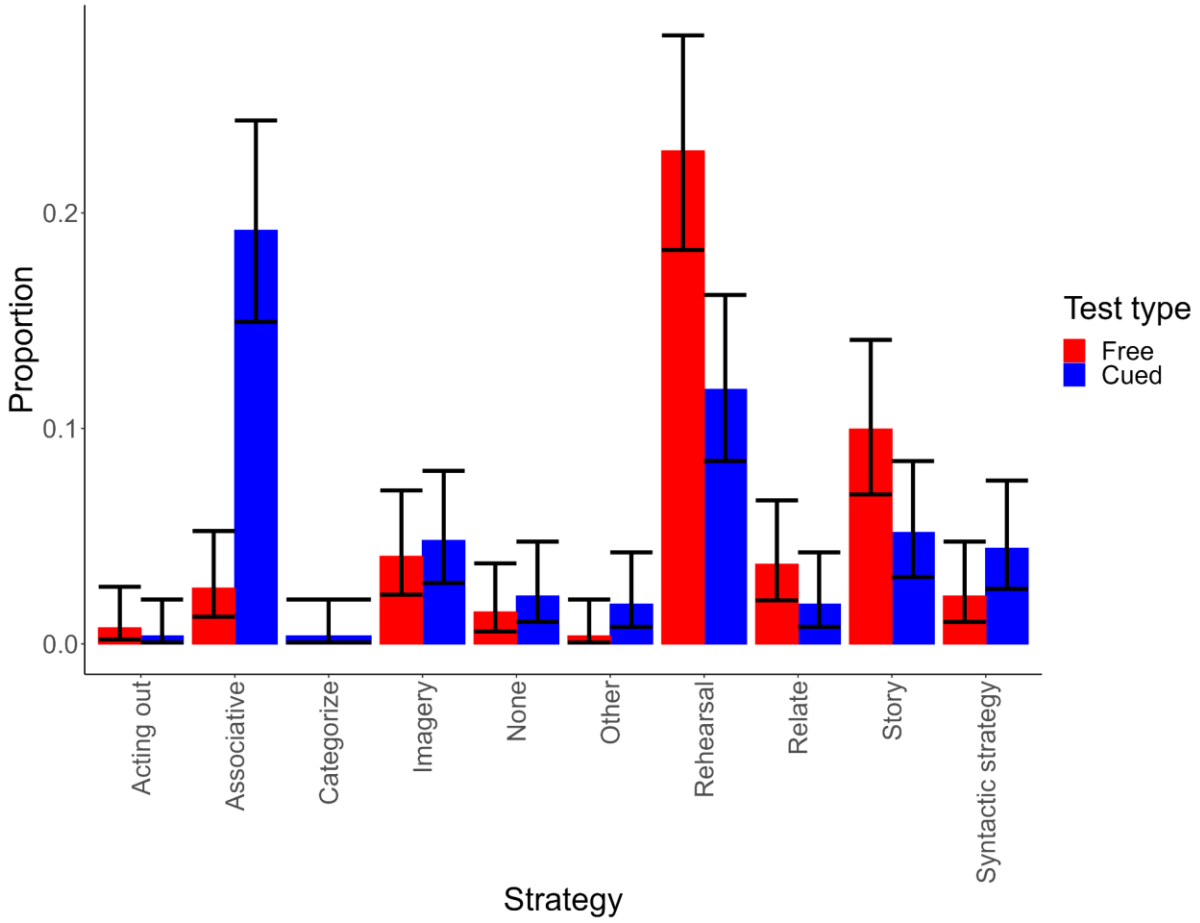
As the 95% CIs on the *SD* estimates did not overlap, this analysis provided evidence for differing FR/CR variances.

D. Order effects: 60 participants completed CR before FR, and 67 completed FR before CR. We examined the possible influence of task order on variability differences and overall accuracy. First, we conducted separate Pitman-Morgan tests in each order condition. The test was significant in the CR -> FR group, $t(58) = 2.01$, $p = .049$, and also in the FR -> CR group, $t(65) = 6.39$, $p < .001$. The bootstrapped CR:FR variance ratio in the CR -> FR group was 1.26 (95% percentile bootstrap CI [1.06, 1.50]), and in the FR -> CR group it was 1.94 (95% percentile bootstrap CI [1.57, 2.39]). Accuracy by test type and test order is shown in the figure below:



In an exploratory analysis of accuracy including both test type and test order factors, the interaction between test type and test order was significant, $\chi^2(1) = 6.83$, $p = .009$. That is, participants who did FR before CR had more similar performance on the tests than participants who did CR before FR.

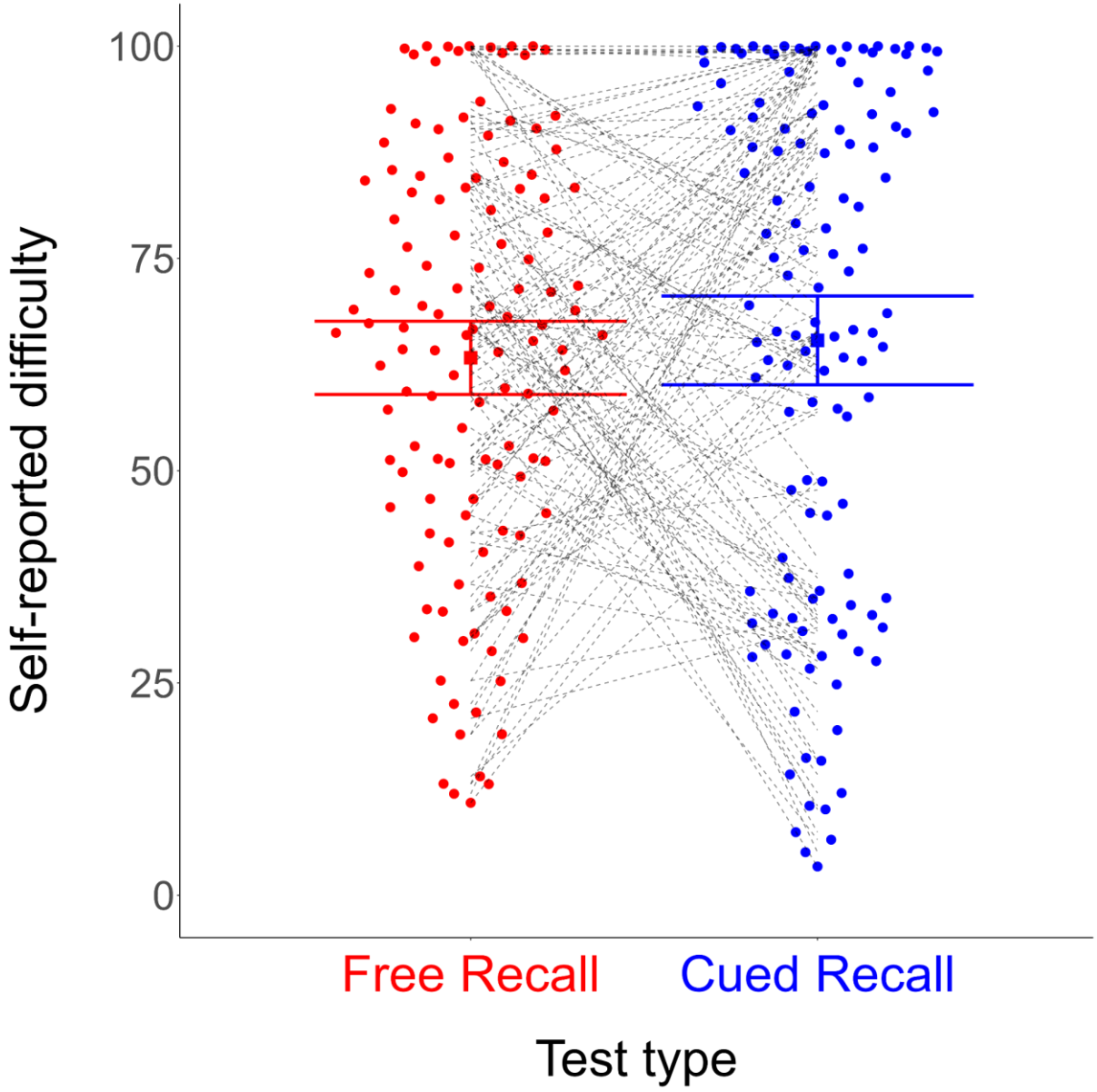
E. Self-reported study strategies. Of 414 coded qualitative strategy responses, the initial two coders agreed on 327. The remaining 87 responses were put to a 3rd coder.



Note. Error bars = proportion 95% CIs (using Wilson's (1927) method).

Unalikeability results were similar in the student sample: slightly higher for CR (.78, 95% percentile bootstrap CI [.73 .82) than FR (.71, 95% percentile bootstrap CI [.64, .77]). The CIs here overlap, but less so than in Experiment 1, and again the difference at least directionally favours CR.

F. Self-reported recall difficulty. Self-reported recall difficulty. The results for difficulty were slightly different in our undergraduate sample:



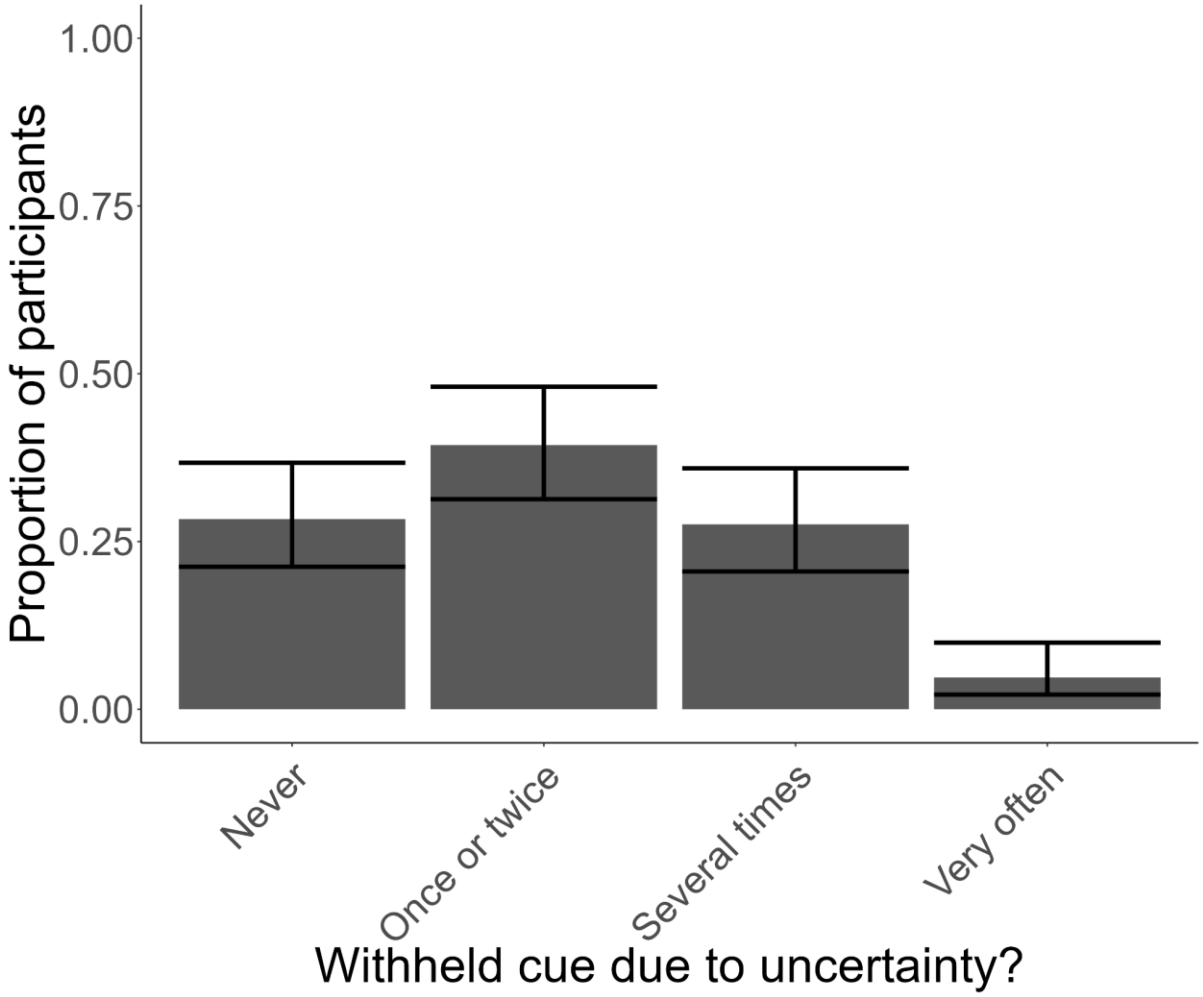
Note. Error bars = 95% CIs (within-subjects). Points jittered horizontally based on relative frequency. Dashed lines connect FR and CR difficulty ratings for individual participants.

