Variability across Subjects in Free Recall Versus Cued Recall

Eric Y. Mah

D. Stephen Lindsay

University of Victoria

This is a pre-print of an accepted manuscript of an article published by Springer Nature in *Memory & Cognition*. Please reference the final publication. The final proofed version might differ from the current document.

Declarations

Funding. This work was supported by an NSERC Discovery grant (#RGPIN-2016-03944) awarded to DSL.

Conflicts of interest/Competing interests: We do not have any conflicts of interest to declare.

Ethics approval. All the experiments reported herein were approved by the ethics review board of the University of Victoria, and were conducted in accordance with the ethical standards as laid down in the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards.

Consent to participate. All participants who took part in the experiment consented to participate.

Consent for publication. Both authors consent to the publication of this manuscript.

Availability of data and materials. All data/experiment programs are available at https://osf.io/3tra5/

Code availability. All analysis scripts/experiment programs are available at https://osf.io/3tra5/

Authors contributions. EYM and DSL conceived of the experiments, EYM programmed the experiments, collected the data, and analyzed the data. EYM and DSL drafted and revised the manuscript.

Acknowledgements. We would like to thank Henry L. Roediger, Colleen M. Kelley, John Dunlosky, Larry Jacoby, Reed Hunt, and Roger Ratcliff for their helpful insights and suggestions.

Open Practices Statement

The data and materials for all experiments are available at https://osf.io/3tra5/, and all experiments were preregistered: Experiment 1 (https://doi.org/10.17605/OSF.IO/XFJ6A), Experiments 2A and 2B (https://osf.io/3w6fm), Experiment 3 (https://osf.io/v67gy), Experiment 4 (https://osf.io/de7bu), and Experiment 5 (https://osf.io/my53w).

Abstract

Memory scientists usually compare mean performance on some measure(s) (accuracy, confidence, latency) as a function of experimental condition. Some researchers have made within-subject variability in task performance a focal outcome measure (e.g., Yao et al., 2016). Here we explored between-subject variability in accuracy as a function of experimental conditions. This work was inspired by an incidental finding in a previous study in which we observed greater variability in accuracy of memory performance on cued recall (CR) versus free recall (FR) of English animal/object nouns (Mah et al., 2023). Here we report experiments designed to assess the reliability of that pattern and to explore its causes (e.g., differential interpretation of instructions, (un)relatedness of CR word pairs, encoding time). In Experiment 1 (N = 120 undergraduates), we replicated the CR:FR variability difference with a more representative set of English nouns. In Experiments 2A (N = 117 Prolific participants) and 2B (N = 127 undergraduates), we found that the CR:FR variability difference persisted in a forced-recall procedure. In Experiment 3 (N = 260 Prolific participants), we used meaningfully related word pairs and still found greater variability in CR than FR performance. In Experiment 4 (N = 360 Prolific participants), we equated CR and FR study phases by having all participants study pairs and again observed greater variability in CR than FR. The same was true in Experiment 5 (N = 120 undergraduates), in which study time was self-paced. Comparisons of variability across subjects can yield insights into the mechanisms underlying task performance.

Variability across Subjects in Free Recall Versus Cued Recall

Suppose that some participants were tested on free recall (FR), in which individually studied words are to be recalled in any order at test, whereas others were tested on paired-associates cued recall (CR), in which random cue-target word pairs are studied and targets are to be recalled in response to cues at test. Would variability across participants in proportion accurately recalled be comparable for FR and CR, greater for FR, or greater for CR? What is your intuition about variability across subjects on these two memory tasks?

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, Mah et al. replicated both of those findings.

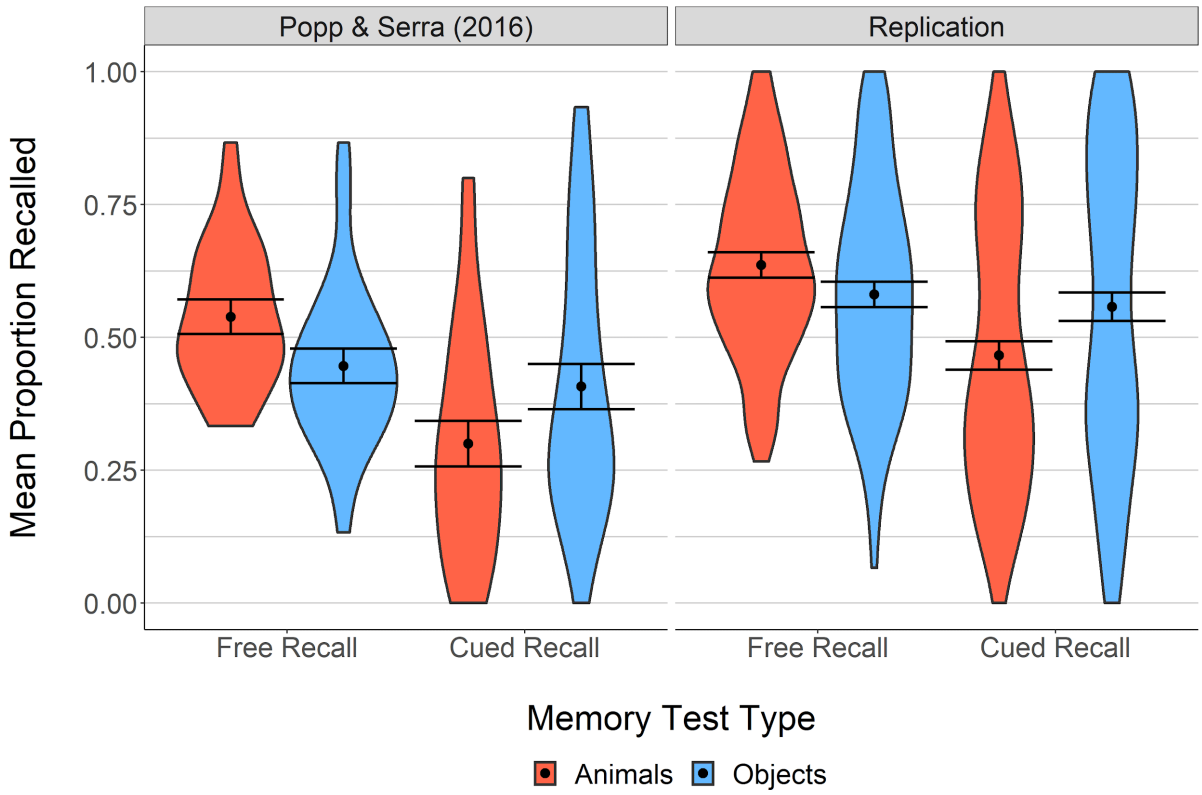
Looking at this figure, we were struck by the greater variability in CR scores than in FR scores, both in Popp and Serra (2016) and in our replication. The Pitman-Morgan test of equal variances for paired samples (Morgan, 1939; Pitman, 1939) indicated that variance in CR proportion correct was greater 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.

Figure 1

FR and CR memory performance in Popp and Serra (2016) and replication data from an ongoing project



Note: Error bars in this figure are 95% confidence intervals for the within-subject comparison between animals and objects, as per Loftus and Masson (1994).

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

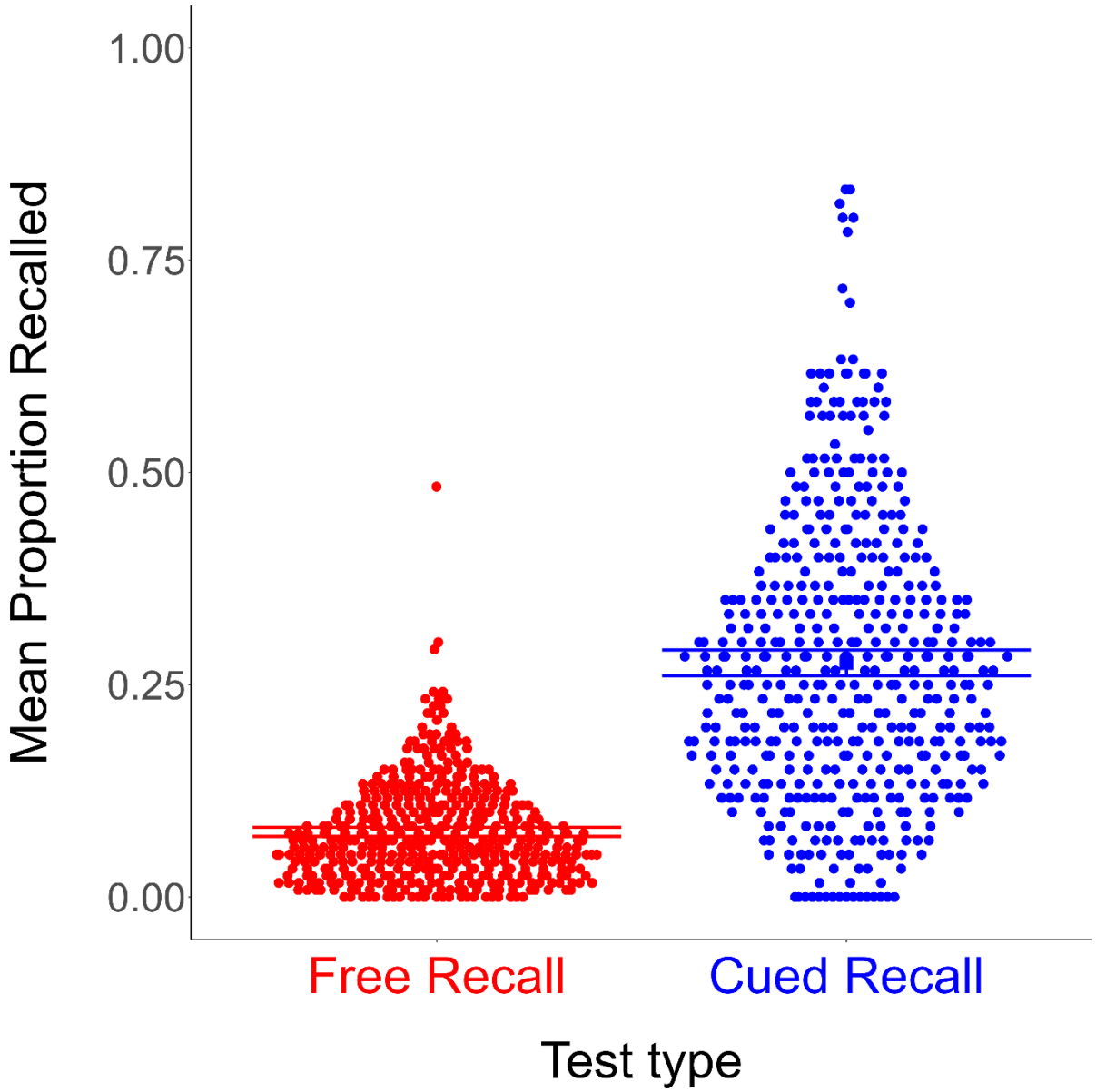
- Colleen Kelley replied "I am actually surprised to think that there would be more strategies available for paired associates than free recall, wouldn't you think there are more constraints in paired associates, so less room for variation?"
- John Dunlosky responded, "If I hadn't read your note first, 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 2

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 difference in inter-individual variability across tasks.

Why might such a difference exist? There are a number of salient differences between FR and CR that are worth considering. Some are more methodological – study time per word/pair (e.g., less time to encode each word in a CR pair than an individual FR target), the nature and strength of associations among words in the list and words in individual pairs, and the number of words studied (e.g., unrelated, meaningfully related), and the total number of words studied (e.g., more total words to encode in a CR task with the same number of pairs as individual FR targets). Some are more theoretical, based on hypothesized underlying mechanisms of encoding and retrieval. In one prominent model of memory – the Search of Associative Memory (SAM) model (Raaijmakers & Shiffrin, 1980) – at encoding, FR items build associations with one another in a buffer as they are studied, whereas in CR, only specifically paired words are associated. At retrieval, FR items serve as cues for one another, whereas in CR, only the paired cue serves as a cue for a particular target. In the SAM model, FR and CR retrieval proceed using general context and particular items as cues. However, the degree to which context versus item cues contribute to retrieval can differ across tasks – if item cues are

not useful, FR retrieval can be based on context cueing alone, but in CR, both item (the specific cue) and context cues inform retrieval. Differences in variability between FR and CR could be due to individual differences in the effectiveness of developing associations between items, pairs, or context (at encoding), or the relative weight given to context versus item cues (at retrieval). These differences could be in turn due to differences in the range of encoding and/or strategies adopted for FR versus CR.

Before attempting to sift through these varied potential theoretical explanations and the potential implications of robustly greater across-subject variability in CR relative to FR, we thought it wise to determine (a) whether the effect was replicable and (b) if so, whether the effect is due to spurious methodological factors (such as the ones described above) or quirks of our experimental designs or stimulus sets.

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, imagability, and familiarity).[1] The experiment program itself was a modified 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/274qd/?view_only=e9336b66e096474a837466f7e4ae3786.

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

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

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[2]

Confirmatory analyses

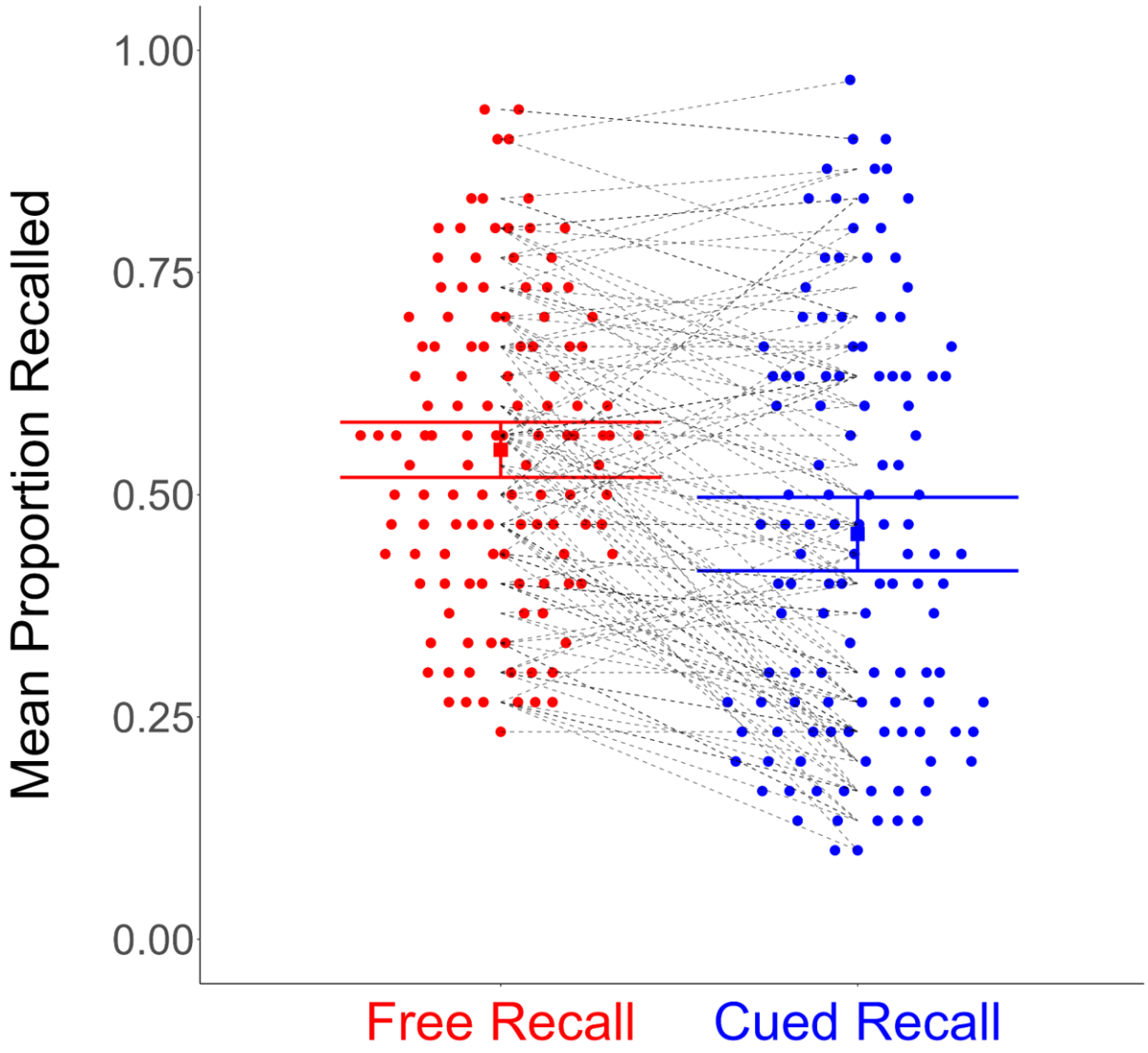
Our critical hypothesis was that inter-individual variability would be greater for CR than

[2] 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/?view_only=e9336b66e096474a837466f7e4ae3786.

for FR. Figure 3 depicts the means, within-subjects 95% CIs, and distributions of CR and FR performance in our sample.

Figure 3

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, $t(118) = 4.27$, $p < .001$. The estimated ratio of CR:FR variance (via bootstrap) 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.[3]

As the above analysis assumed roughly normal distributions in proportion recalled, 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).[4] Thus, we found evidence for the CR variability effect with a general noun wordset.

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*) and found a non-significant difference in variability in strategy use (See Supplementary Material 1 for a description of the coding process and Supplementary Material 3D for the specific Experiment 1 results). 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 3E).

---

[3] 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.

[4] 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)

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."

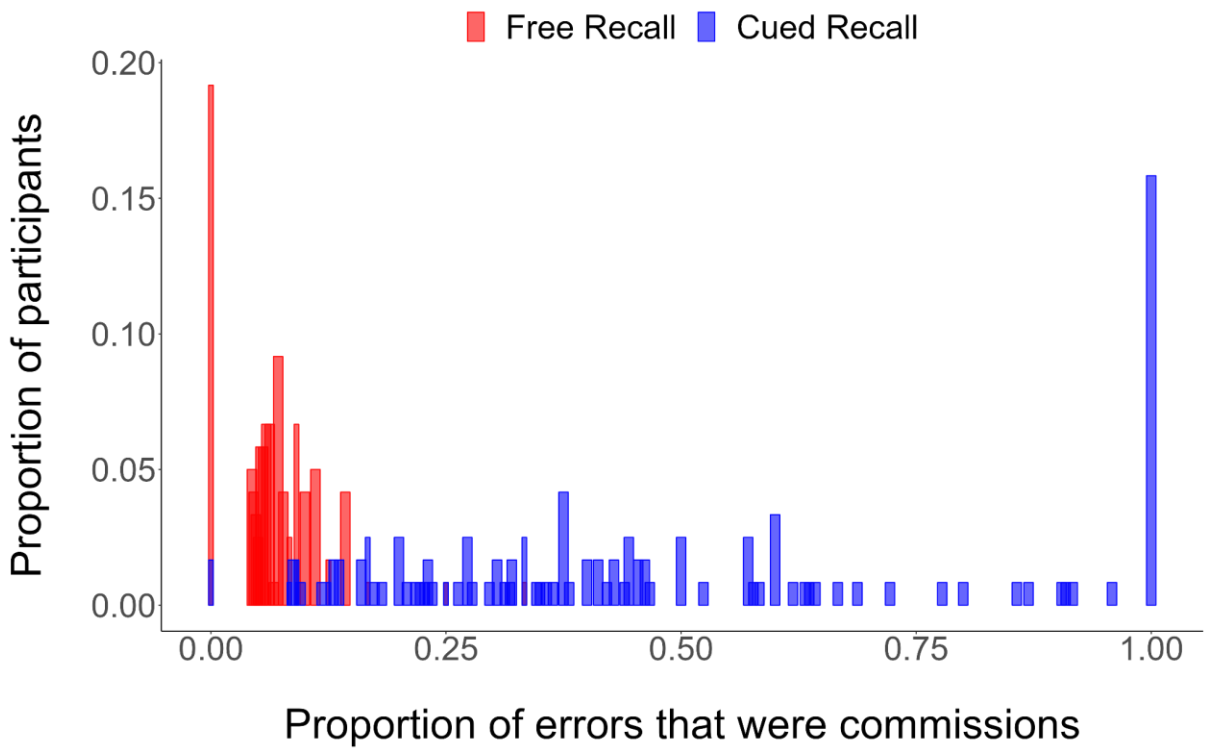
To investigate this possibility, we conducted an exploratory analysis of participants' tendency to make errors of *omission* (i.e., a failure to recall a given target) versus errors of *commission* (i.e., an incorrectly recalled target – from a current list, previous list, or new intrusion not previously 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 4.

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." To better address these possibilities, we conducted new Experiment 2.

Figure 4

Experiment 1: Commission error proportion by recall type



Experiment 2A & 2B: "Forced recall"

As we suggested in Experiment 1, 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.[5] 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/?view_only=da1006ae5b064efca0f3997461a56525.

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 recall does not matter for them to be counted correct.

…and for CR:

[5] 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.

> 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 three 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[6], from which we excluded: 12 participants who

[6] 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

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[7], 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 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

---

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.

[7] 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.

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

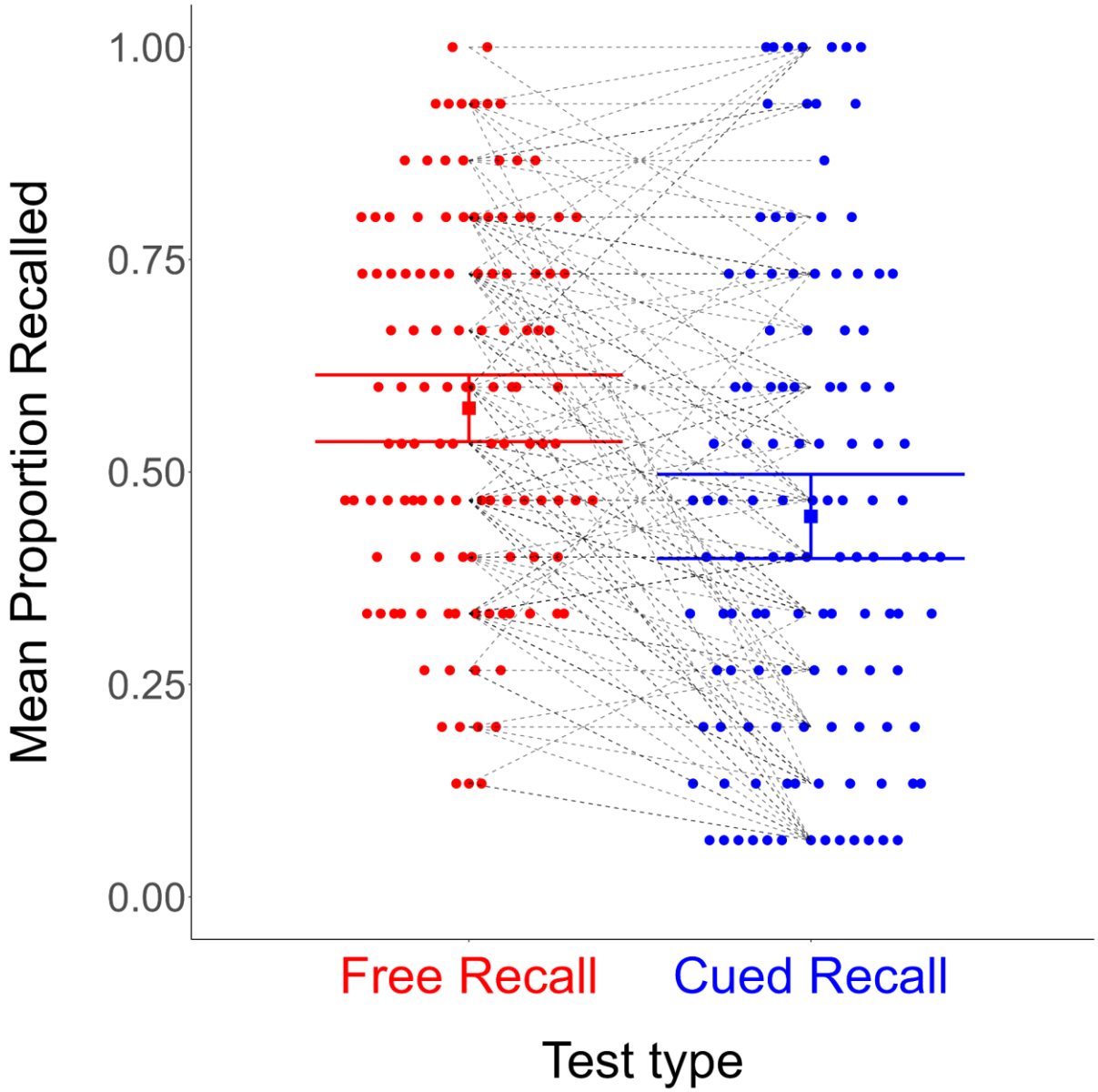
Data files and analysis scripts for Experiments 2A and 2B are available at https://osf.io/yv3b7/?view_only=da1006ae5b064efca0f3997461a56525.

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 5:

Figure 5

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]).[8] This ratio is slightly smaller than that observed in Experiment 1 (by about 5%).[9] 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.

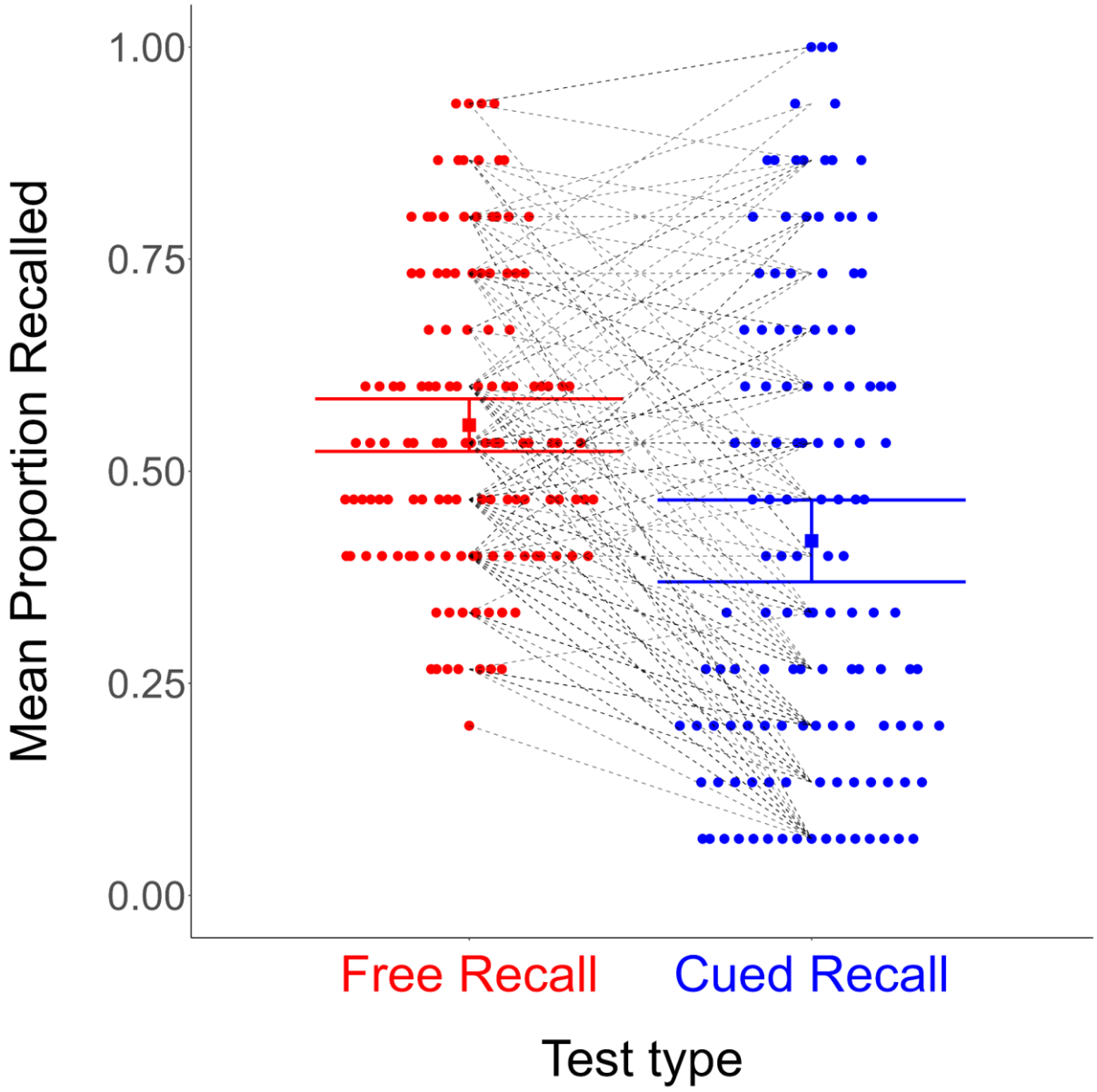
Experiment 2B. We conducted the same analyses for Experiment 2B, in our undergraduate sample. Memory performance as a function of recall test type (only treating targets recalled with matching cues as correct) is shown in Figure 6.

[8] 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).

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

Figure 6

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%.[10] The corresponding generalized mixed-effects logistic regressions also provided clear evidence of a CR:FR variance difference (See Supplementary Material 5C).[11] 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 5 and 6 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 variability effect cannot be explained purely in terms of

[10] 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.

[11] 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]).

instructional ambiguity or differences 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.

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. 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).

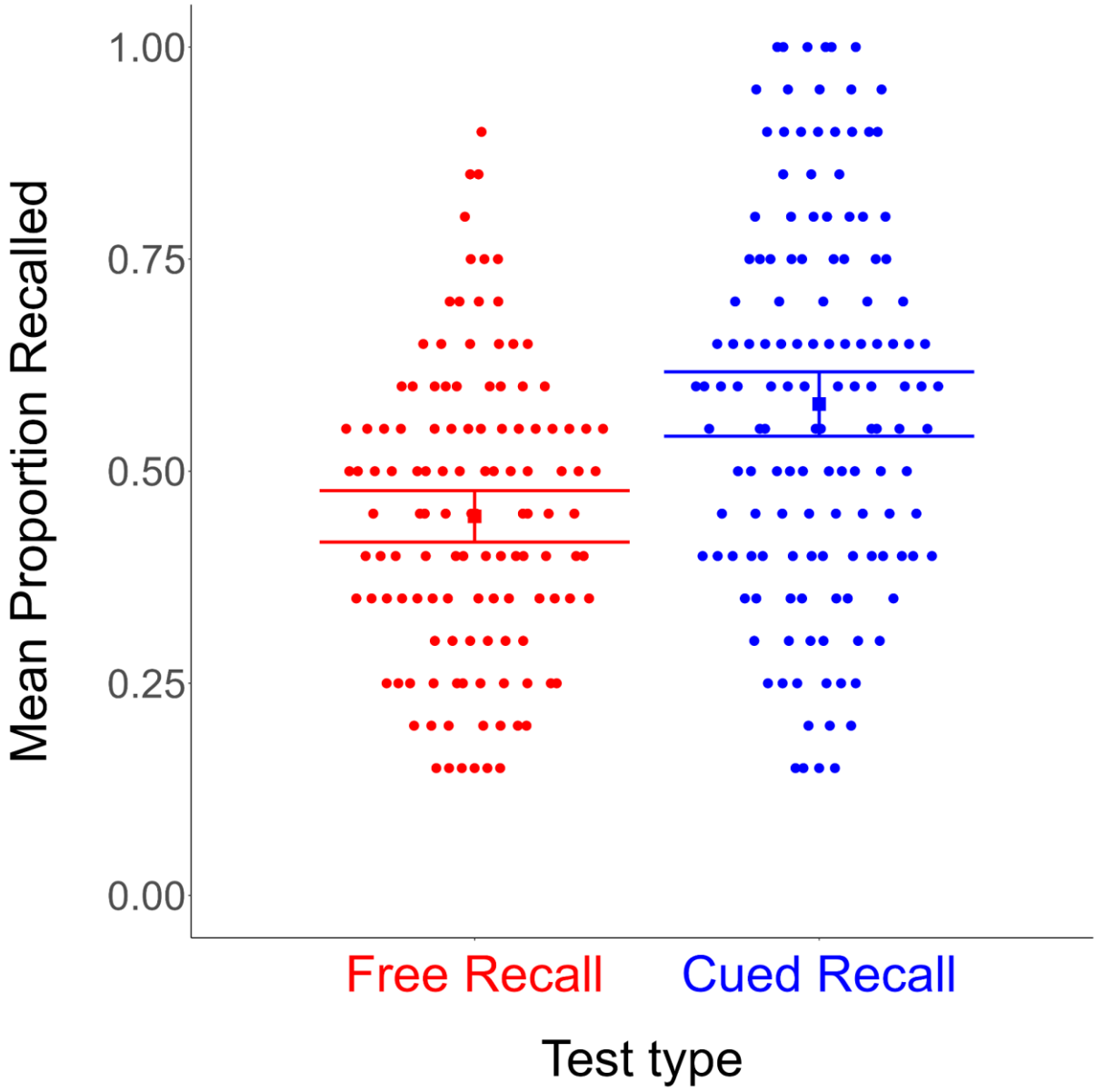
Results

Data files and analysis scripts for Experiment 3 are available at https://osf.io/pfhu9/?view_only=5e75abb06f1348f28f0d13c0a0fc6e08.