Here, the variability difference was significant, Pitman-Morgan $t(124) = 2.14$, $p = .03$, with a bootstrapped CR:FR variance ratio of 1.21 (95% percentile bootstrap CI [1.07, 1.38]). However, the Bayesian model comparison did not produce clear results; the model with *differing FR/CR variances* was only slightly favoured over the *equal FR/CR variances* model, ΔLOO = 2.17 (*SE* = 1.87, 95% CI [-1.51, 5.84]), with the 95% CI containing 0.

G. Self-reported frequency of cued recall answers

a. Unsure of correct cue

Participants were asked to self-report the general frequency with which they withheld a CR target that they thought of because they were unsure of whether it was paired with the current cue. Possible responses included: *Never*, *Once or twice*, *Several times*, and *Very often*. The figure below shows proportions of responses to this question:

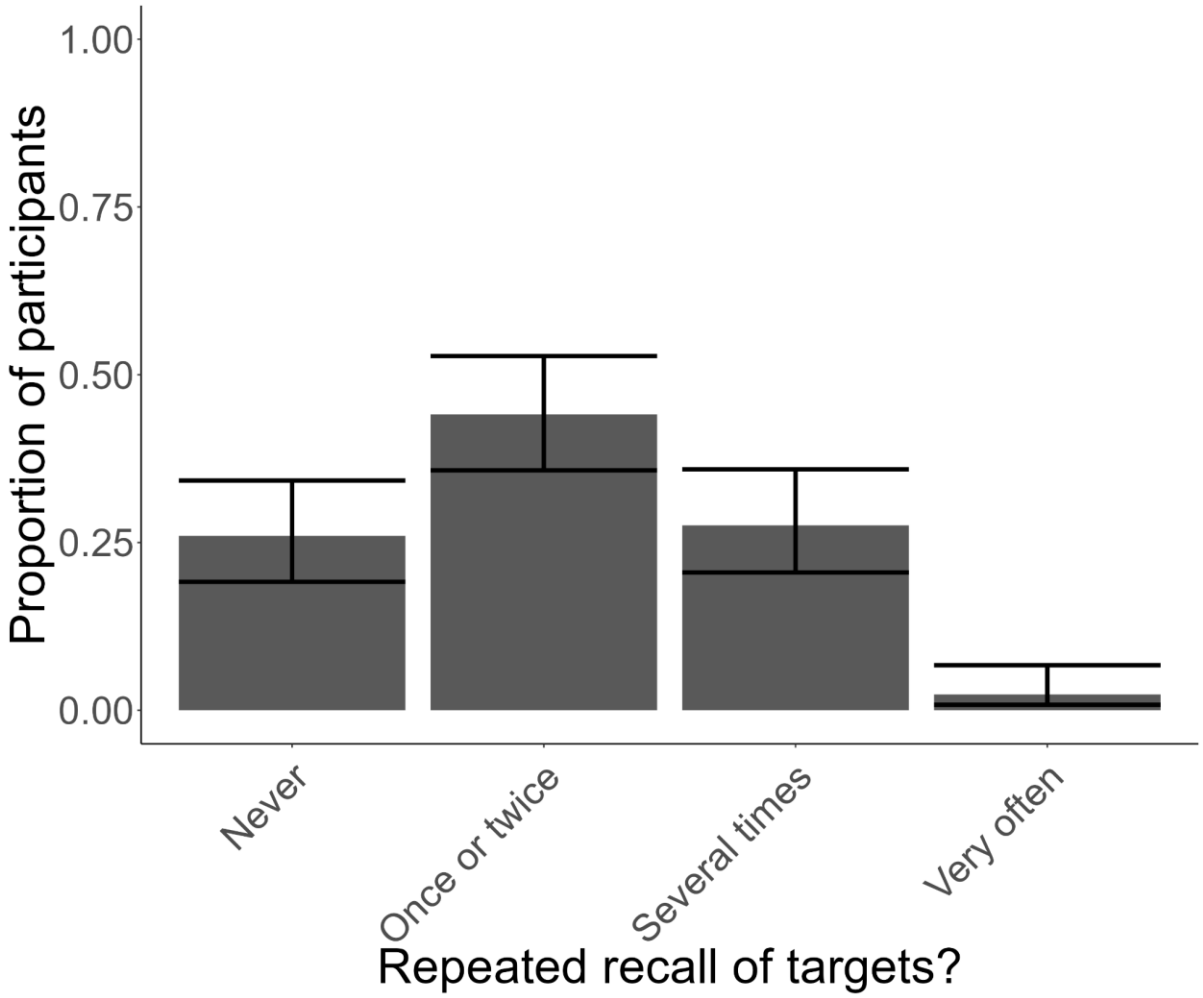


Note. Error bars = 95% CIs on the proportions (Wilson method)

b. Repeated recall of targets

Participants were asked a similar question about the repeated recall of targets, i.e., whether they later recalled a target they had already given because they realized that the previous

recall instance was to the incorrect cue. Proportions are shown below:

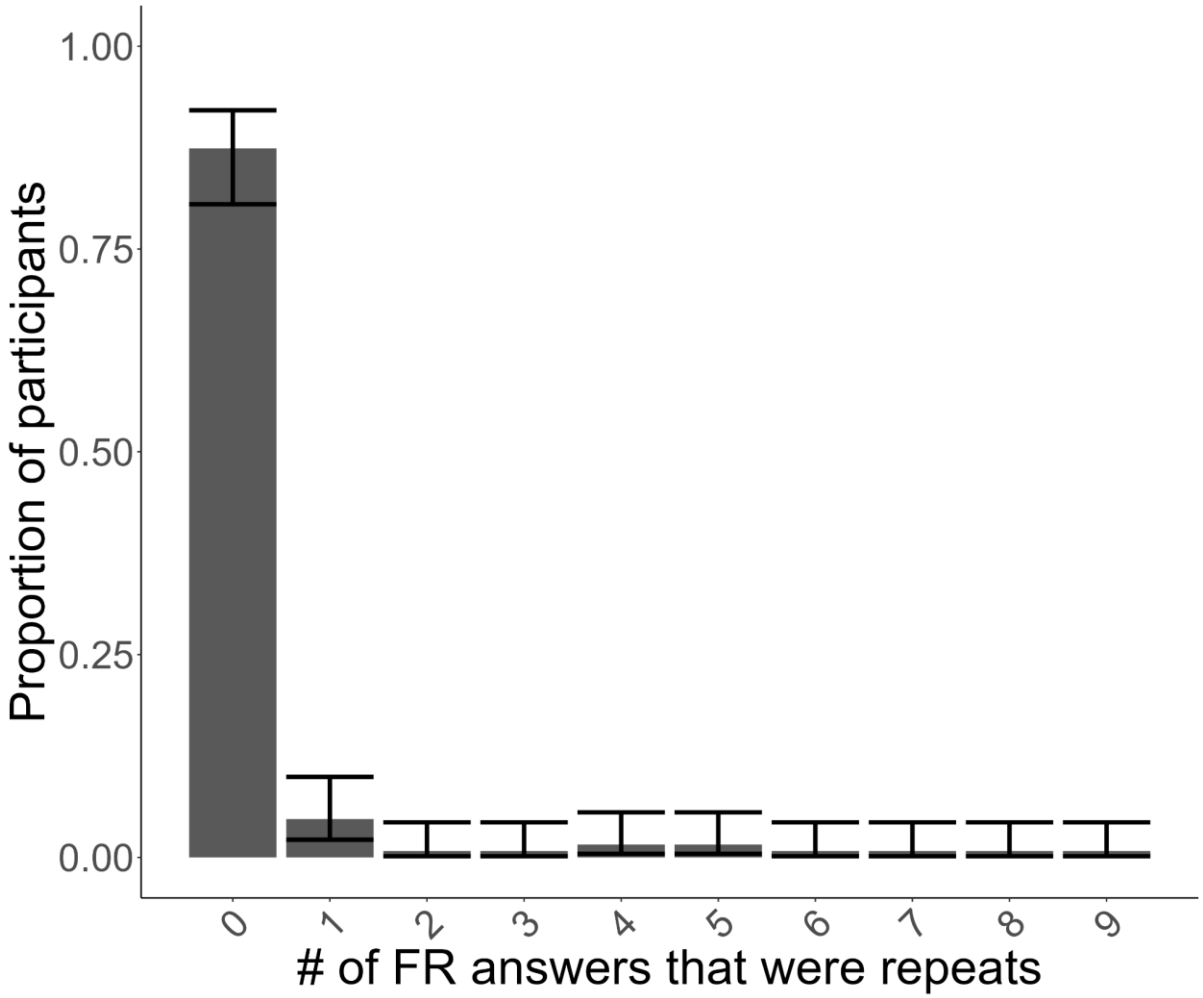


Note. Error bars = 95% CIs on the proportions (Wilson method)