As in previous experiments, we compared variability in CR and FR memory performance. Memory performance by test type is shown in Figure 7.

Figure 7

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.

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.

Experiment 4: "Equivalent study lists"

In the following two experiments, we considered possible methodological explanations for the CR:FR variance effect. In comments on an earlier version of this paper, Larry Jacoby (personal communication) and action editor Jen Coane independently noted that 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/?view_only=58e5ae97bbef433d8c32f617145bb3f7). 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

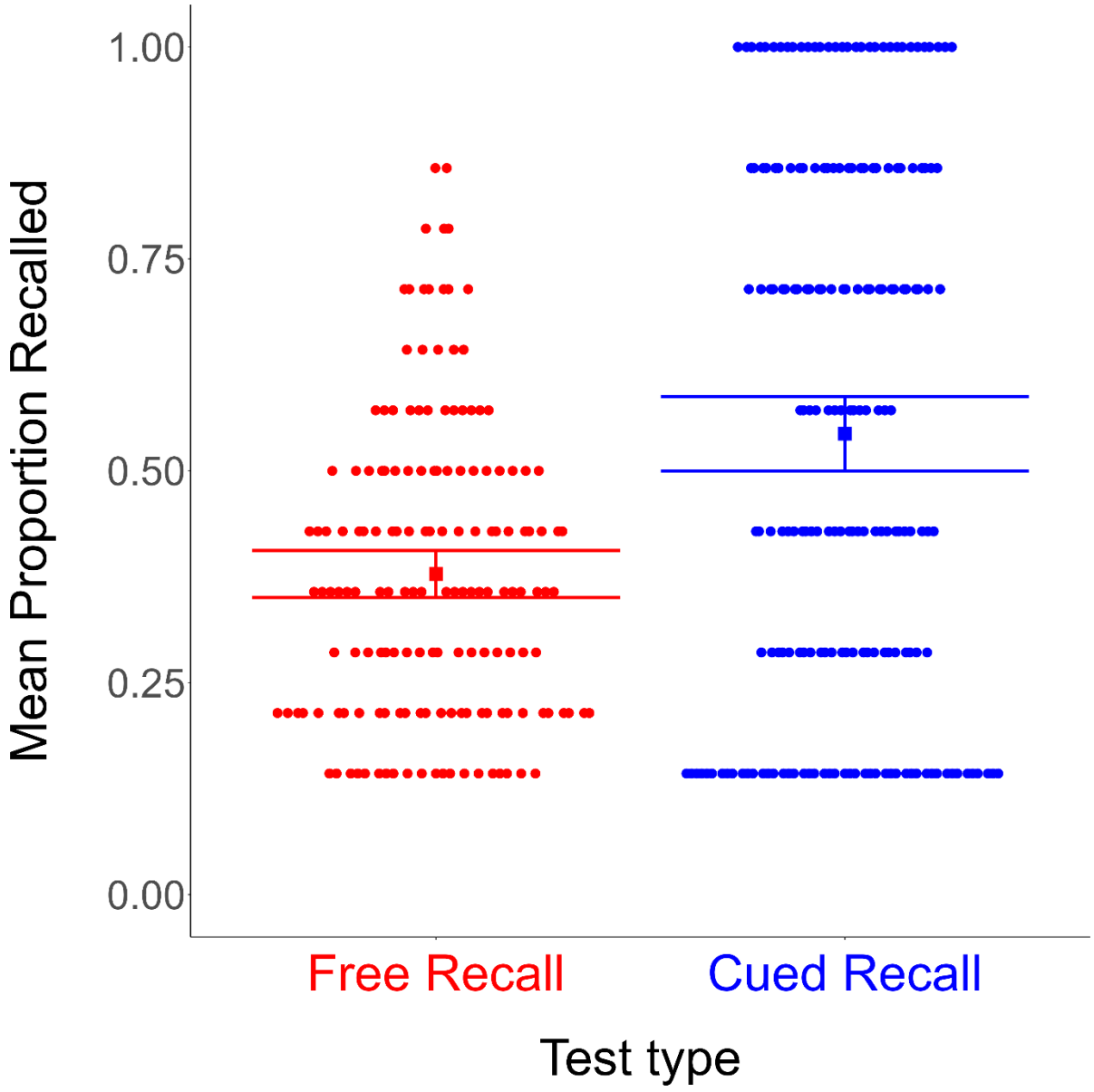
Via power simulations, we set a post-exclusion target $N = 360$. We collected data from a total of $N = 492$ Prolific.co participants and based on preregistered criteria excluded: 82 participants who did not get at least one (two) correct on CR (FR), 52 participants who reported understanding less than 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

First, a visualization of CR and FR performance:

Figure 8

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/?view_only=58e5ae97bbef433d8c32f617145bb3f7), 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[12]. 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

[12] 90% because our hypothesis of CR > FR variability was one-sided.

number of studied words does not account for the CR:FR variability difference.

Experiment 5: "Self-paced study"

In their signed reviews of an earlier version of this paper William Hockley and Adam Osth drew our attention to an additional potential methodological confound—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/yv3b7/?view_only=da1006ae5b064efca0f3997461a56525).

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[13]. The experiment was programmed in PsychoPy (see https://osf.io/xdsk8/?view_only=05243dd0b65e4c4193e413030e802939) 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 ($M = 20.27$, $SD = 3.40$)[14], with 68 completing CR first and 56 completing FR first.

Results

Recall accuracy

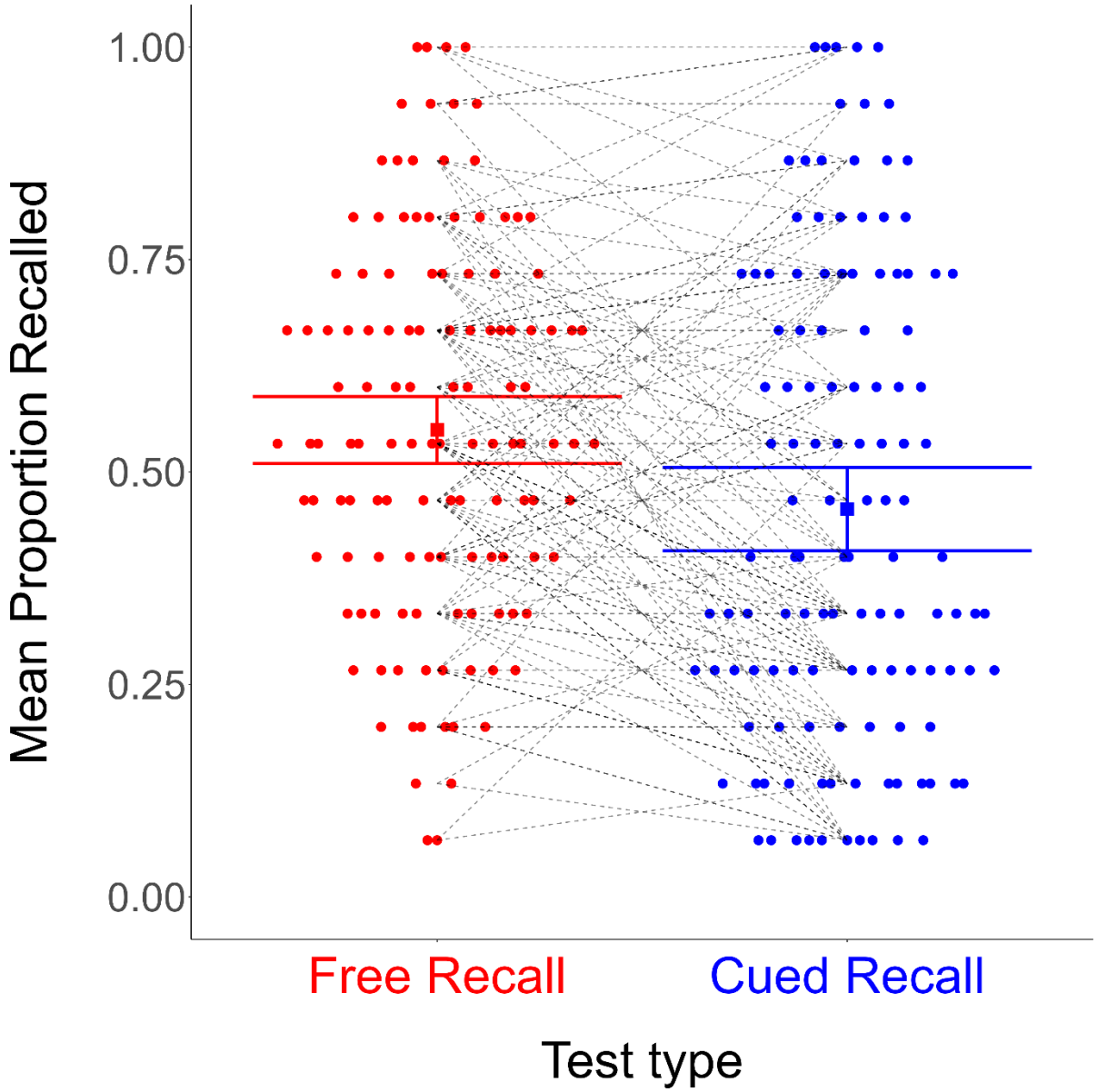
Figure 9 shows proportion correct for FR and CR when the study phases were self-paced.

[13] 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.

[14] 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 9

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.

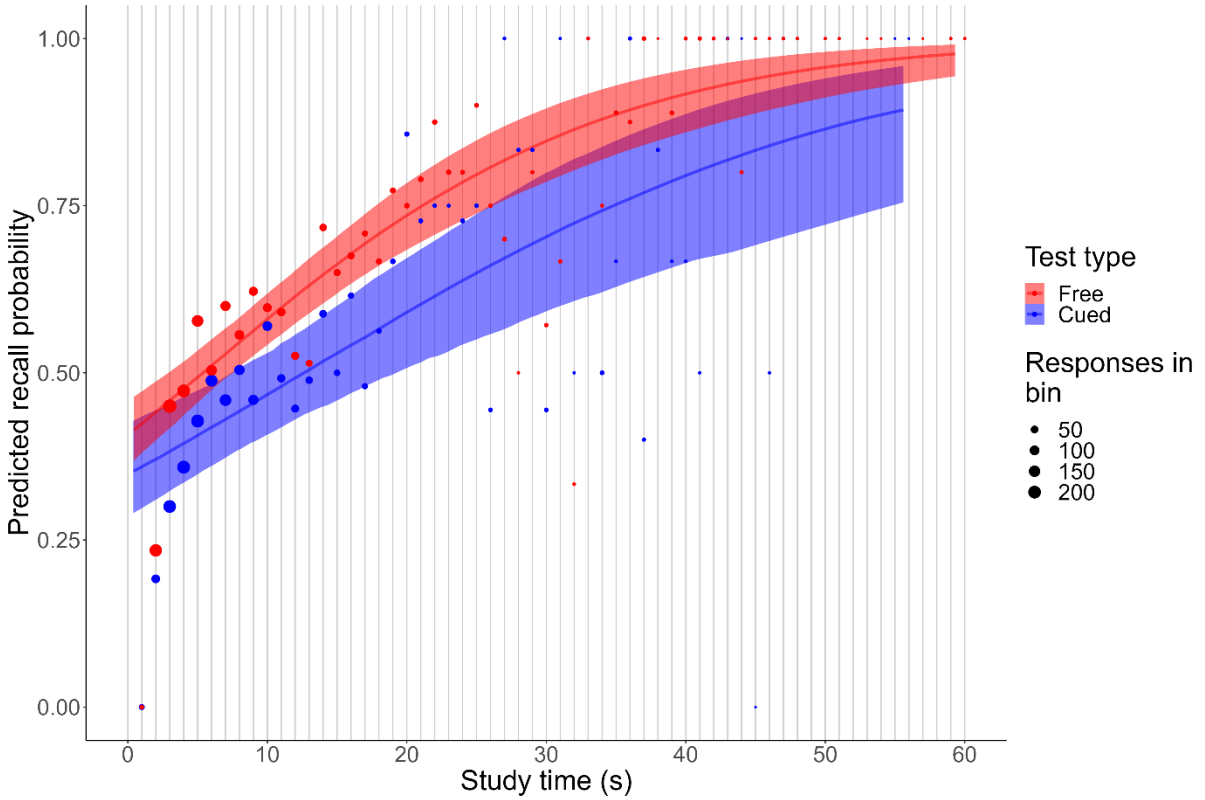
Study time

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 10 shows accuracy (empirical, model-predicted) as a function of study time.

Figure 10

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[15] GLMM posterior means and 95% credible intervals on the predicted accuracy mean at each study time. A total of 17 trials with study time > 60s were removed.

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

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

when excluding study times longer than 30s (*p*s < .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.

General Discussion

Across three experiments, we investigated an apparent difference in inter-individual variability in free recall versus cued recall memory performance. We observed this effect with a general set of English nouns in both student and non-student populations (Experiment 1). Our various manipulations suggested some preliminary boundary conditions and ruled out some potential explanations for the effect. Specifically, we found evidence against the possibilities that increased CR variability was due to greater variability in interpretation of CR versus FR instructions, or more propensity to guess on CR versus FR (Experiments 2A and 2B). We also found evidence against the possibility that the CR variability effect occurs only with random/unrelated word pairs (Experiment 3). Finally, we found evidence against the possibility that greater CR variability is a result of more words that must be held in memory (Experiment 4) or shorter per-word encoding time (Experiment 5).

Having ruled out what we believe to be the primary possible methodological explanations, we can return to potential theoretical explanations. As previously suggested, the variability difference could be due to differences in the strategies that participants adopt. Given the lack of overt retrieval support in the FR task, participants may have gravitated toward reliance on temporal context for support. This aligns with our findings that participants predominantly reported using rehearsal/repetition for FR, and that participants varied slightly more in study strategies for CR (albeit only directionally). Although we did not find compelling

evidence for greater variability in CR than FR strategies in the experiments where we queried this, it is also possible our questions were not fine-grained enough to capture meaningful differences. For instance, there could be substantial variability in the strategies participants used to associate the words in the pairs (e.g., static versus interactive imagery). One might observe greater variability in CR strategies with more pointed questions (such as those used by Morrison et al., 2016).

Drawing again on the SAM model (Raaijmakers & Shiffrin, 1980), it could be that use of context versus item cues at encoding and/or retrieval differed across tasks. For example, in all our experiments, FR lists consisted of mostly unrelated words. In this case, item-item associations may have been less useful and participants more consistently bound items to general context during encoding, and used general context (which is purely a function of time spent in the buffer) as a cue at retrieval. For CR, item cues were inherent to the task and had to be used in concert with context cues. Here, factors like pre-existing semantic associations between cues and targets (which may differ across individuals) could have played a greater role, contributing to more pronounced individual differences. We hope that future research can uncover the nature of these individual differences and the factors that underlie them. For instance, in some experiments (1 and 4 in particular) there were striking qualitative differences in the shapes of the FR and CR distributions, with potential multimodality in CR accuracy[16]. This might suggest that the factors underlying individual differences in CR performance correspond to unidentified subgroups in our samples, or that the factors are more dichotomous than continuous.

[16] We conducted formal analysis of multimodality for FR and CR accuracy in all experiments via Hartigan's Dip Test of Unimodality (Hartigan & Hartigan, 1985), and observed evidence against unimodality for FR and CR in all experiments other than Experiment 1 (see SOM 9)

Another question that remains is whether the differences between CR and FR variability occur primarily at study (e.g., encoding processes) or test (e.g., recall strategies). In Experiment 5, we attempted to equate the study phases by having all participants study pairs (as did Cox et al., 2018). The presence of the variability difference despite this manipulation might suggest that the factors responsible occur primarily at test (e.g., variability in retrieval strategies or use of cues). In an additional, unreported experiment (full details and a link to the preregistration in Supplementary Material 6), we attempted to equate test phases by giving participants standard FR/CR study phases but then surprising CR participants with an FR test in which they simply had to recall as many targets as possible, in any order. In this experiment, we did not observe the CR:FR variability effect, but this interpretation was clouded by a floor effect for CR participants that likely limited variability. We conducted a series of additional pilots testing potential manipulations for getting performance off floor (see https://osf.io/n7c9h/wiki/home/), to no avail. Thus, the study-vs-test question remains a potential avenue for further examinations of the effect.