H. Repeats in recall

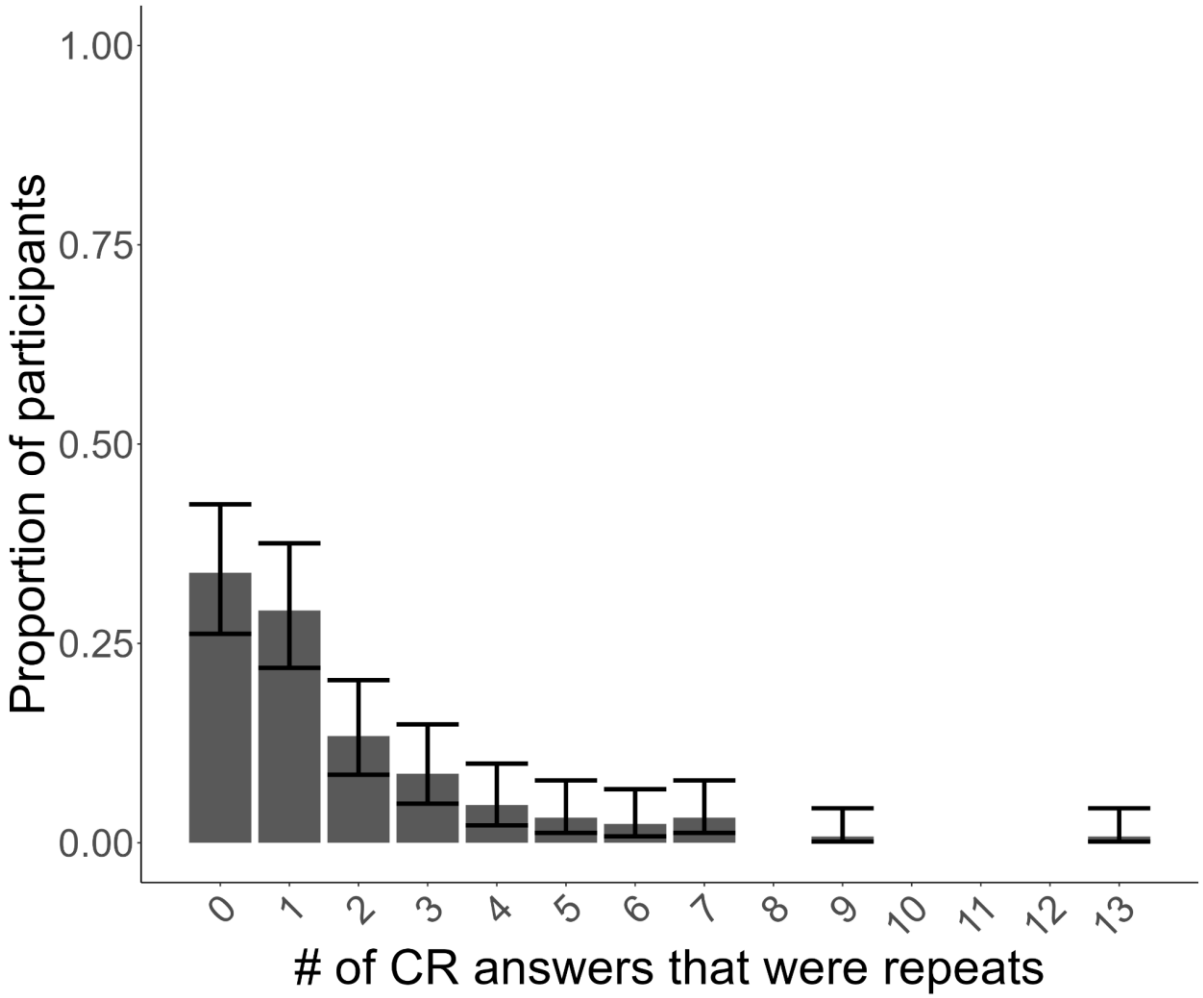
a. Free recall

We examined the number of free recall responses that were repeats. The vast majority of participants (87.4%, 95% CI [80.5%, 92.1%]) did not repeat any answers:



b. Cued recall

We examined the number of cued recall answers that were repeats. Though zero was the modal number of repeats (33.9%, 95% CI [26.2%, 42.5%]), most participants repeated one or more cued recall response:



Of the repeats, about half were of studied targets (57.4%, 95% CI [50.7%, 63.8%]).

6. Unreported Experiment summary and results: "Surprise free recall"

A. Summary, materials, procedure, and results

The tendency for participants to vary more on CR than FR could be due to greater individual differences in cognitive processes during the study phase, test phase, or both. The commission error proportion results in Experiment 1 hinted at a greater variability in CR recall strategies than in FR recall strategies.

We conducted Experiment 2 to investigate this possibility. In a within-subjects design, participants completed standard FR and CR study phases but were tested on FR

regardless of study instructions. That is, regardless of the study instructions, and regardless of whether they had studied individual words or word pairs in that block, they were instructed to recall as many studied targets as they could, in any order. Our objective here was to approximately equate the conditions at test for FR and CR. If the CR:FR variability effect persisted under these conditions, then that would suggest that the effect is due at least in part to differential variability in FR and CR processes at study. If the CR:FR variability effect disappeared on the "surprise free recall" test, then it is likely that the effect is due to differential variability in FR and CR processes at test. We did not have an explicit hypothesis favouring one of these possibilities over the other, but preregistered our design, materials, and analyses (viewable at https://osf.io/3tra5/?view_only=65b1552b17144c1ca6c401d5d325ec18, under a registration titled "Performance variability in free recall and paired-associates learning: Encoding vs. Test").

Methods

Materials

We made several changes to the materials for Experiment 2. First, we re-examined and reduced the set of 120 nouns used in Experiment 1, excluding any words with salient non-noun meanings and any words we thought participants might not be familiar with (e.g., HIND). Word exclusions were based on the subjective ratings of three research team members,[20] The reduced wordset contained 83 words. The reduced wordset and experiment program (now made in PsychoPy & run via Pavlovia) can be found at https://osf.io/z47r3/?view_only=39b351e7e98a4c7c80fe619a9556c12B.

Procedure

In Experiment 2, participants completed one FR and one CR study-test cycle (order

[20] Specifically, if two out of three raters considered a word to have a salient non-noun meaning or to be too obscure, that word was removed from the pool.

counterbalanced), each consisting of 15 words/word-pairs. As in Experiment 1, words/word-pairs were presented for 5s each at study, with standard FR/CR study instructions. Our crucial manipulation was to the CR test phase, when participants were given the "surprise free recall" test. Specifically, participants were told:

"On the next page, you will be tested on the word list that you just studied. Although we told you that we would present the first word of the pairs that you studied and ask you to recall the second word, we will simply ask you to recall as many of the second words of the pairs as you can, in any order, until you cannot remember any more. So, if you studied 'guitar - spoon' and 'lion - fish', you would only need to freely recall 'spoon' and 'fish', in any order (you wouldn't need to recall 'guitar' or 'lion'). You will type as many of the words as you can into the computer, one at a time, until you cannot remember any more. Each word you enter will be displayed on the screen after you enter it. You will type one word at a time and press the ENTER key to enter it. Remember that the order of the words you recall does not matter for them to be counted correct; simply try to recall as many as you can."

After completing both study-test cycles, participants completed the same questions as in Experiment 1, with the only differences being the addition of a cheating question ("Did you take notes?") and the removal of the qualitative strategy questions. This experiment was conducted as a combined experiment in collaboration with Bottesini et al. (2021), who added to the end of our experiment a meta-science experiment. This combined experiment was run online via Amazon mTurk.

Sample

Based on our power analysis for Experiment 1, we preregistered the same target sample size (N = 120), and collected data until we reached this N after exclusions, in this case

a total sample of 195 mTurk participants who each received $5 USD for participating. Participation was restricted to mTurk participants age 18+ who self-reported English fluency and had an mTurk HIT approval rate > 90% and at least 10 approved HITs. From our sample of 195, we excluded 73 participants on the basis of preregistered exclusion criteria: 12 participants who indicated experiencing a major distraction during the study, 22 participants who reported understanding less than 75% of the studied words, 62 participants who did not get at least one correct on both lists, 15 participants who reported cheating, and 13 participants who reported a major technical difficulty (note that many participants were excluded on multiple criteria). Our final sample included 122 participants ages 21-66 (M = 38.16, SD = 10.89).

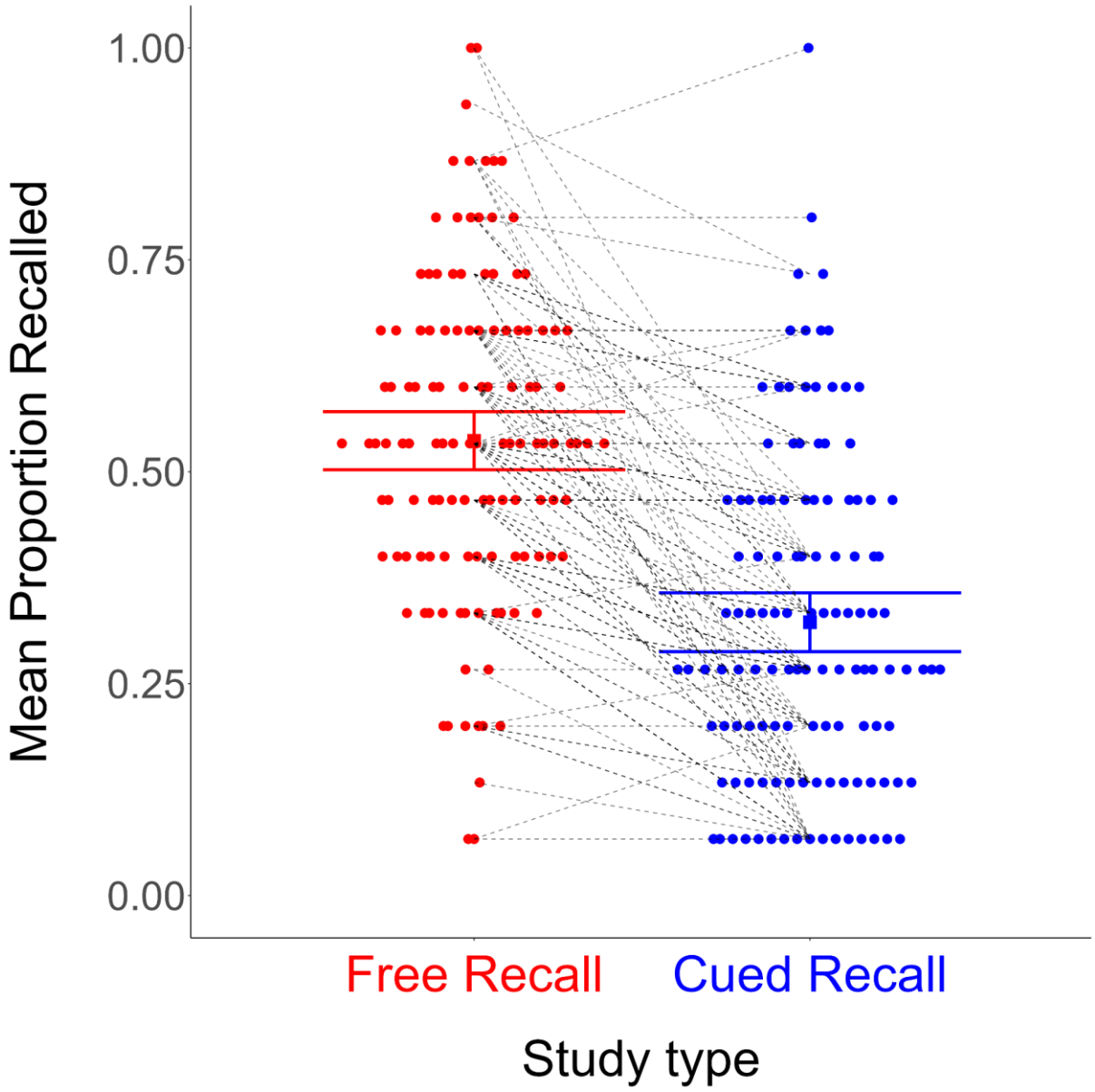
As with Experiment 1, we also manually checked and coded participant commission errors on FR and CR. In total, 106 FR errors (out of 1,241 total FR responses) and 446 CR errors (out of 1,100 total CR responses) were manually checked by two independent coders. Of these errors, the coders disagreed on 51 FR errors (45 accepted corrections) and 28 CR errors (35 corrections accepted). All disagreements were resolved by the 2nd coder.