Finally, one salient methodological factor that may still be worth investigating is output order. In all of our experiments, participants were free to recall FR targets in any order they desired. In contrast, for CR, order was determined (randomly). It is possible that this factor may have contributed to the CR:FR variability difference. We conducted several exploratory analyses of output order across experiments and found that a) FR output order was generally consistent with prior literature and more homogenous across participants than CR output order (strong primacy and moderate recency effects, serial recall otherwise, e.g., Cowan et al., 2002; SOM 8A) and b) CR accuracy declined over the course of the test (i.e., evidence for output interference; SOM 8Ba).

However, further analyses cast doubt on the idea that output order explains our findings. First, variability in the *correspondence* between study order and recall order (via correlation) was greater for FR than for CR – despite the correspondence correlations having similar ranges for FR and CR (see SOM 8C). This was likely driven by a positively-skewed but wide distribution for FR (i.e., most participants tended toward serial recall, but varied, with some participants recalling in reverse serial order). Second, we considered the possibility that by chance, some CR participants ended up with test orders that might be more conducive to recall (and some with test orders that might be less conducive to recall). For each experiment, we determined the 'normative' FR output order (i.e., for each study position, the most common recall position) and for CR, each participant's deviation from that order. We reasoned that if output order played a role in the observed effects, participants with orders closer to the normative might have higher accuracy (and vice versa). This was not the case (see SOM 8Bb). Together, these results suggest that random CR test order does not explain the observed effects. However, these analyses were exploratory, and a future experiment directly manipulating CR test order would provide a stronger test of this possibility.

Limitations

One potentially damning criticism of this work might be that we have merely added another entry to the burgeoning catalogue of memory effects. Or that we are continuing a long tradition of studying research paradigms as opposed to advancing a theoretically grounded account of how and why memory functions as it does. This line of research grew out of looking at a raincloud plot of distributions of CR and FR scores and thinking "That's funny….why is variability greater for CR than FR?" In addition to evidence that this difference in variability is replicable, here we reported studies aimed at illuminating the underlying mechanism of that

difference. Our efforts to date met with limited success. Perhaps this approach is doomed to failure (Jamieson et al., 2016; van Rooij & Baggio, 2021).

Despite this, unexpected phenomena are inherently interesting and studying them has often turned out to be productive (Watkins, 1990). Although our examination is a clear case of putting the effect before the theory (van Rooij & Baggio, 2021), the effect we describe is potentially relevant to broad/general theories of memory that purport to explain associative and item memory (e.g., ACT-R, Anderson et al., 1998; SAM, Raaijmakers & Shiffrin, 1980). Our results here require that theories take into account the increased variability in cued recall relative to free recall, and delineate the mechanisms responsible for this difference (e.g., greater number of processes involved in associative memory relative to item memory). In other words, the CR variability effect serves as another constraint on theory and narrows the space of possible theories (Eronen & Bringmann, 2021). Formal computational models have been proposed as one way to constrain and specify psychological theories (Oberauer & Lewandowsky, 2019). Variability in memory test performance could serve as additional datum for such models, allowing the estimation of a greater number of theoretically-relevant parameters. We have drawn upon the SAM model (Raaijmakers & Shiffrin, 1980) as a theoretical framework for interpreting our results. However, we did not conduct model-fitting or simulation analyses to determine whether SAM and other relevant memory models such as ACT-R (Anderson et al., 1998) or TODAM (Murdock, 1982) make explicit predictions about differences in performance variability across the memory tasks. It is possible that these models can accommodate, or indeed predict this effect. Such an enterprise is beyond the scope of the current paper, but a comprehensive simulation and model-fitting analysis across varied datasets, models, and tasks represents a valuable next step in exploring the effect we describe here. Aside from the possible value for

theory development and testing, our work here reinforces the idea that performance variability is a worthwhile metric to examine in memory studies (Zerr et al., 2018).

Constraints on Generality

The terms “free recall” and “cued recall” can refer to a wide range of tasks (e.g., memory of a real-world experience might be assessed by asking for free recall of that episode or by providing cues to recall particular details). Our observations and speculations are specific to performance on the particular (quite artificial) tasks used in our experiments (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). Variability might be quite different under other conditions (e.g., incidental learning, long delays, more complex/naturalistic materials).

Our experiments tested English-speaking participants sourced from x and y, with English concrete nouns with x and y characteristics. We have shown generalizability across some sets of words, but it is possible that variability would differ with participants from different cultures and/or with different sorts of words or with non-word stimuli.

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).

Conclusions

We have presented evidence that individuals vary more from one another on cued recall tasks than on free recall tasks. Examinations of differences in inter-individual variability have proven fruitful in other domains (e.g., Christensen et al., 1999, LaPlume et al., 2021; Yao et al.,

2016), and application of these analyses to free and cued recall could provide greater insight into underlying mechanisms and overarching theories of memory.

References (* denotes Supplementary Material reference)

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

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

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.

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

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

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

*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.

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.

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

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

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

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

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1*(4), 476–490. https://doi.org/10.3758/BF03210951

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

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

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

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

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

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

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

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

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., Shoda, Y., & Lindsay, D. S. (2017). Constraints on Generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science, 12*, 1123-1128. DOI:10.1177/1745691617708630

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

*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

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

*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

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

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

Variability Across Subjects in Free Recall Versus Cued Recall:

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][17].

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

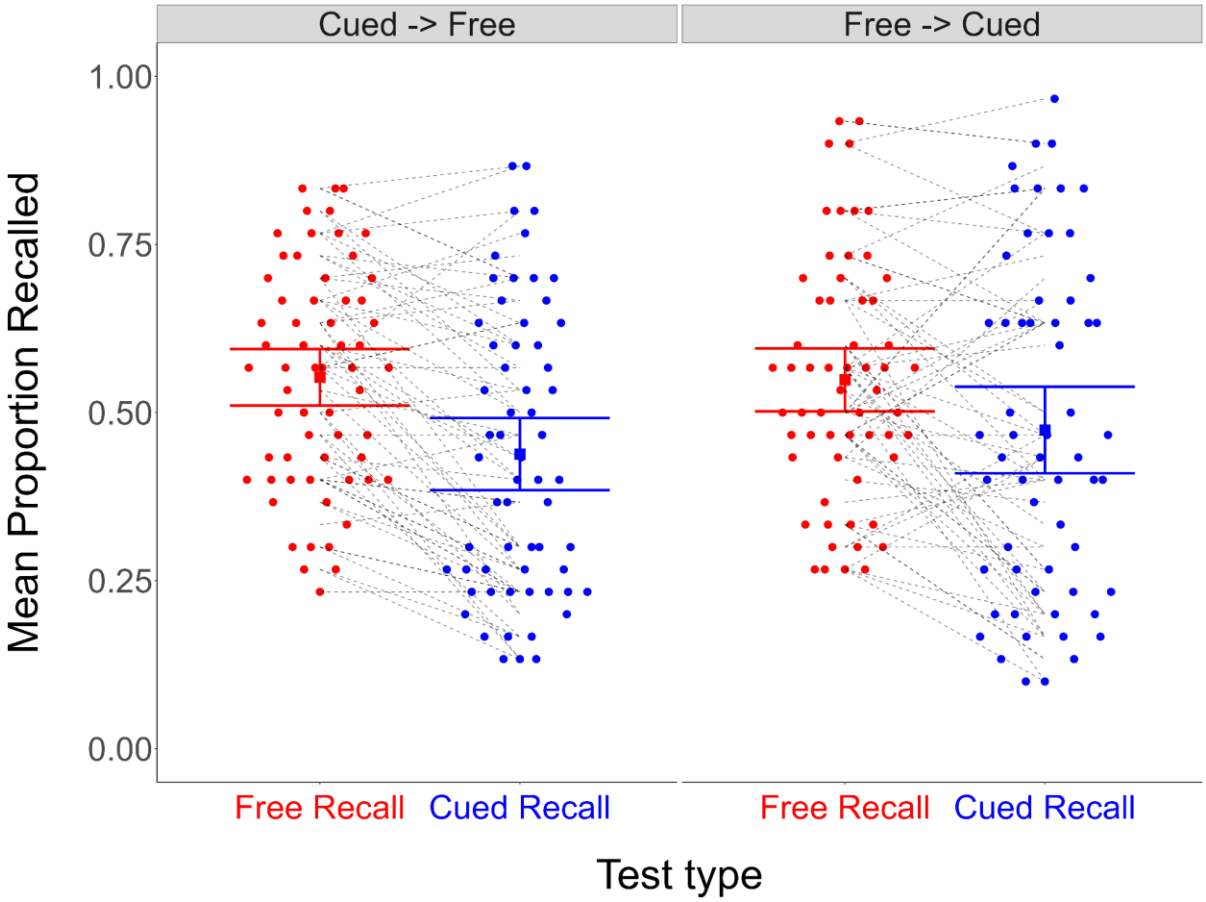
[17] 95% CI on the LOO difference was estimated by multiplying the SE of the difference by +/- 1.96.

memory types:

Test type	***SD* (Logit units)**	**95% CI lower**	**95% CI upper**
FR	.65	.54	.78
CR	1.00	.85	1.17

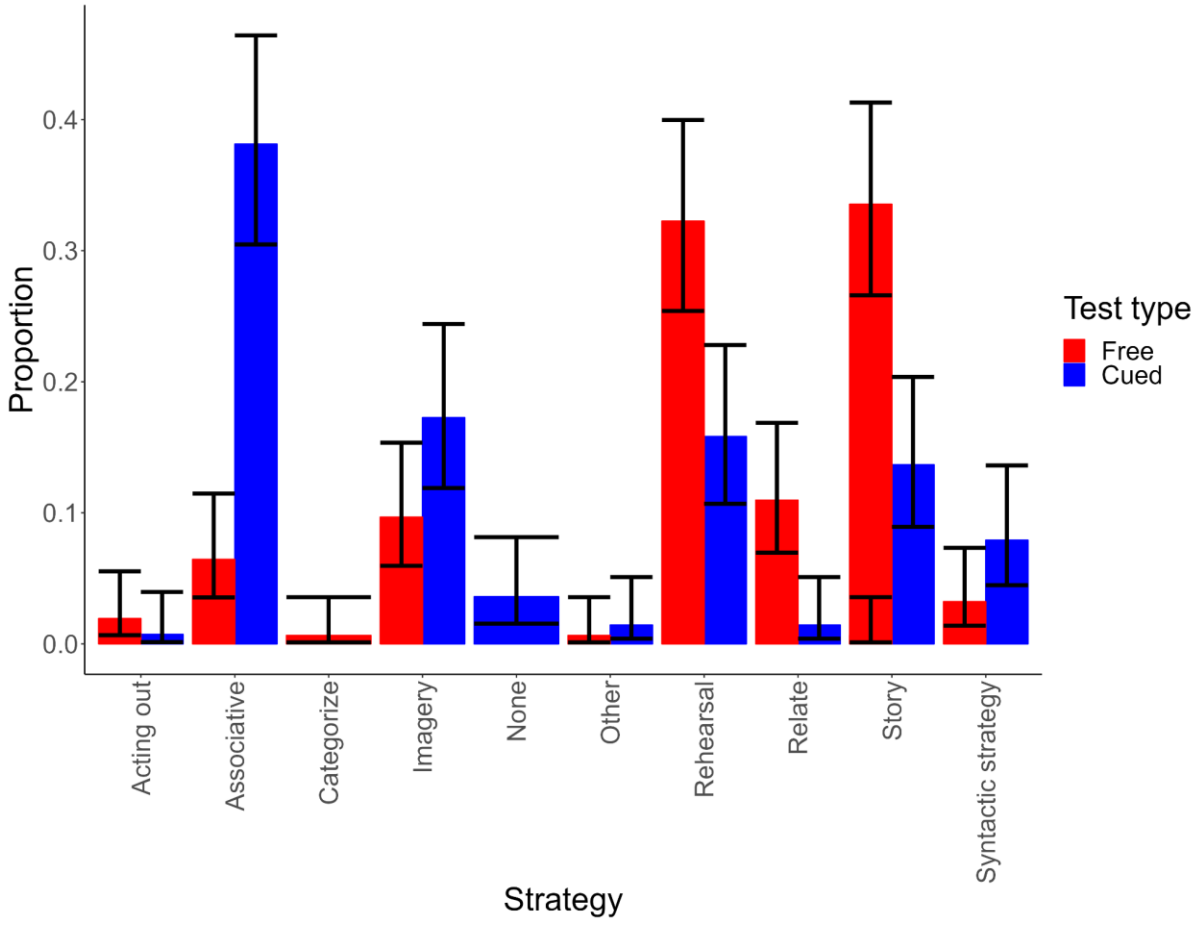
As the 95% CIs on the SD estimates did not overlap, this analysis provided evidence for differing FR/CR variances.

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, $\chi^2(1) = 1.74$, $p = .19$.

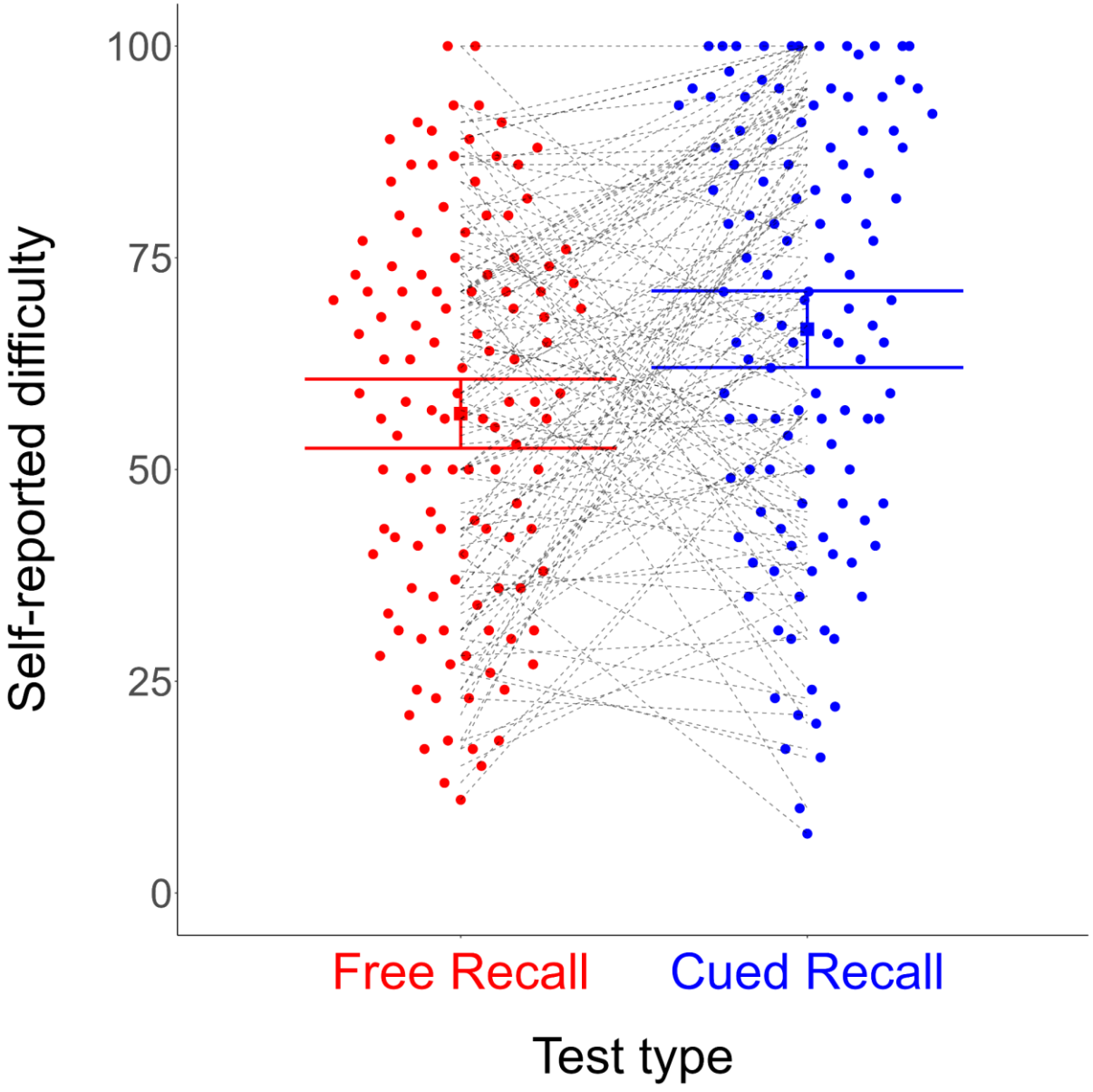
D. Self-reported study strategies. Of 330 coded responses (i.e., one FR and one CR response for all participants in the full sample), the two coders agreed on 177. The remaining 163 disagreements were put to a third coder. Final coded strategies were those that were reported by at least two out of three 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:



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]).