Results

Confirmatory analyses

Our primary analyses were the same as in Experiment 1 (data files and analysis scripts available at https://osf.io/z47r3/?view_only=39b351e7e98a4c7c80fe619a9556c12B). The figure below depicts the means, within-subjects 95% CIs, and distributions of FR and CR (surprise free recall) performance in our sample.

Experiment 2: Memory performance as a function of study type



Note. Error bars = 95% CIs (within-subjects). Points jittered horizontally based on relative frequency. Dashed lines connect FR and CR performance for individual participants.

A paired Pitman-Morgan test of unequal variances was not significant, $t(120) = .19$, $p = .85$, with an estimated ratio of CR:FR variance (via bootstrap) of 1.02 (95% percentile bootstrap CI [.85, 1.22])[21]. The generalized mixed-effects logistic regressions also failed to provide

[21] Results were similar when looking at accuracy separately by test order (i.e., for those who did CR first vs. second), see Supplementary Material X.

evidence for differing CR/FR variances (see Supplementary Material 6Bb.). As in Experiment 1, we also conducted an exploratory analysis of variability in self-reported recall difficulty. Other than much higher difficulty ratings for the "surprise free recall" test, we did not find any evidence for differences in variability (See Supplementary Material 6Bd.).

These results suggest that the variability difference in Experiment 1 had more to do with differences in the variability of processes at test (e.g., recall strategies) than at study (e.g., encoding strategies). However, performance on the surprise FR test following CR study was low, with a number of responses at (post-exclusion) floor. It is possible that CR variability in this case was constrained by a potential floor effect (although most CR proportions were above floor).

Experiment 2 provided some evidence that the CR:FR variability difference has more to do with processes occurring at test. Our analyses of commission proportions in Experiment 1 (Figure 4) showed that the proportion of errors that were commissions was more variable for CR than for FR, which suggests that participants may vary more in their propensity to guess on CR (where many participants guessed incorrectly and some left answers blank) than on FR (where most participant errors were omissions).

B. Unreported Experiment Supplementary Results

a. **Bayesian computational modelling analysis.** The corresponding Bayesian computational modelling analysis via PSIS-LOO slightly favoured the model with *equal FR/CR variances* over the model with *differing FR/CR variances*, ΔLOO = .22 (*SE* = 1.14, 95% CI [-2.02, 2.44]), but the 95% CI on the difference contained 0. However, as with the Experiment 1 subjective difficulty ratings, in this case the lack of support for the *differing variances* model may reasonably be interpreted as support for the equal-variances model on the grounds of parsimony.

b. Generalized mixed-effects logistic regression results. Due to issues computing confidence intervals on random-effects variance estimates in a full GLMM, we instead estimated variances in intercept-only models for FR and CR response data separately, i.e.:

Level 1: $\text{logit}(y_{ti}) = \beta_{0i} + e_{ti}$

Level 2: $\beta_{0i} = \gamma_{00} + \mu_{0i}$

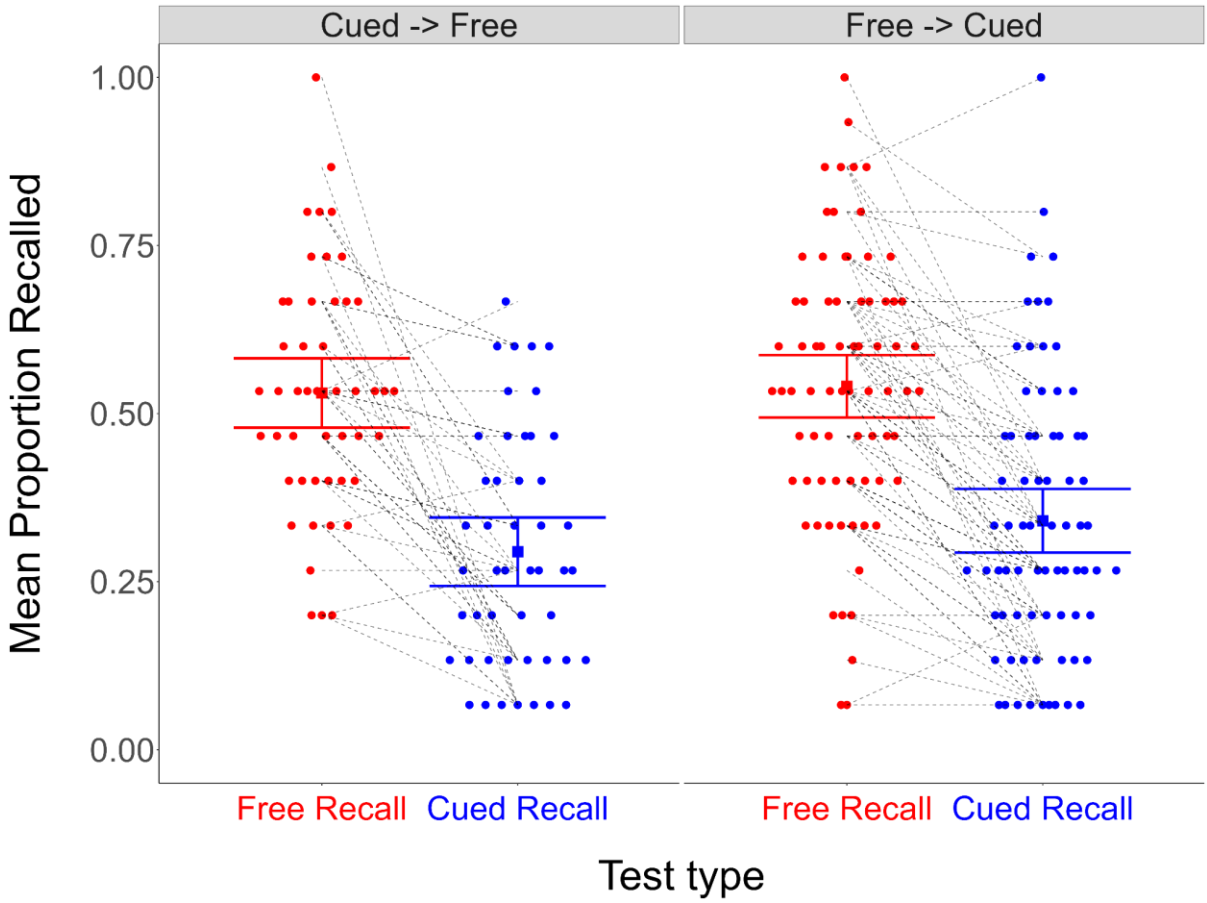
These models resulted in the following estimates:

Test type	***SD* (Logit units)**	**95% CI lower**	**95% CI upper**
FR	.63	.49	.80
CR	.77	.61	.95

As the 95% CIs on the *SD* estimates overlapped, this analysis did not provide evidence for differing FR/CR variances.

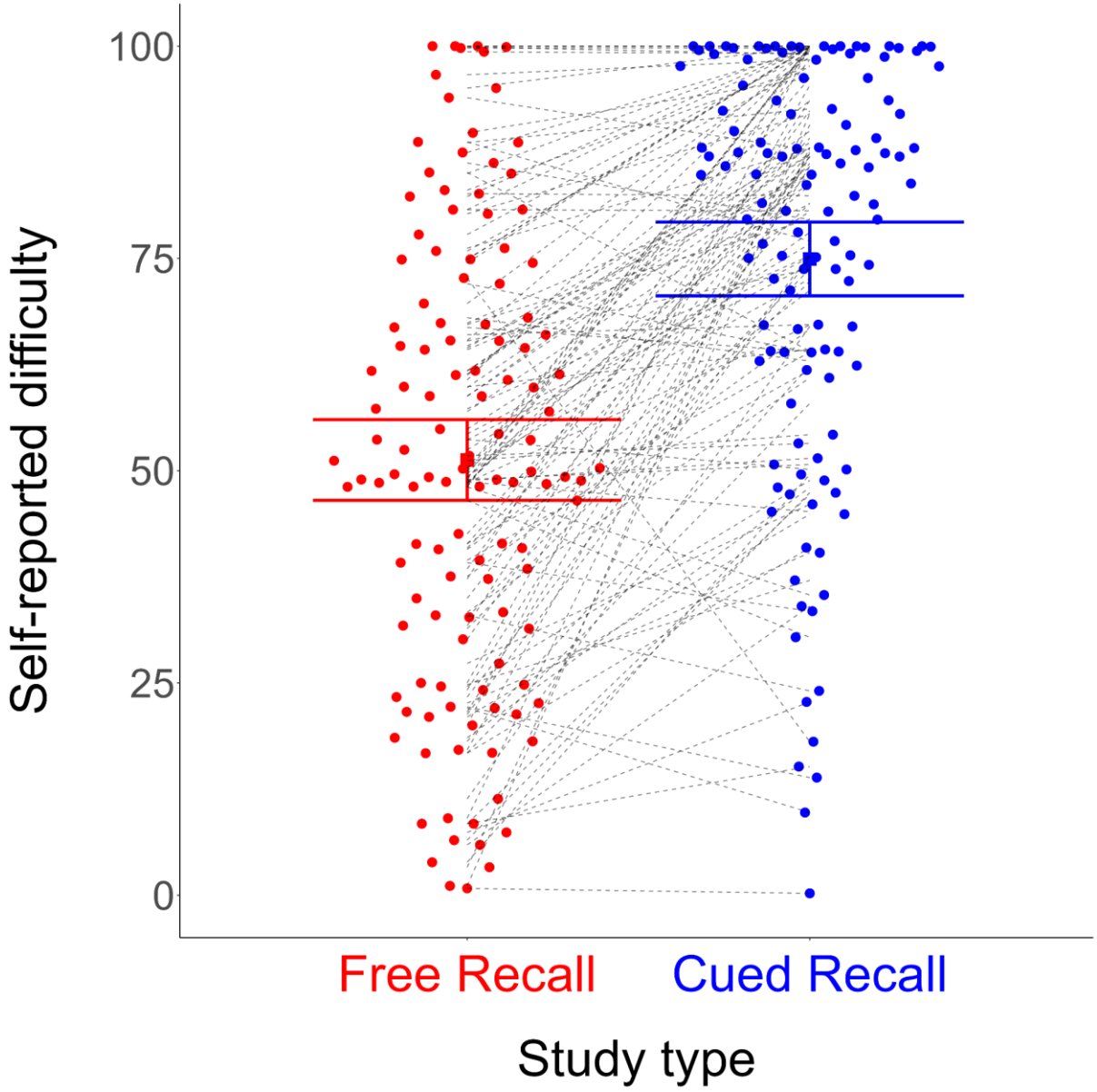
c. Order effects. 48 participants completed CR before FR, and 74 completed FR before CR. We examined the possible influence of task order on variability differences and overall accuracy. First, we conducted separate Pitman-Morgan tests in each order condition. The test was non-significant in the CR -> FR group, $t(46) = .09$, $p = .93$, and also in the FR -> CR group, $t(72) = .21$, $p = .84$. The bootstrapped CR:FR variance ratio in the CR -> FR group was 1 (95% percentile bootstrap CI [.77, 1.28]), and in the FR -> CR group it was 1.03 (95% percentile

bootstrap CI [.81, 1.29]). Accuracy by test type and test order is shown in the figure below:



In an exploratory analysis of accuracy including both test type and test order factors, the interaction between test type and test order was not significant, $\chi^2(1) = .88$, $p = .35$.

d. Self-reported recall difficulty. As in Experiment 1, we examined variability in self-reported recall difficulty. Subjective difficulty ratings are shown in the figure below:



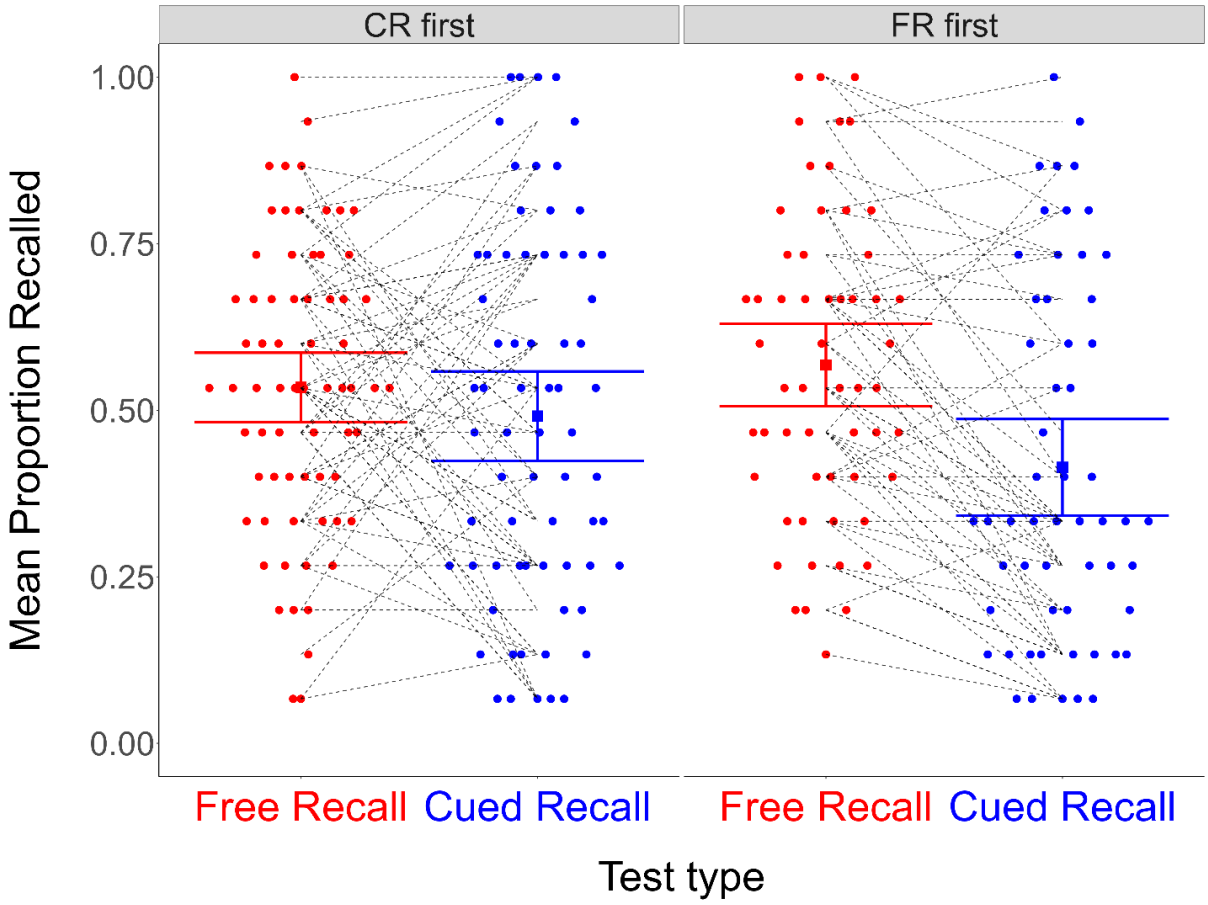
Note. Error bars = 95% CIs (within-subjects). Points jittered horizontally based on relative frequency. Dashed lines connect FR and CR difficulty ratings for individual participants.

Our analyses of variability yielded non-significant results, Pitman-Morgan $t(120) = 1.17$, $p = .24$, with an bootstrapped CR:FR variance ratio of .91 (95% percentile bootstrap CI [.77, 1.07]) and a Bayesian model comparison slightly favouring the model with *equal FR/CR variances* over the model with *differing FR/CR variances*, ΔLOO = .68 (*SE* = 1.03, 95% CI [-1.34, 2.70]), although the 95% CI on the difference contained 0. The lack of variability

differences is less interesting than the striking difference in difficulty ratings, with participants rating the surprise FR test after CR study as much more difficult than the FR test after FR study. This mirrors the behavioural results and suggests that participants may have been thrown off by our manipulation.

7. Experiment 5 Supplementary Results

A. Accuracy by task order



a. ***CR first***

Pitman-Morgan: $t(66) = 2.16$, $p = .03$

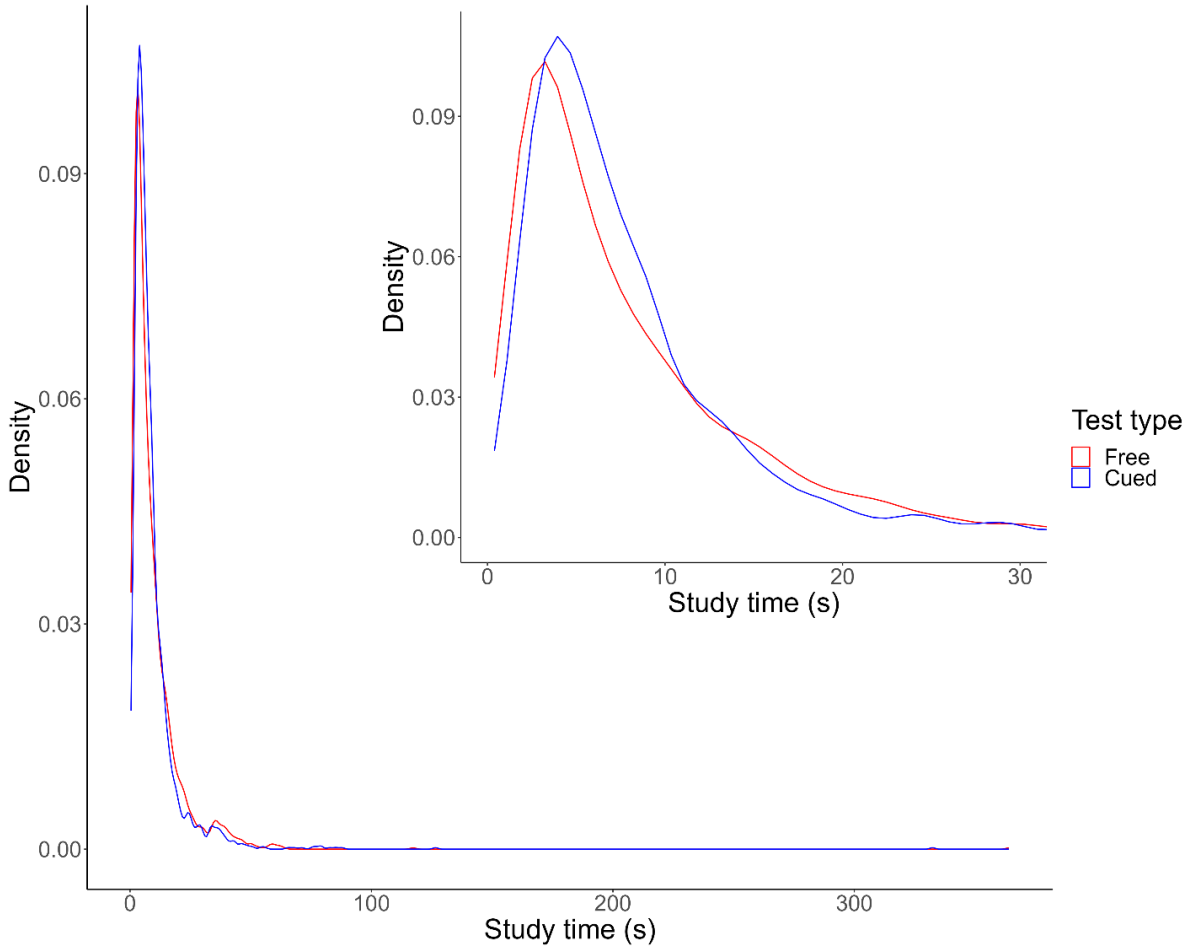
Bootstrapped CR:FR variance ratio = 1.29 [95% CI: 1.07, 1.56]