E. 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.

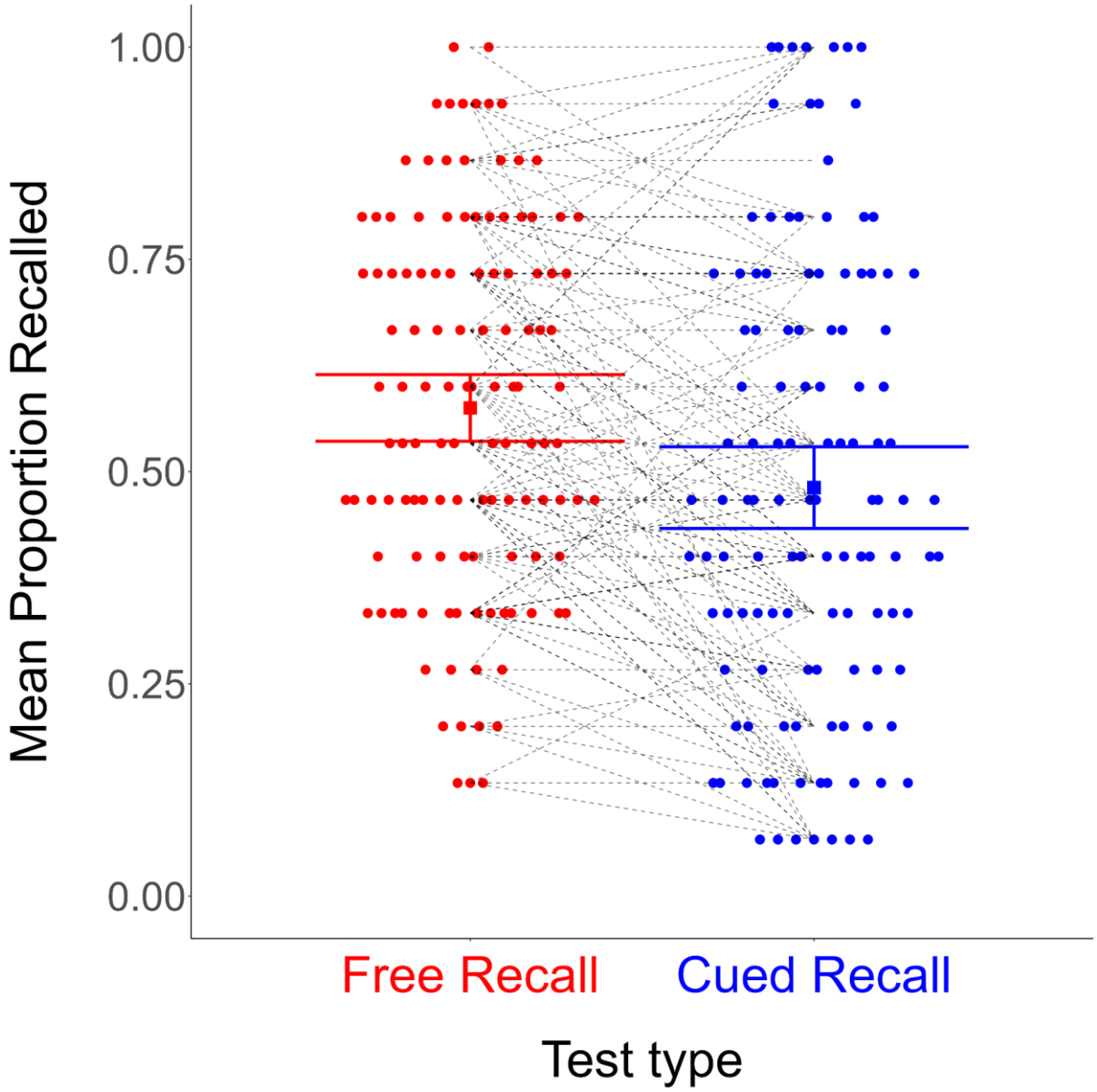
Other than a potential ceiling effect for CR difficulty ratings, there is no immediately obvious difference in variability ratings. Indeed, a Pitman-Morgan test failed to reject the null hypothesis of equal variances, $t(118) = 1.15$, $p = .25$, with a bootstrapped CR:FR variance ratio = 1.11 (95% percentile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

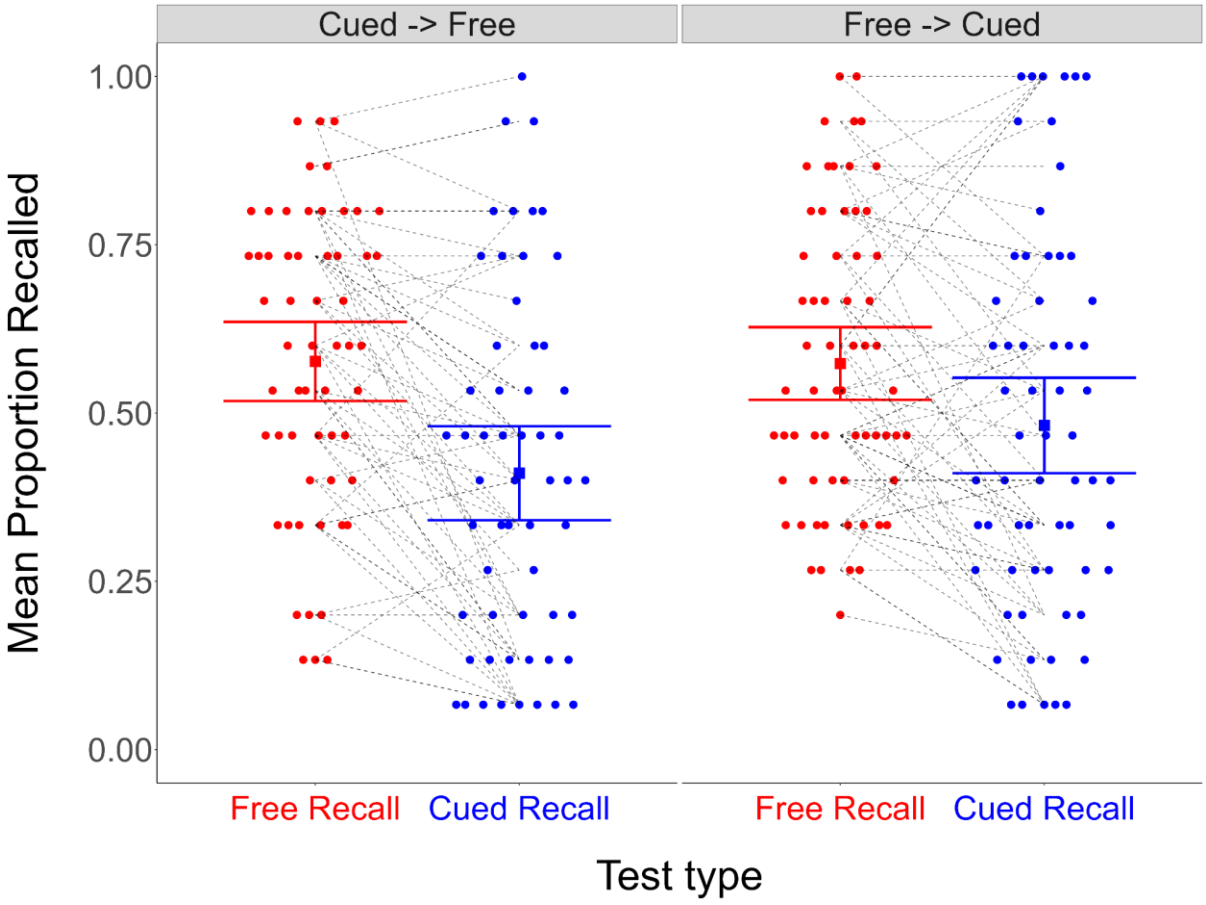
C. Generalized mixed-effects logistic regression results. We used the same GLMM model as in Experiment 1 to compare FR and CR variability (for standard CR proportion recalled and proportion recalled when same-list commission errors were counted as correct):

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

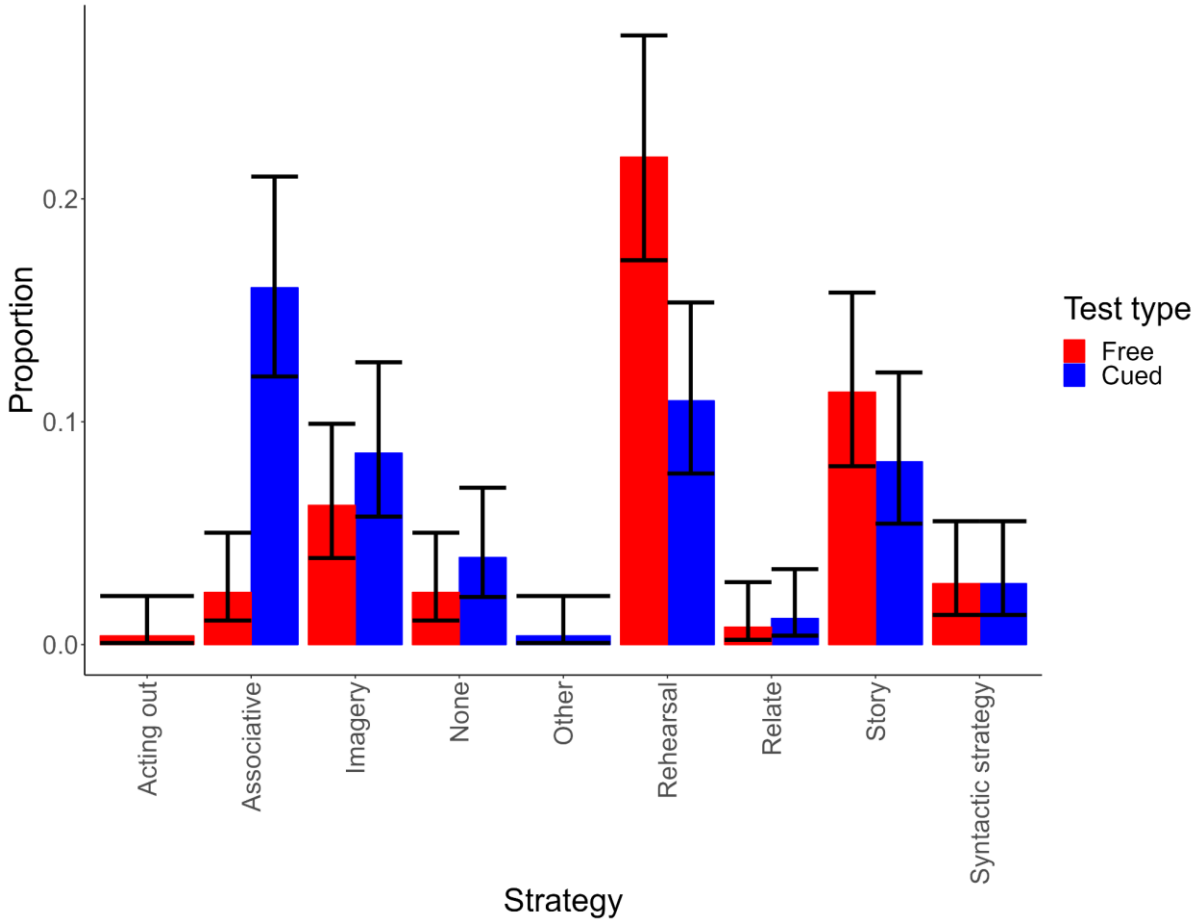
was significant in the FR -> CR group, $t(61) = 2.62$, $p = .01$. The bootstrapped CR:FR variance ratio in the CR -> FR group was 1.19 (95% percentile bootstrap CI [.97, 1.46]), and in the FR -> CR group it was 1.32 (95% percentile bootstrap CI [1.09, 1.58]). Accuracy by test type and test order is shown 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.

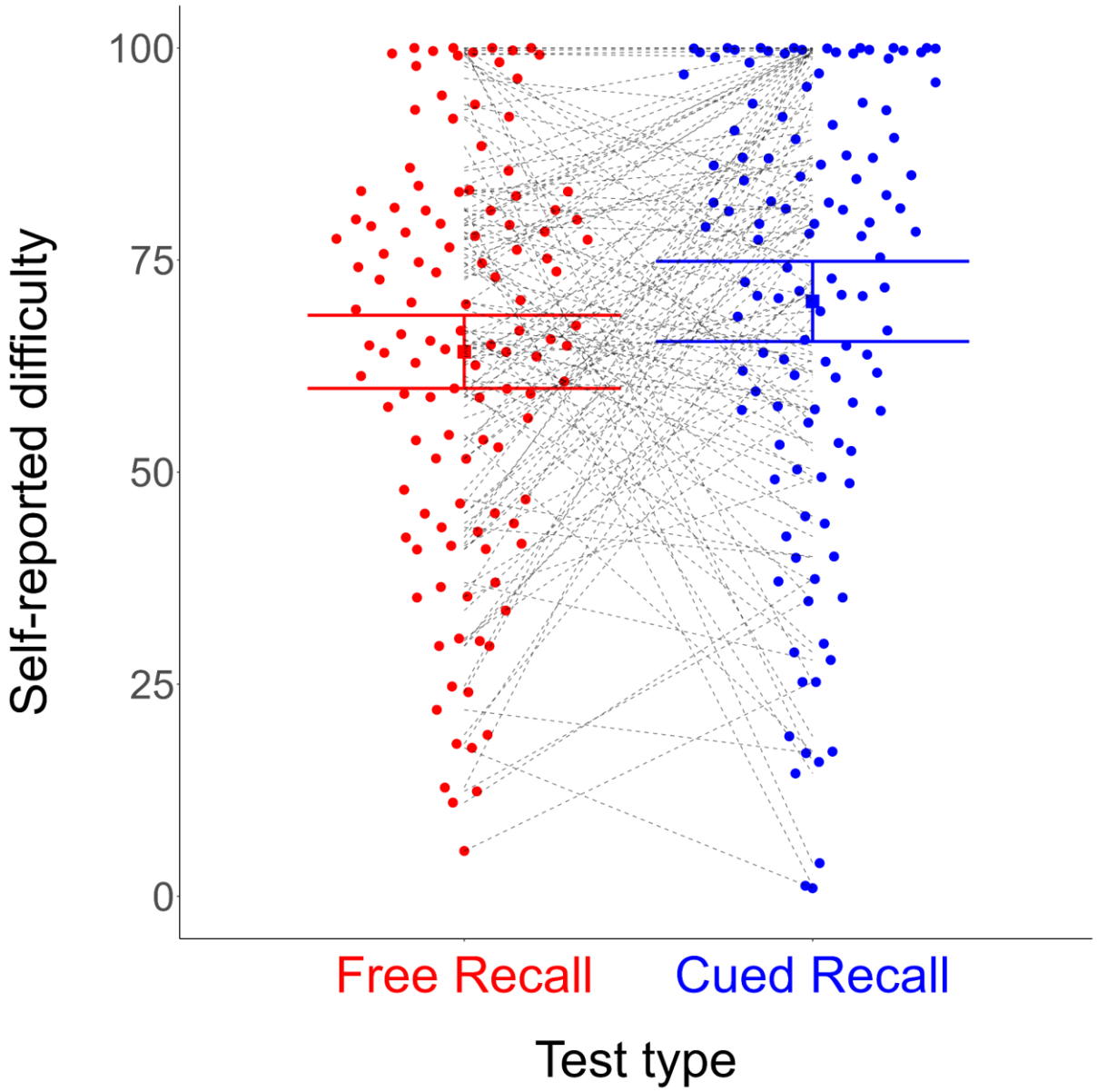


Note. Error bars = proportion 95% CIs (using Wilson's (1927) method).