b. ***FR first***

Pitman-Morgan: $t(54) = 1.55$, $p = .13$

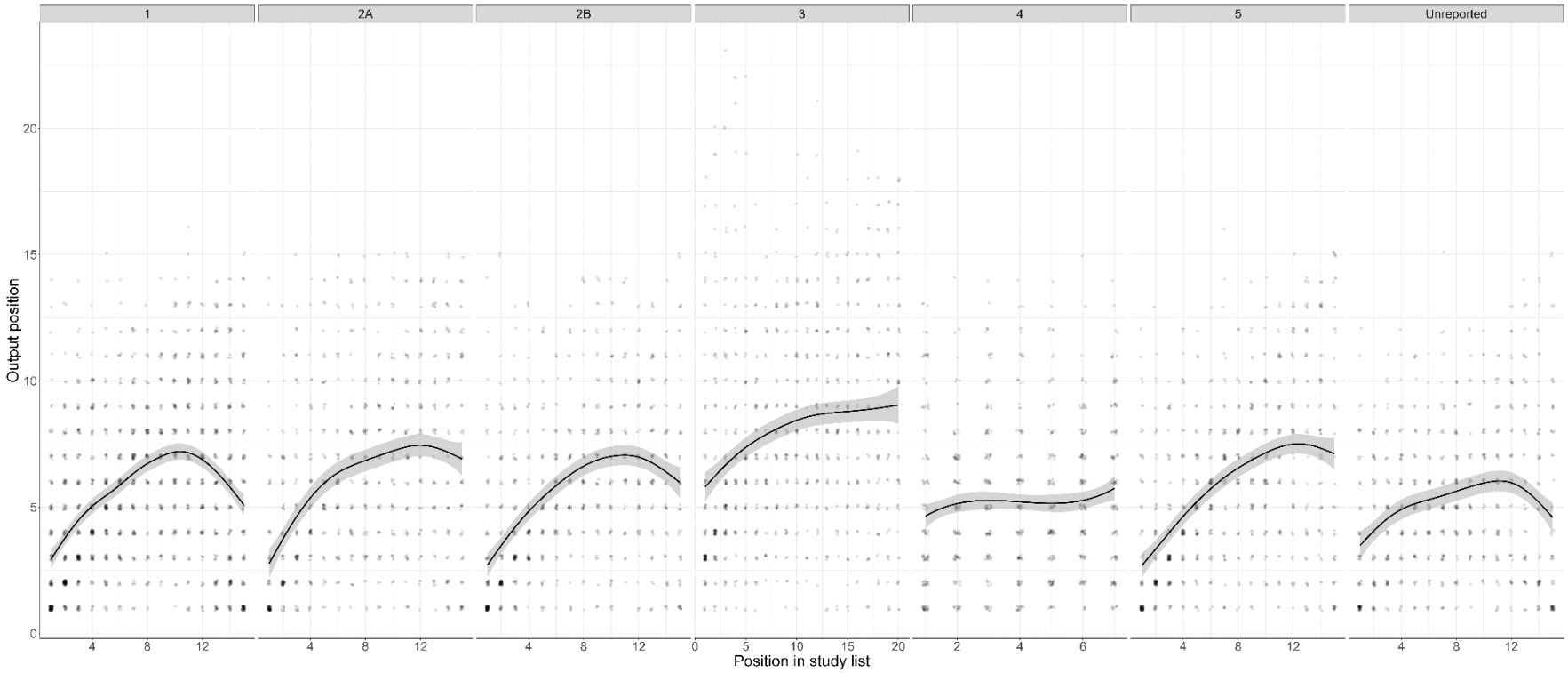
Bootstrapped CR:FR variance ratio = 1.18 [95% CI: .96, 1.43]

B. Study time



8. Output order

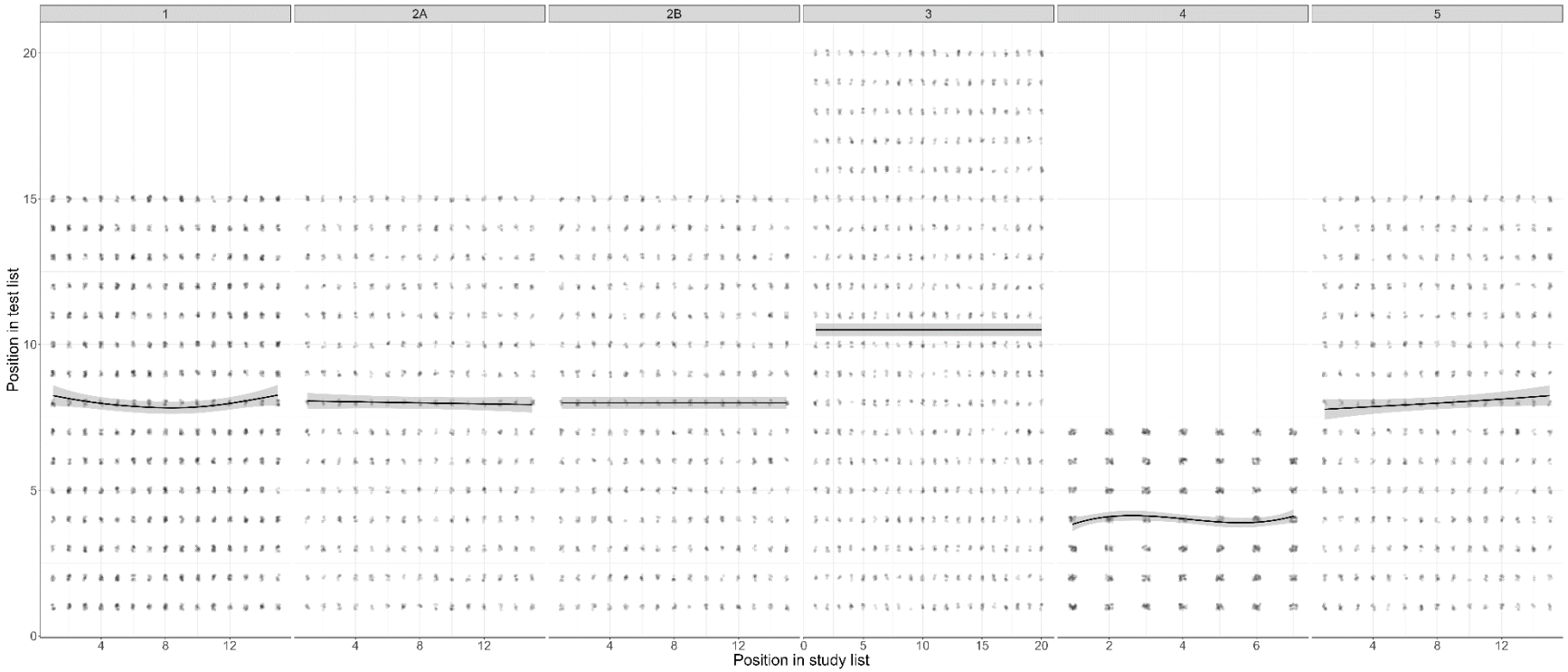
A. Free recall



The figure above depicts (correct) output position for free recall as a function of position in study list. Jittered points represent individual words, and lines and ribbons represent regression lines (local polynomial regression fitting/LOESS) with 95% confidence intervals. The data here show

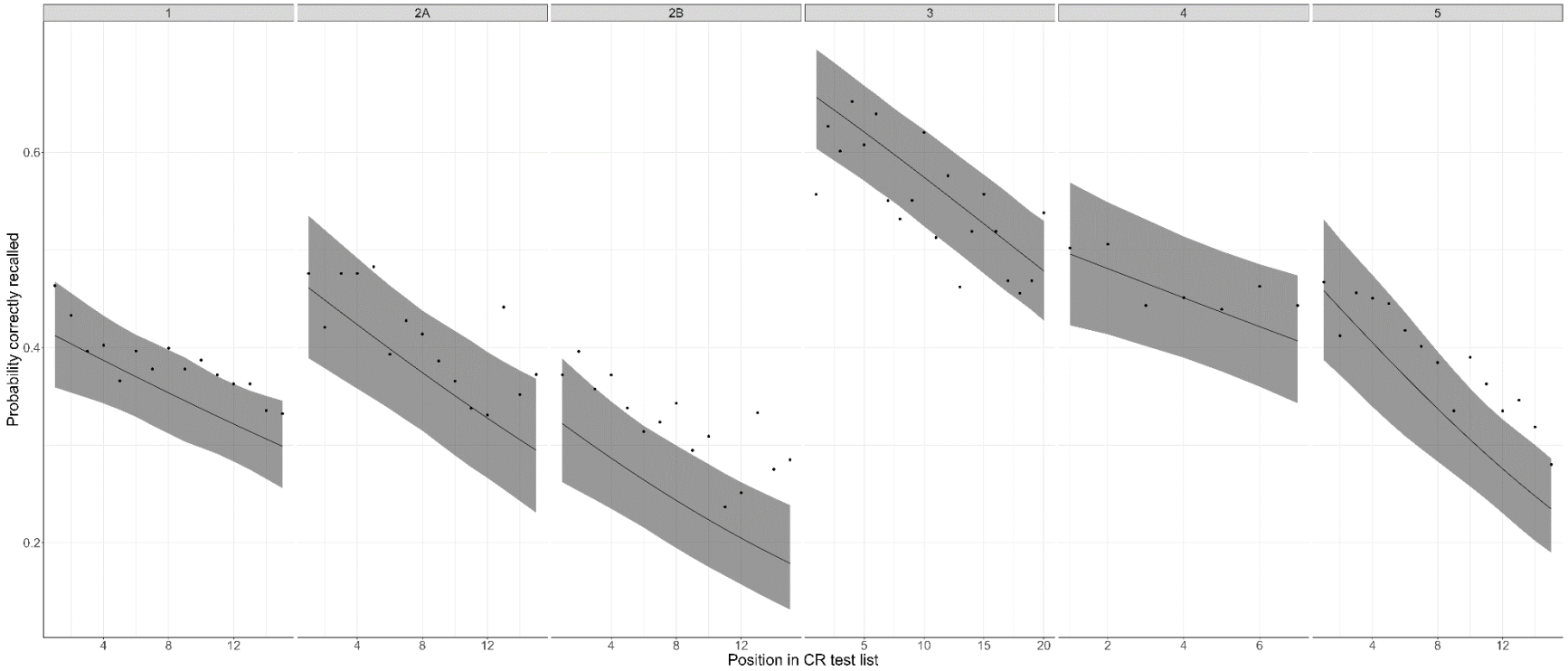
a pattern consistent with prior findings in the literature – i.e., a tendency to recall words serially, with the exception of recalling recently studied words earlier in the list.

B. Cued recall



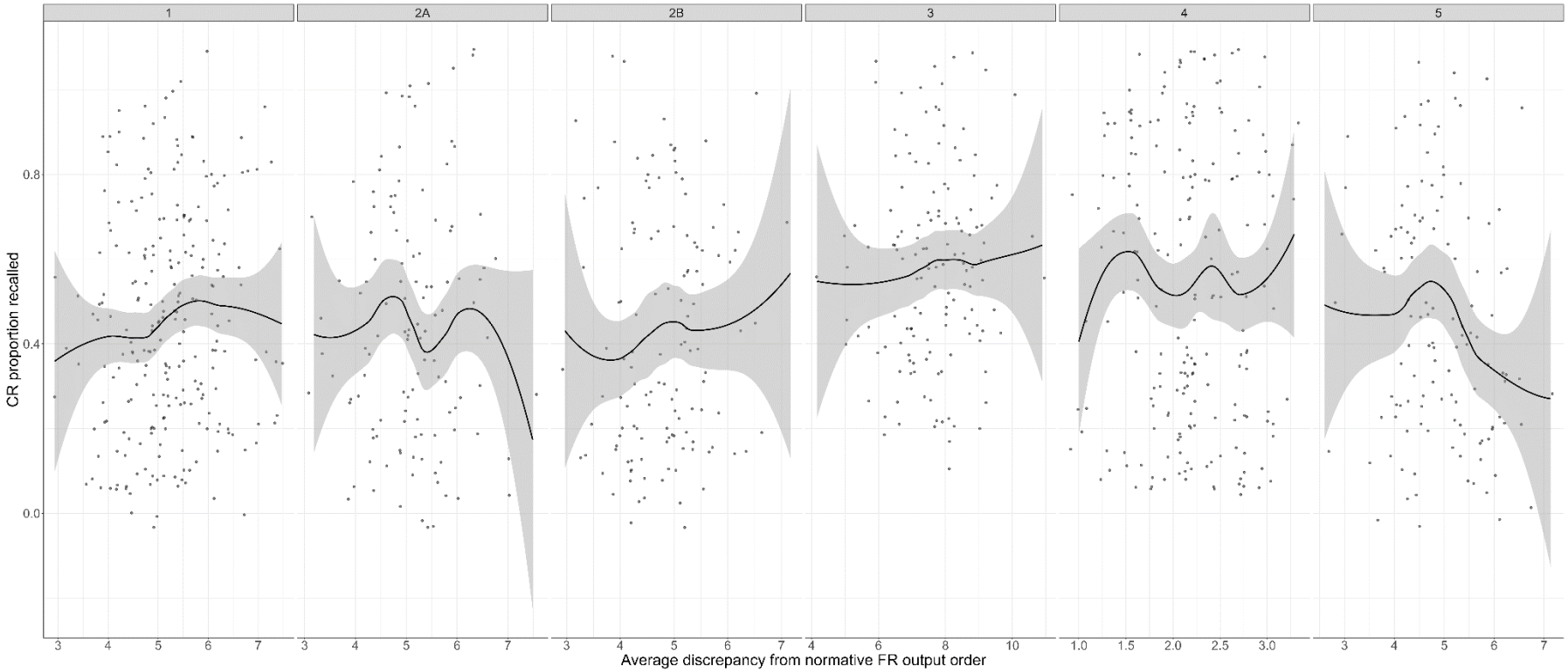
Similar figure for CR showing that, as expected, there was no relationship between position in the study list and likelihood of recall at a particular test position.

a. CR output interference



This figure depicts probability correct (regression lines and 95% CIs estimated via GLMM, points representing averaged accuracy at each position) as a function of position in the CR test list, and clearly shows output interference (i.e., declining performance over the course of the CR test).