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:



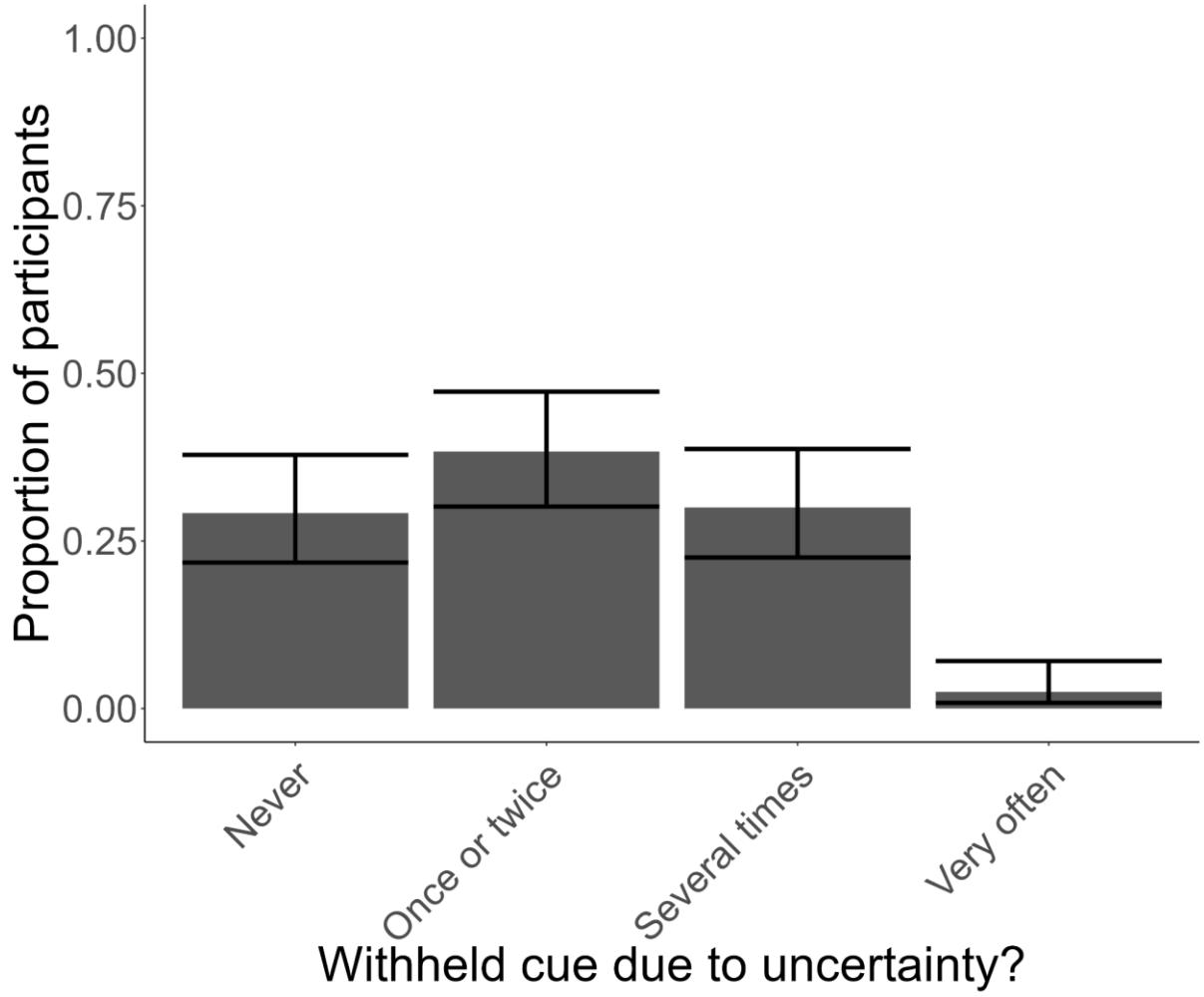
Note. Error bars = 95% CIs (within-subjects). Points jittered horizontally based on relative frequency. Dashed lines connect FR and CR difficulty ratings for individual participants.

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, $\Delta LOO = .58$ (*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

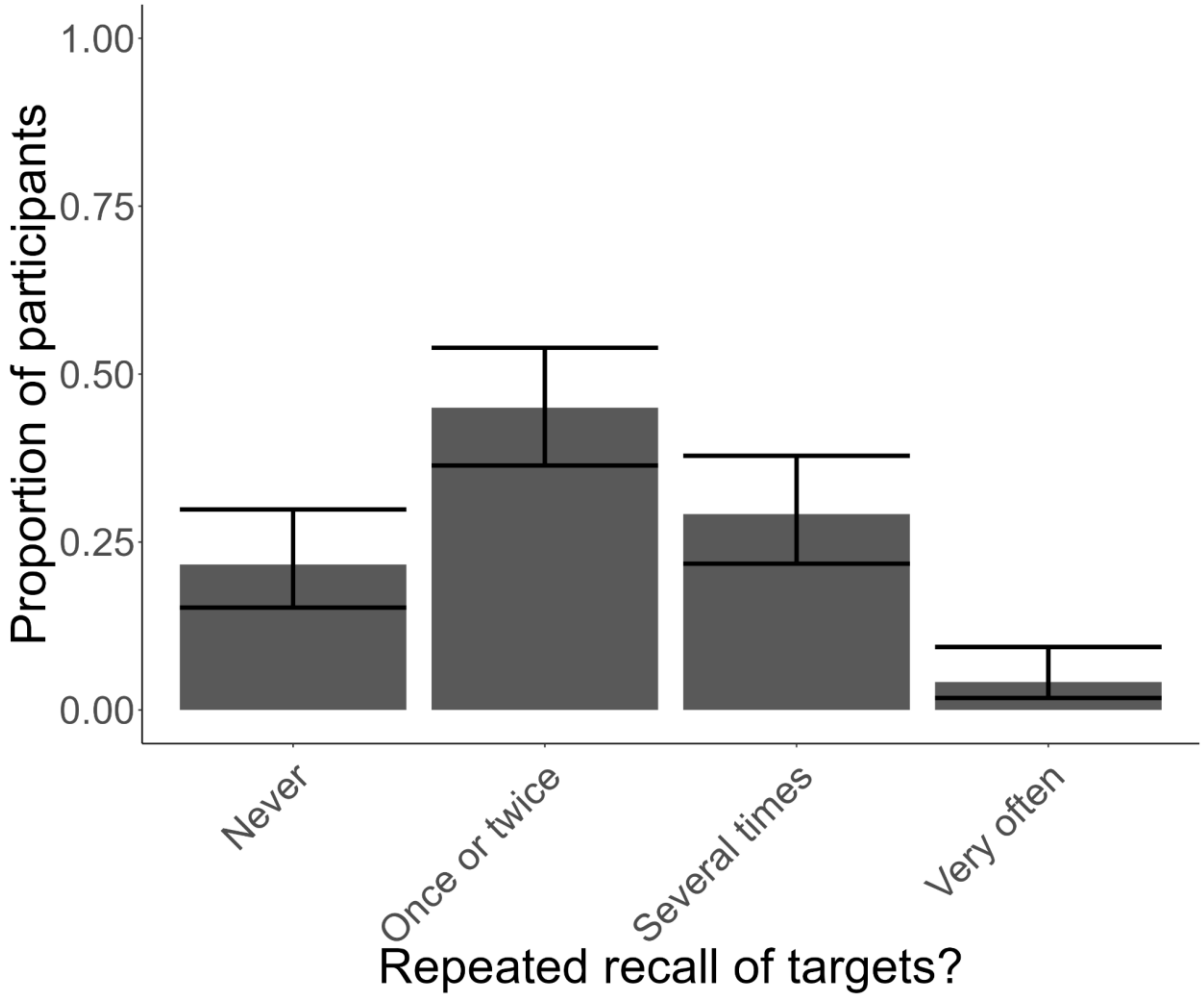
Participants were asked to self-report the general frequency with which they withheld a CR target that they thought of because they were unsure of whether it was paired with the current cue. Possible responses included: *Never*, *Once or twice*, *Several times*, and *Very often*. The figure below shows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

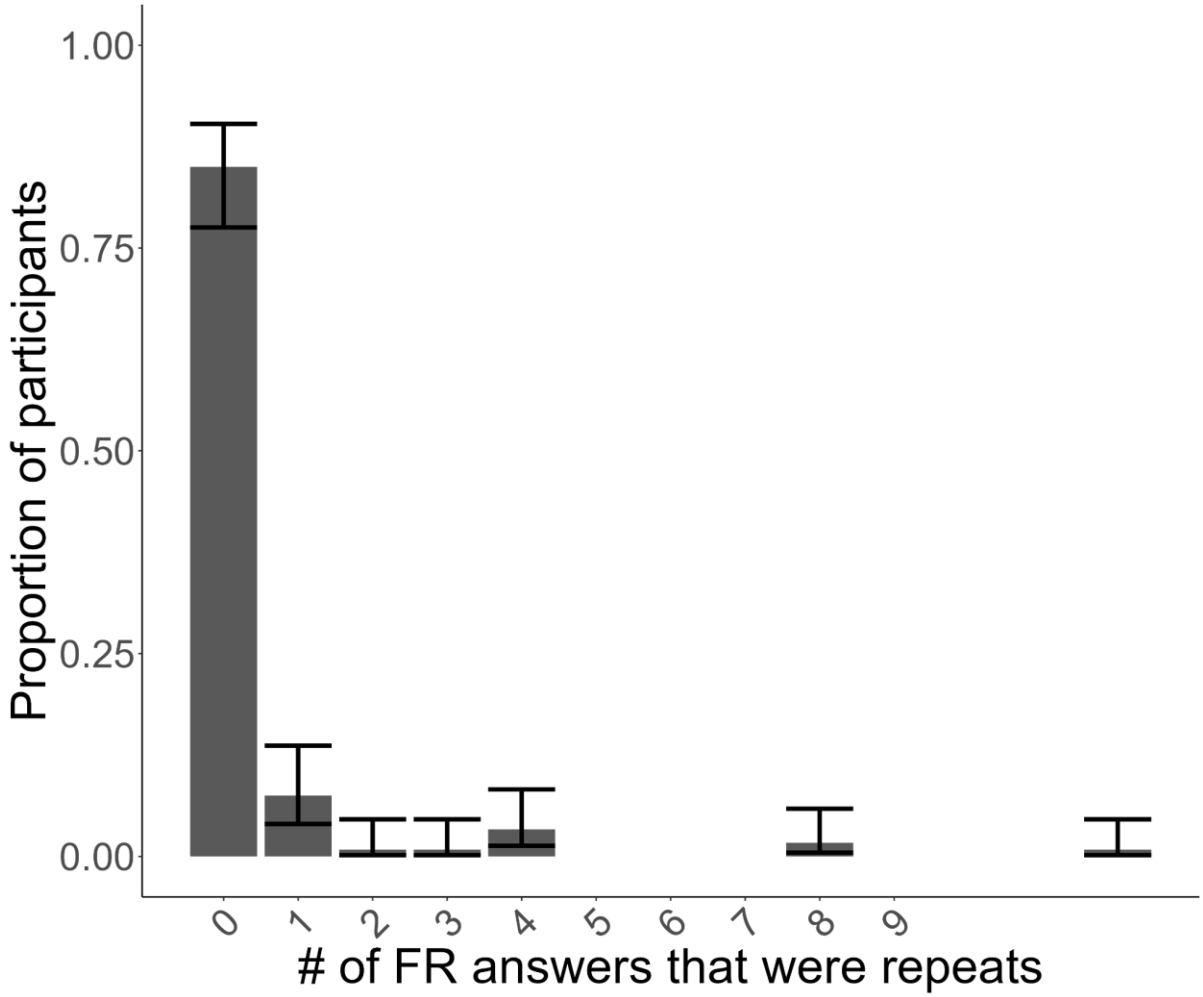


Note. Error bars = 95% CIs on the proportions (Wilson method)

H. Repeats in recall

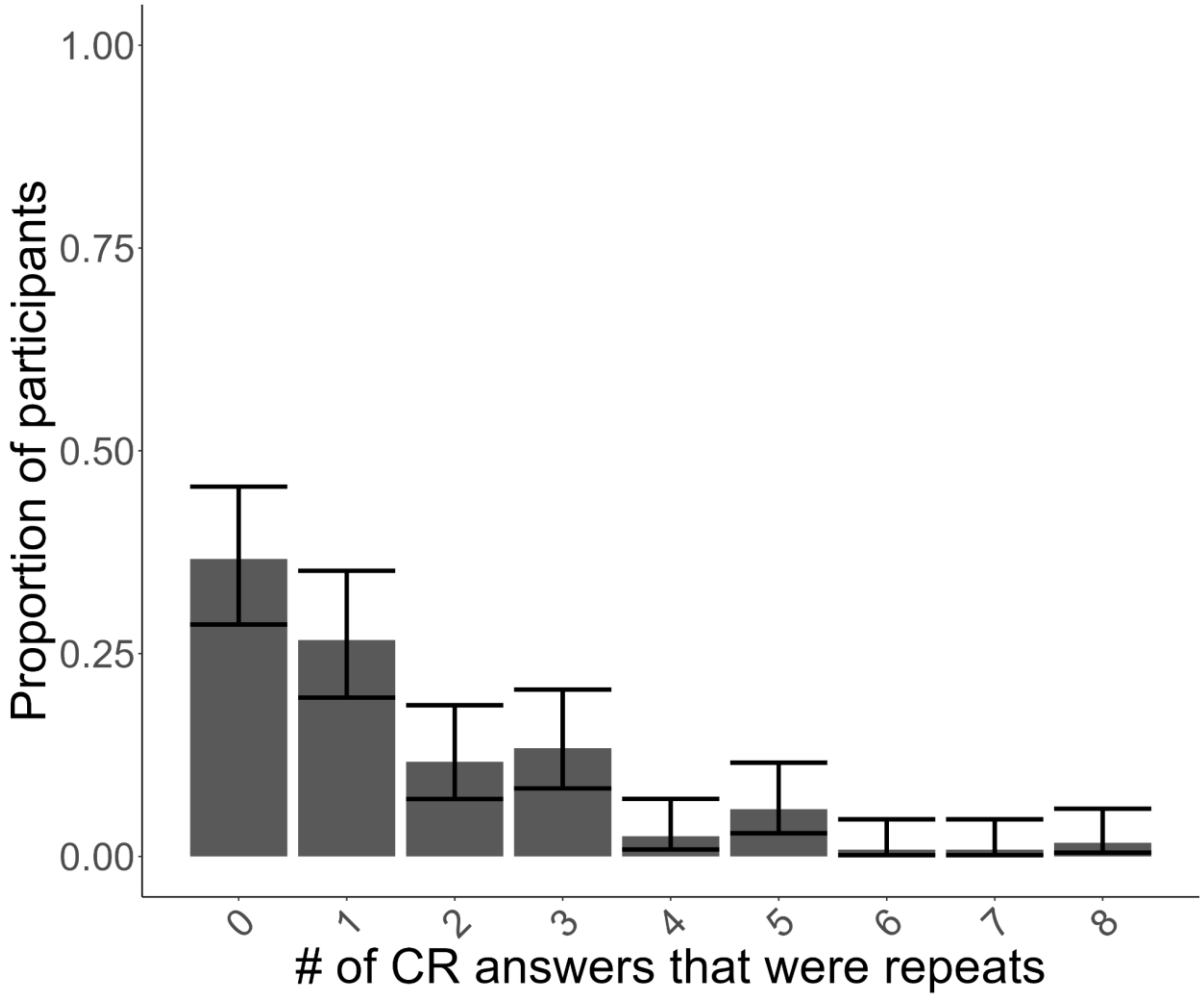
a. Free recall

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:



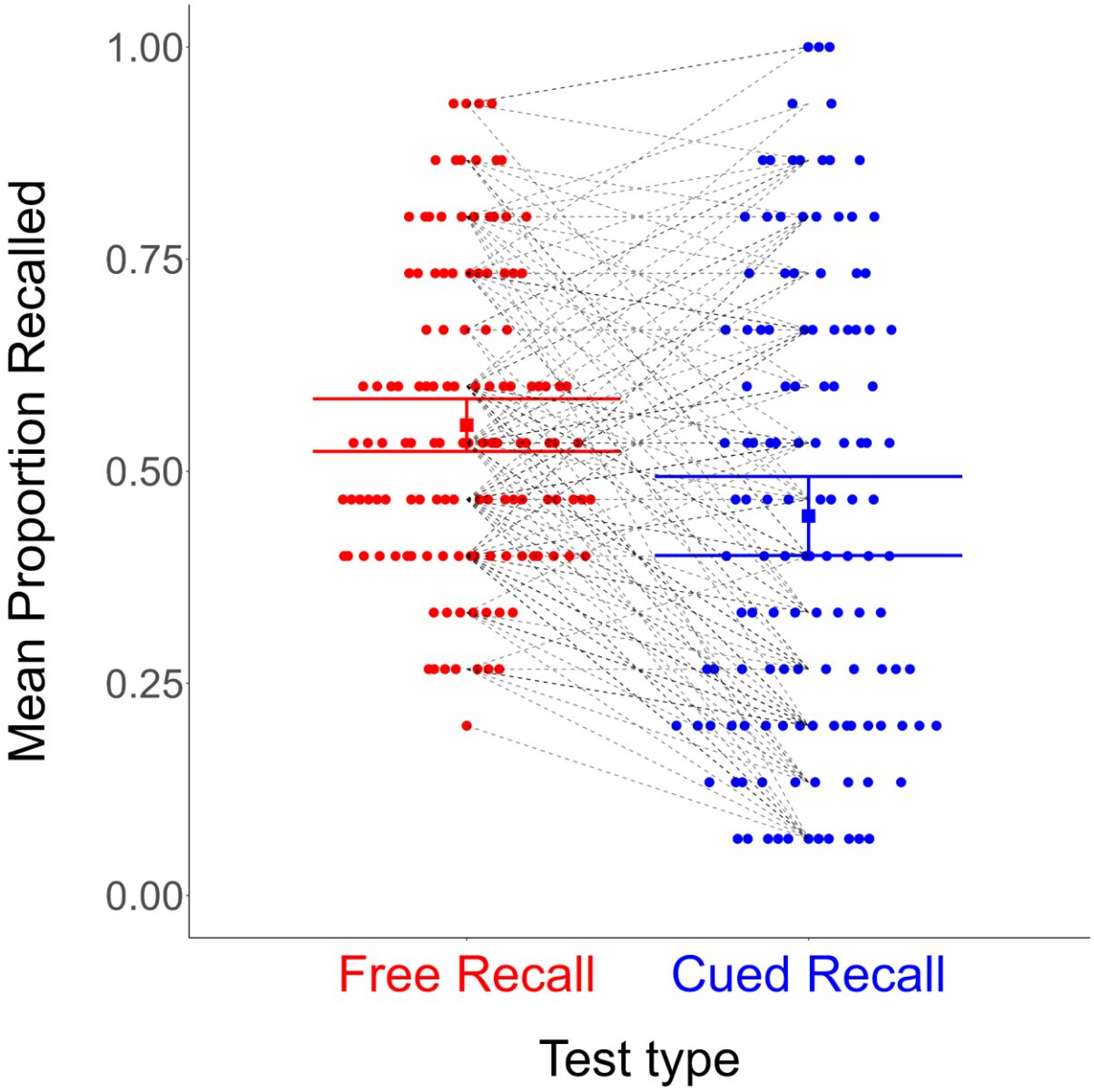
The majority of repeats (71.2%, 95% CI [64.8%, 77.7%]) were of studied targets.

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]).

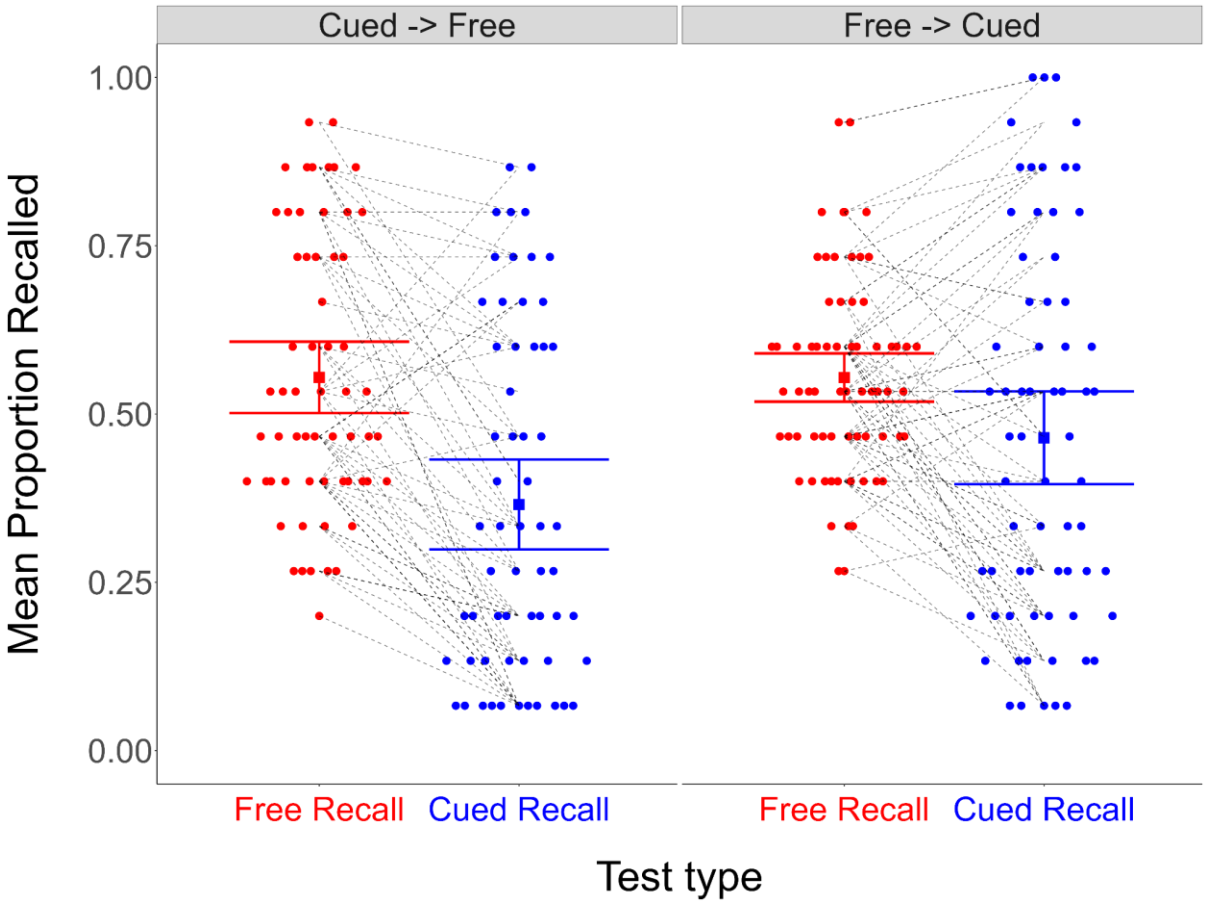
C. Generalized mixed-effects logistic regression results. We used the same GLMM model as in Experiment 1 to compare FR and CR variability (for standard CR proportion recalled and proportion recalled when same-list commission errors were counted as correct):

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.

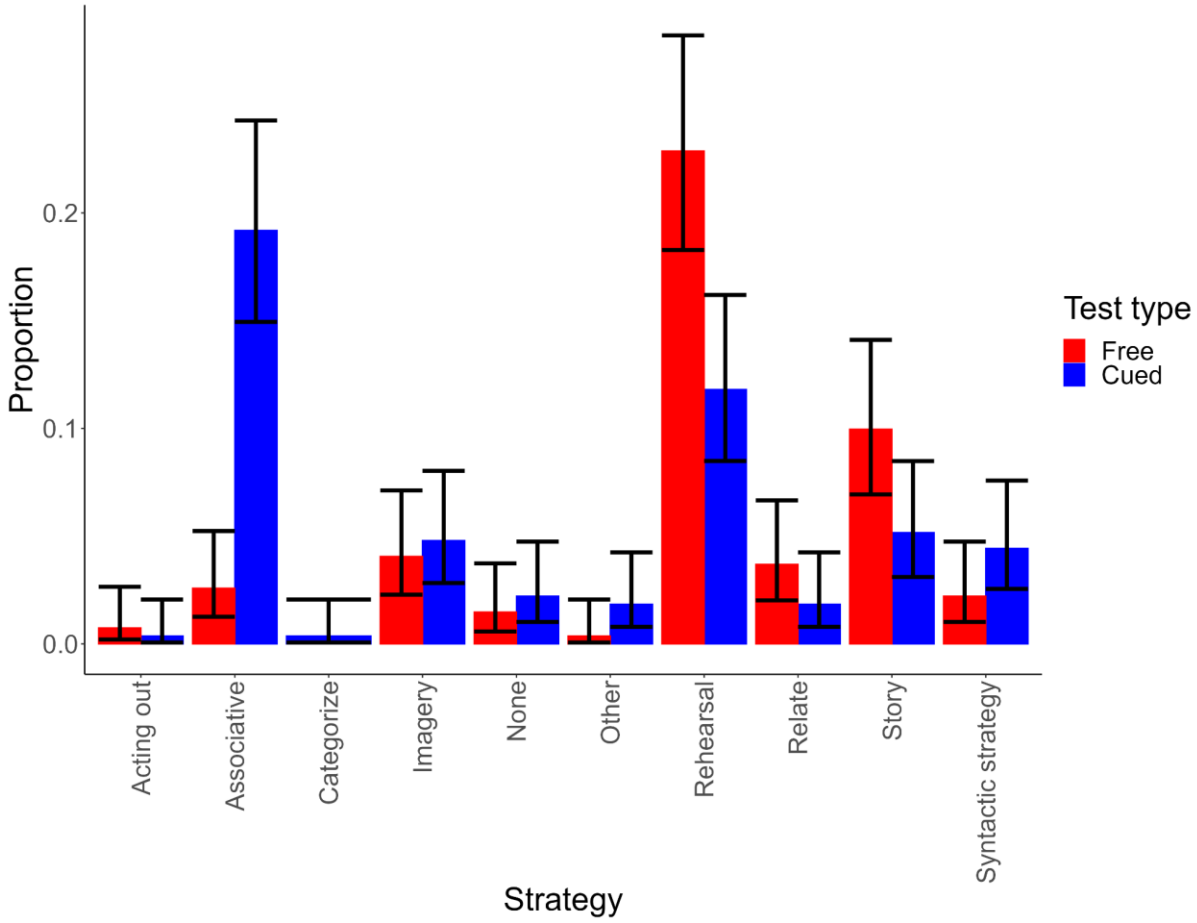
D. Order effects: 60 participants completed CR before FR, and 67 completed FR before CR. We examined the possible influence of task order on variability differences and overall accuracy. First, we conducted separate Pitman-Morgan tests in each order condition. The test was significant in the CR -> FR 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.

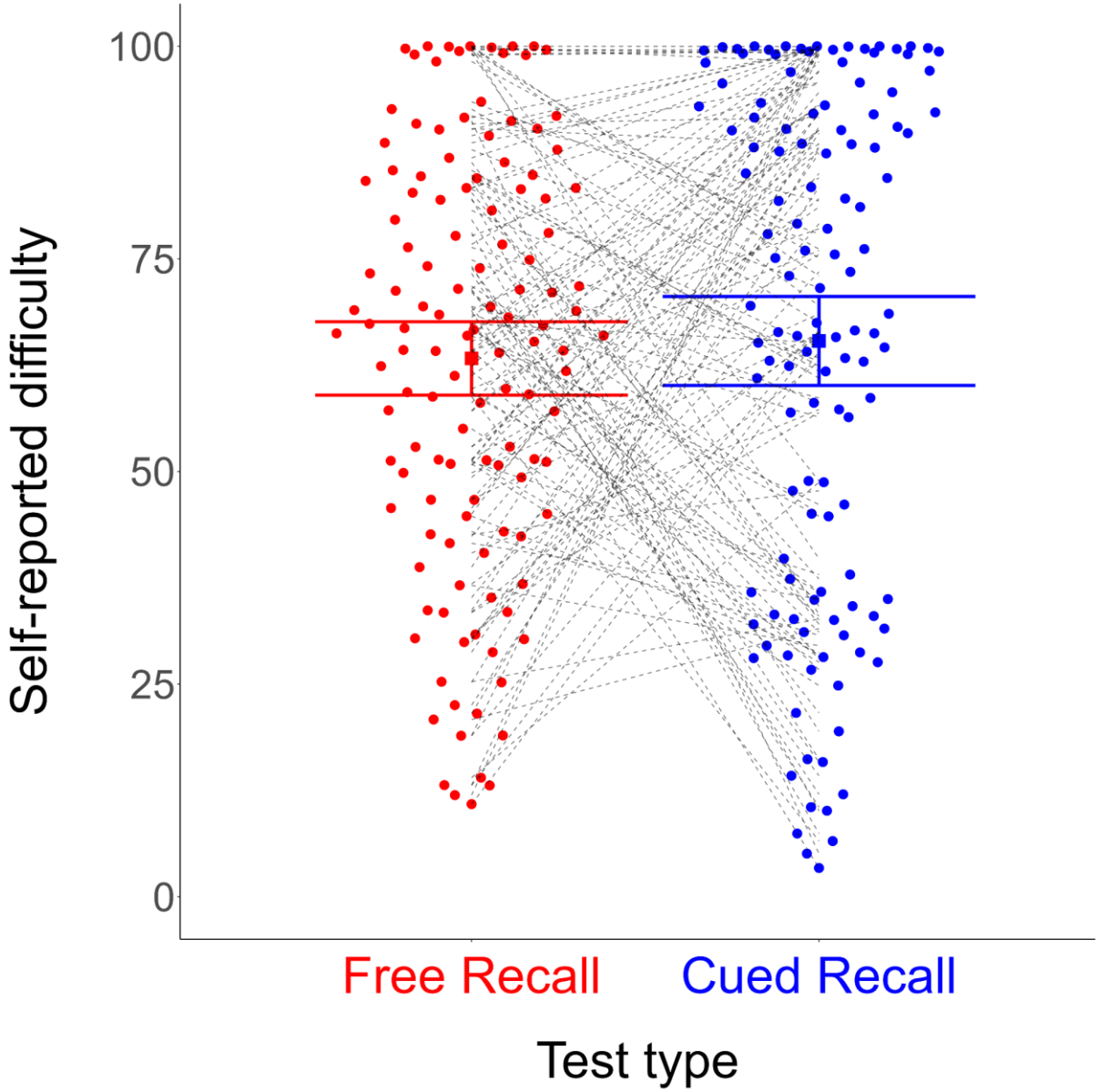
E. Self-reported study strategies. Of 414 coded qualitative strategy responses, the initial two coders agreed on 327. The remaining 87 responses were put to a 3rd coder.



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.

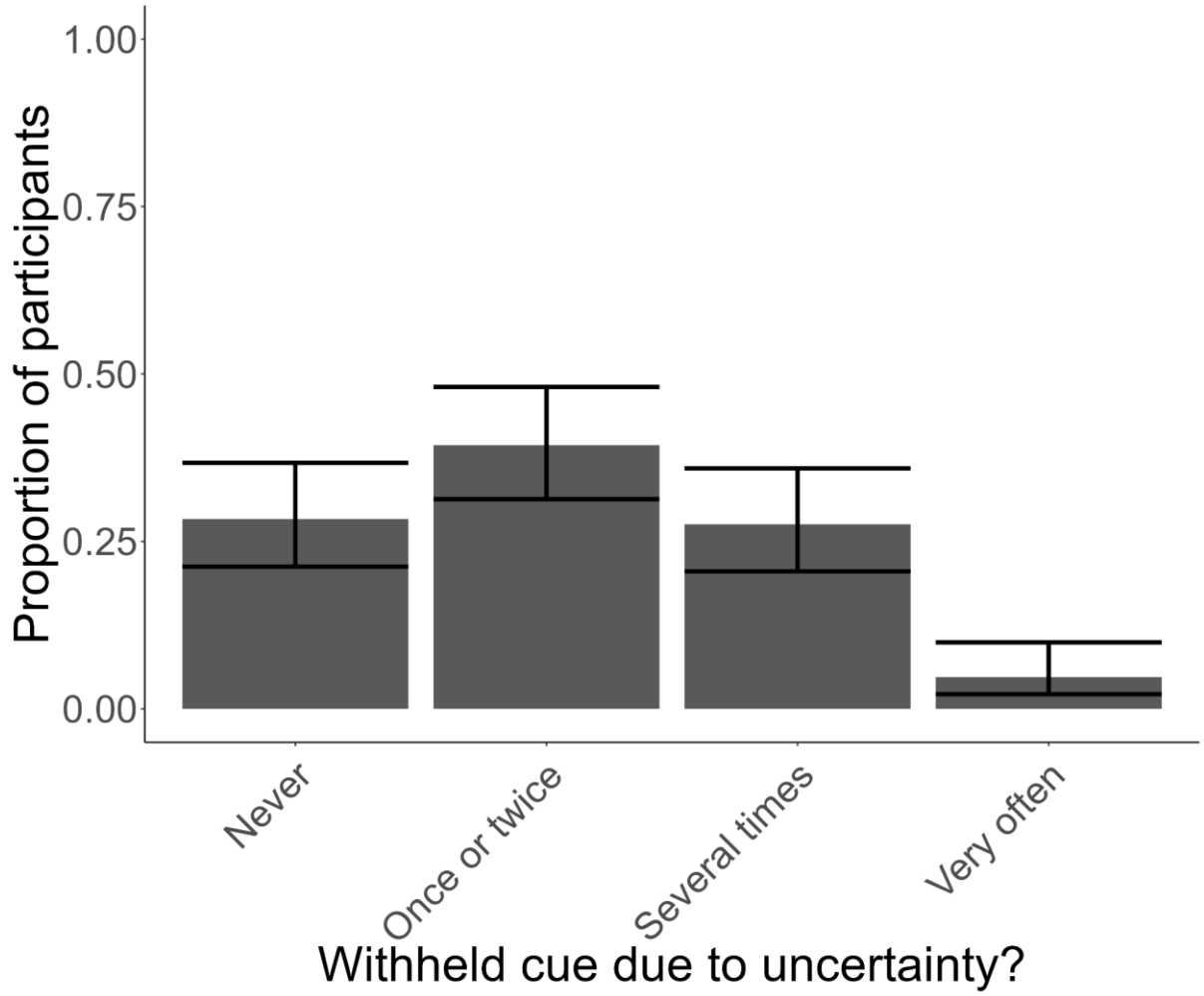
Here, the variability difference was significant, Pitman-Morgan $t(124) = 2.14$, $p = .03$, with a bootstrapped CR:FR variance ratio of 1.21 (95% percentile bootstrap CI [1.07, 1.38]). However, the Bayesian model comparison did not produce clear results; the model with *differing FR/CR variances* was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