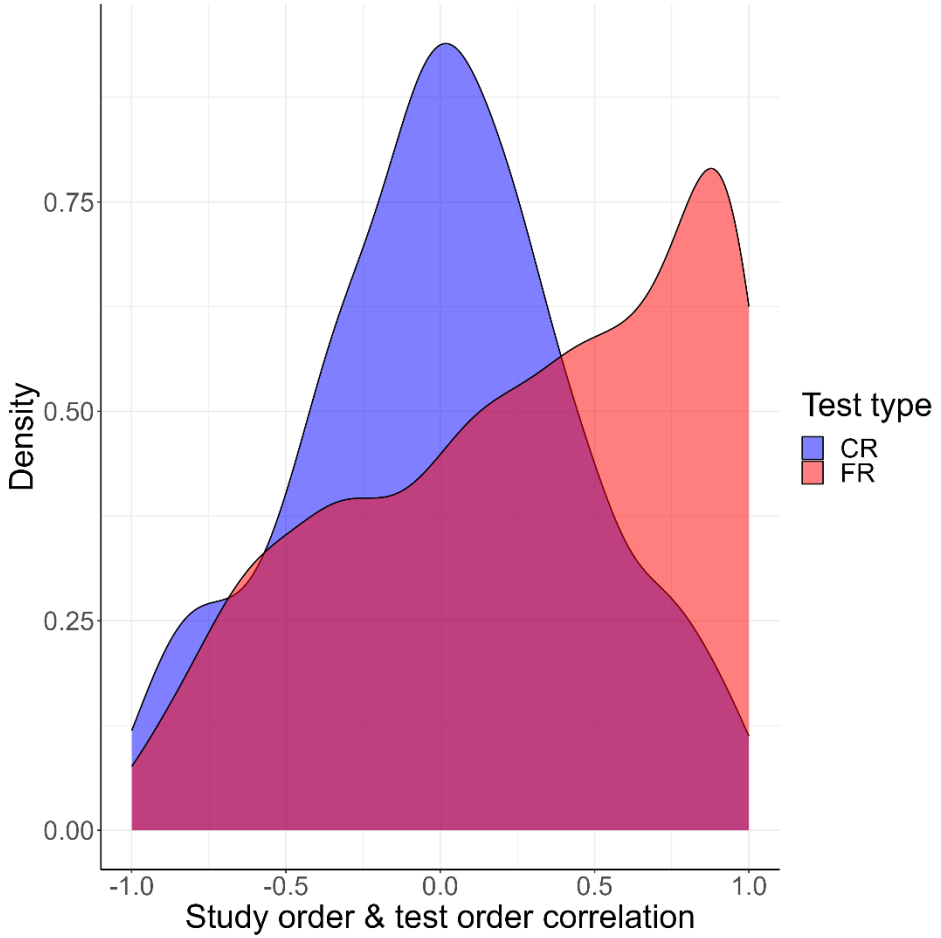
b. Concordance with 'normative' recall order



The figure above shows CR proportion recalled at the list level as a function of *CR discrepancy from the 'normative' FR output order*. By 'normative FR output order', we mean the order assembled by computing the most common position in the study list recalled at each position in the test list for free recall, for each experiment. We then computed each participant's discrepancy from this order, by obtaining the absolute difference between the actual test position and the normative test position. E.g., if the 1st studied CR pair was presented 5th, and if the normative recall position for the 1st studied FR word was 1st, the absolute difference for that CR pair would be 4. We then averaged the discrepancies for

each list, and predicted that list's accuracy from the average discrepancy, reasoning that participants who by chance ended up with an order closer to the normative order might have higher performance. We did not find compelling evidence for such a relationship.

C. Correspondence between study order and test order



The figure above shows the distributions of computed Pearson's *r* correlations between study and test order (at the list level), for all experiments. As the figure suggests, variance in the correlation coefficients was greater for FR than for CR, $F(662, 852) = .65$, $p = 6.11$, CR:FR variance = .65.

9. Unimodality vs. Multimodality

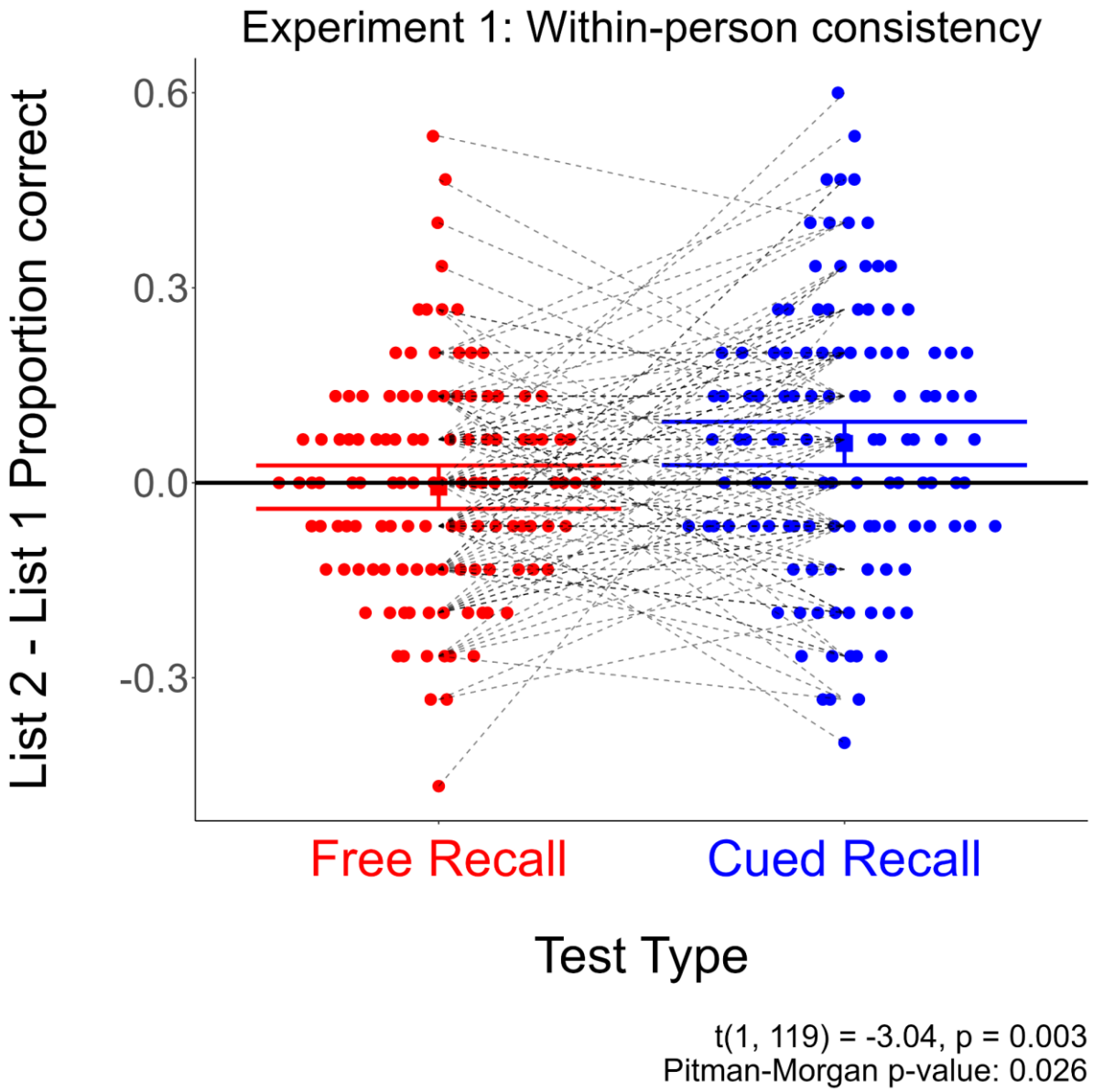
	Hartigan's Dip Test for Unimodality			
	Free Recall		Cued Recall	
Experiment	D_n	p	D_n	p
1	0.045	0.0745	0.041667	0.1465
2A	0.060504	0.0025	0.055672	0.009
2B	0.07874	< .001	0.059055	0.001
3	0.060976	0.002	0.051095	0.0075
4	0.078616	< .001	0.09204	< .001
5	0.067204	< .001	0.056452	0.0075

Unreported	0.061475	0.0015	0.06694	< .001

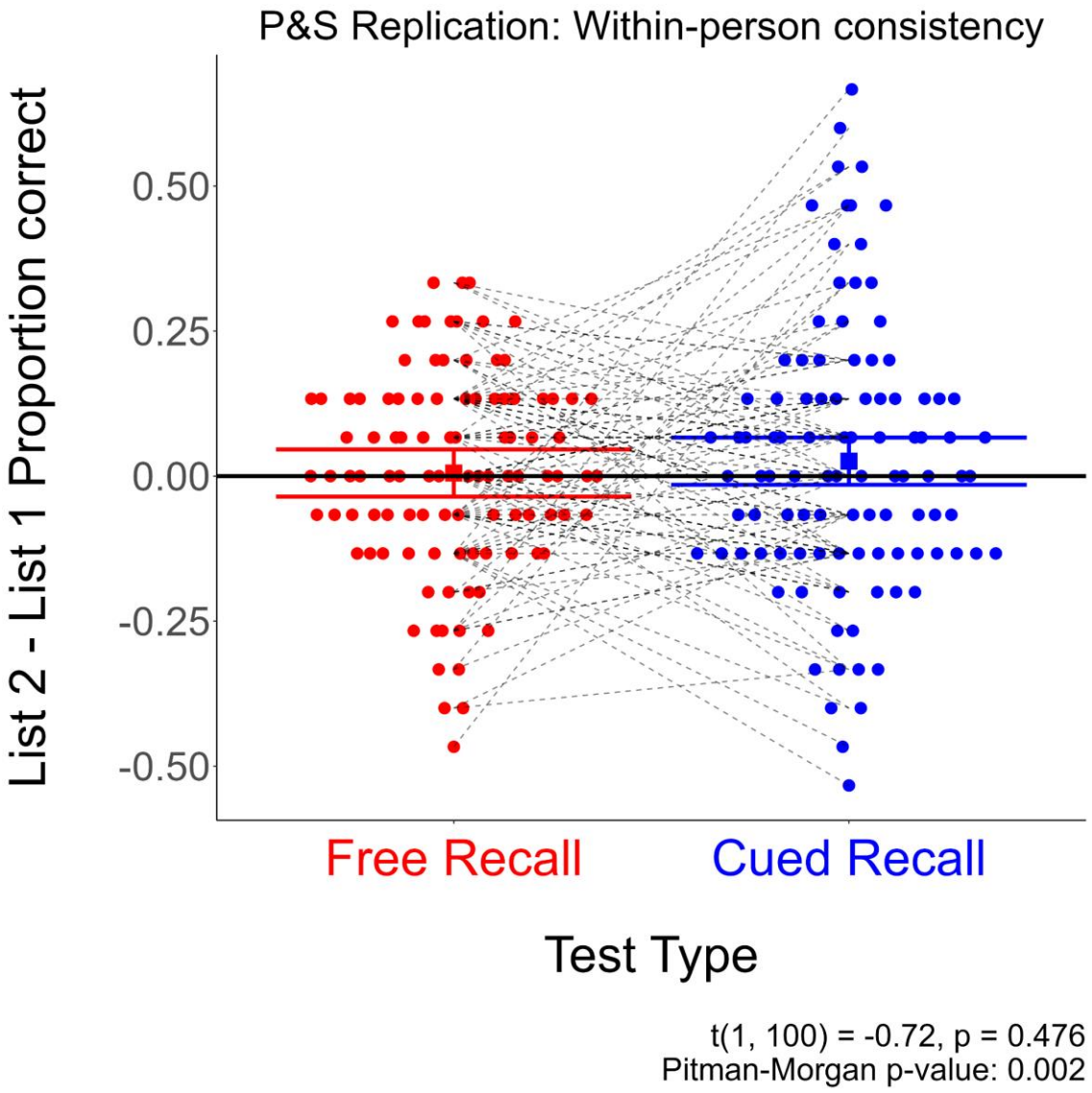
Note. D_n = Dip statistic. Significant *p*-values indicate evidence against the null hypothesis of unimodality.

10. Within-person consistency

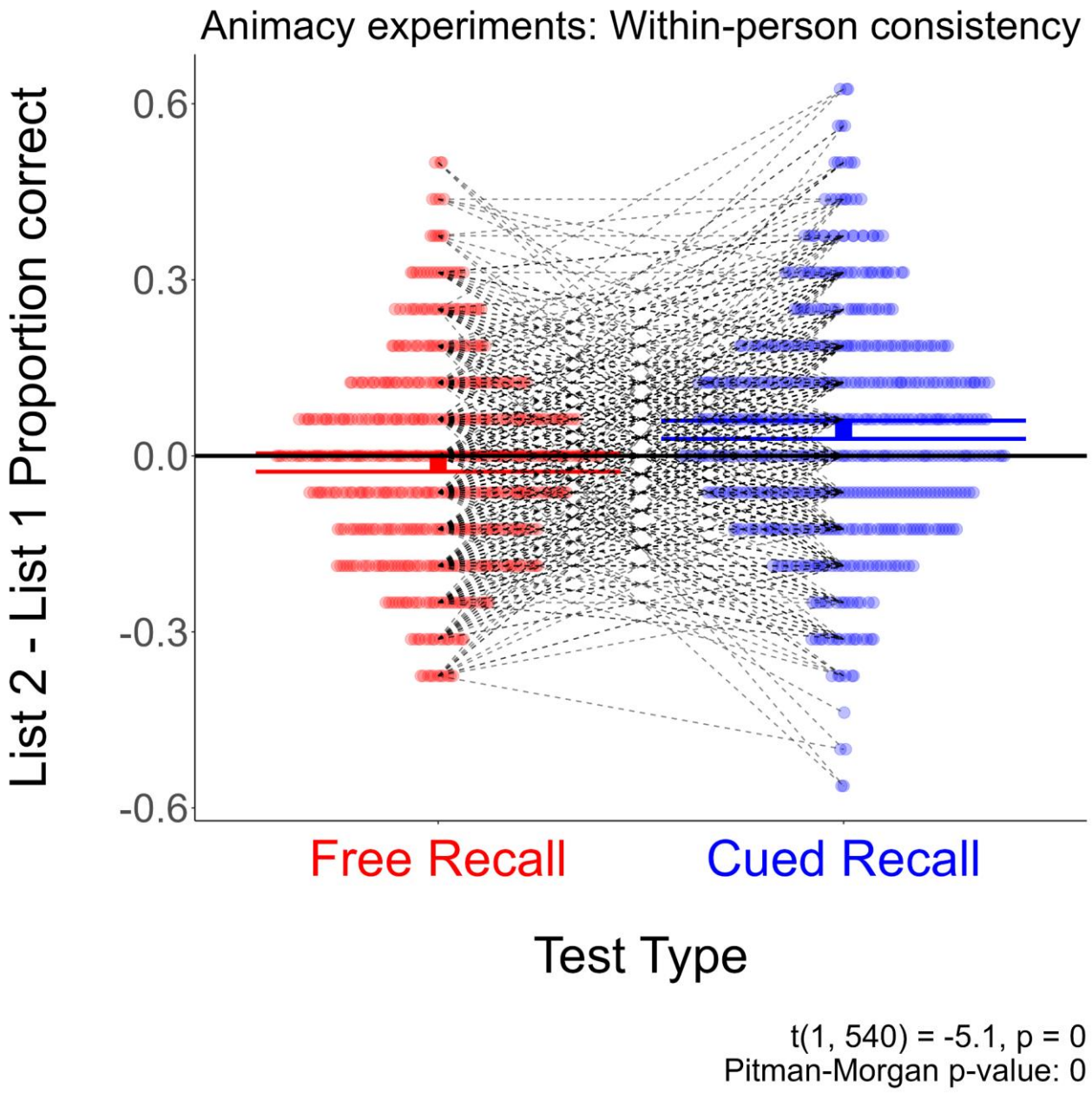
A. Experiment 1


Experiment 1: Within-person consistency
List 2 - List 1 Proportion correct
0.6
0.3
0.0
-0.3
Free Recall
Cued Recall
Test Type
t(1, 119) = -3.04, p = 0.003
Pitman-Morgan p-value: 0.026

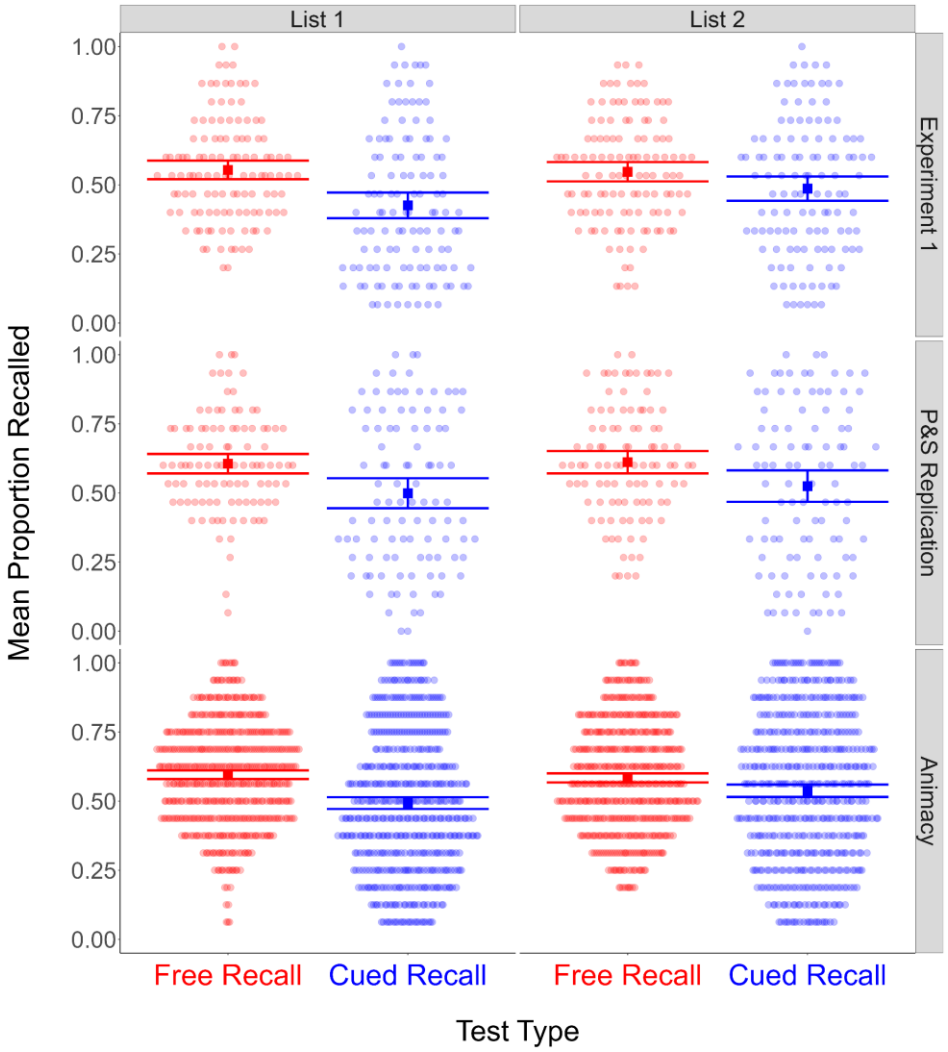
B. Popp & Serra (2016) Replication


P&S Replication: Within-person consistency
List 2 - List 1 Proportion correct
0.50
0.25
0.00
-0.25
-0.50
Free Recall
Cued Recall
Test Type
t(1, 100) = -0.72, p = 0.476
Pitman-Morgan p-value: 0.002

C. Animacy experiments


Animacy experiments: Within-person consistency
List 2 - List 1 Proportion correct
0.6
0.3
0.0
-0.3
-0.6
Free Recall
Cued Recall
Test Type
t(1, 540) = -5.1, p = 0
Pitman-Morgan p-value: 0

11. CR:FR accuracy by list in multi-list experiments



Supplementary References

*Gabry, J., & Cesnovar, R. (2021). cmdstanr: R Interface to 'CmdStan'. https://mc-stan.org/cmdstanr, https://discourse.mc-stan.org.

*Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., & Rubin, D.B. (2013). *Bayesian Data Analysis: Third Edition*. Chapman & Hall/CRC.

*Kader, G.D., & Perry, M. (2007). Variability for categorical variables. *Journal of Statistics Education, 15*(2), 4. DOI: 10.1080/10691898.2007.11889465

*Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*(5), 1413-32. DOI 10.1007/s11222-016-9696-4

*Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association, 22*(158), 209-212. DOI: 10.1080/01621459.1927.10502953