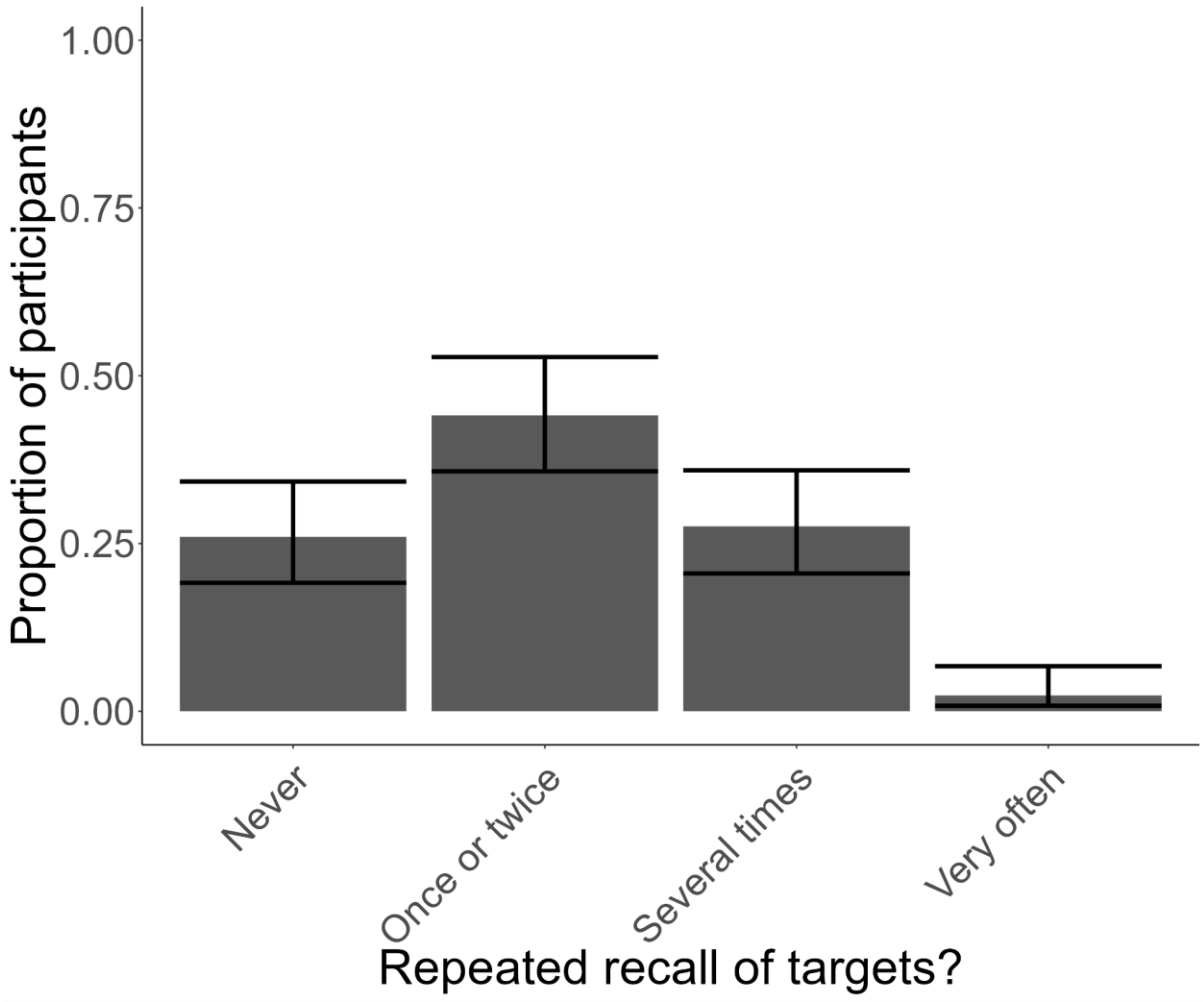
Participants were asked to self-report the general frequency with which they withheld a CR target that they thought of because they were unsure of whether it was paired with the current cue. Possible responses included: *Never*, *Once or twice*, *Several times*, and *Very often*. The figure below shows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



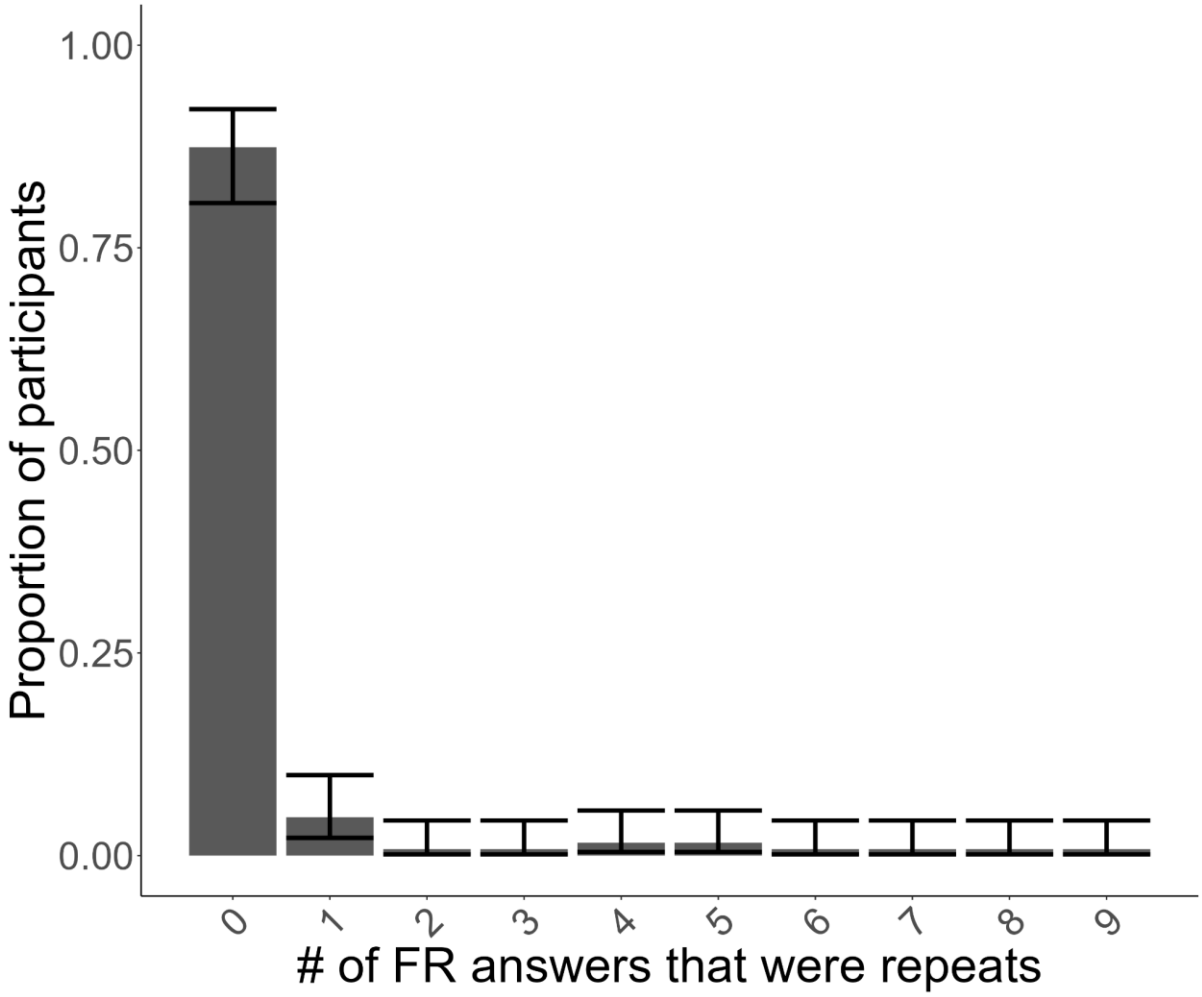
Note. Error bars = 95% CIs on the proportions (Wilson method)

H. Repeats in recall

a. Free recall

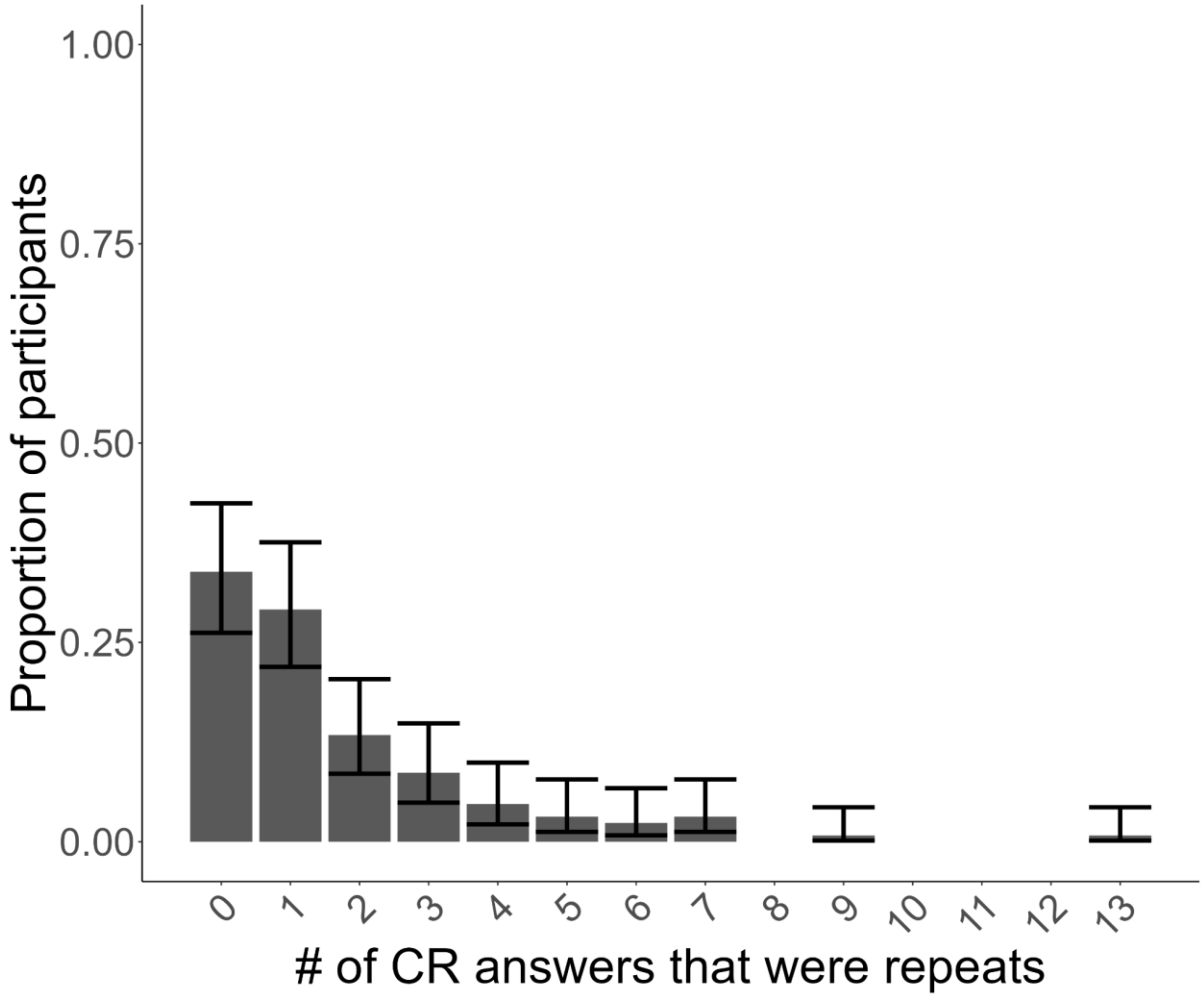
We examined the number of free recall responses that were repeats. The vast majority of

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 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 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,[18] 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.

[18] 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.

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. 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.

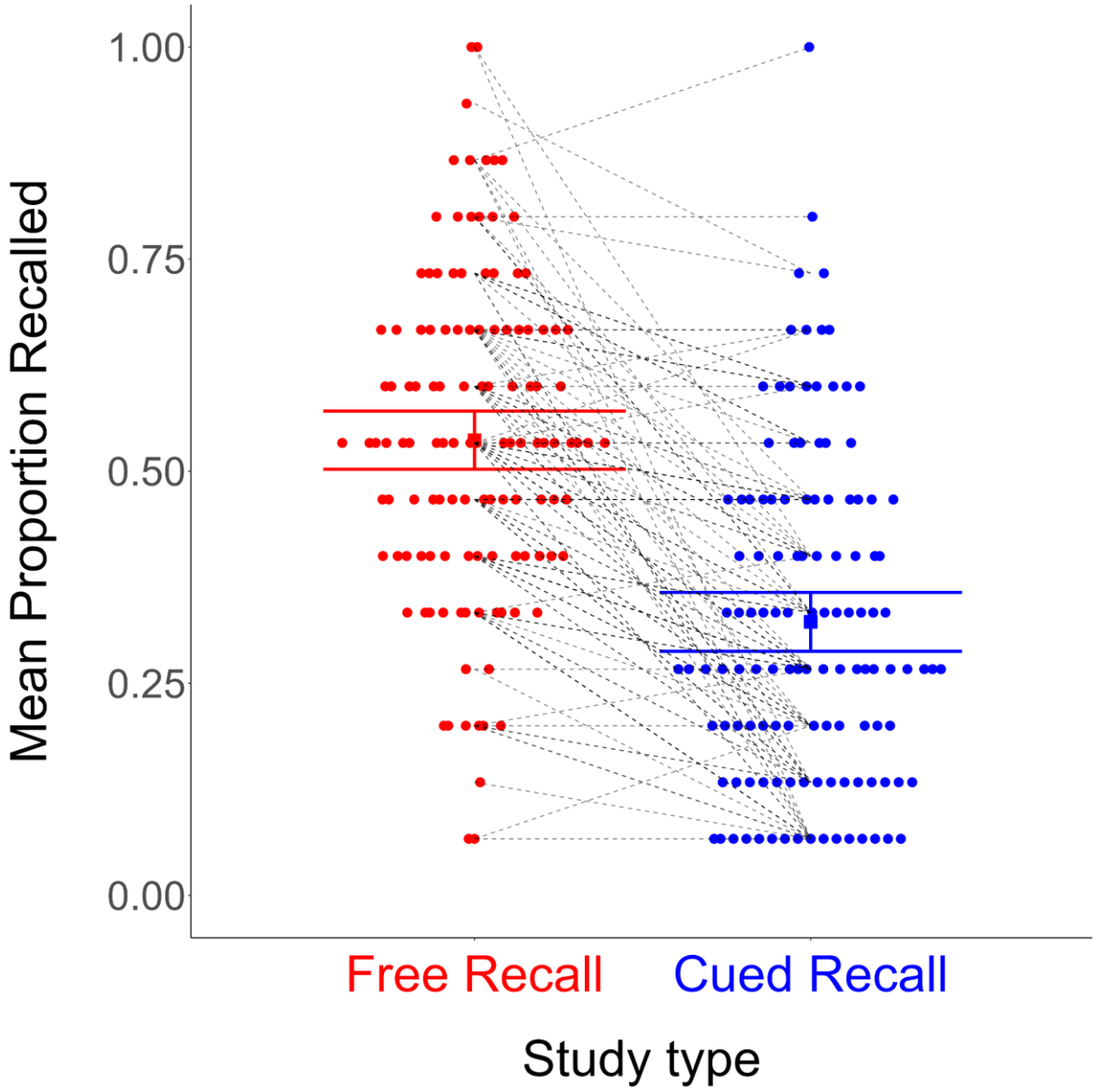
Results

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])[19]. The generalized mixed-effects logistic regressions also failed to provide 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 3) 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

[19] 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.

(*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: $\beta_{0i} = \gamma_{00} + \mu_{0i}$

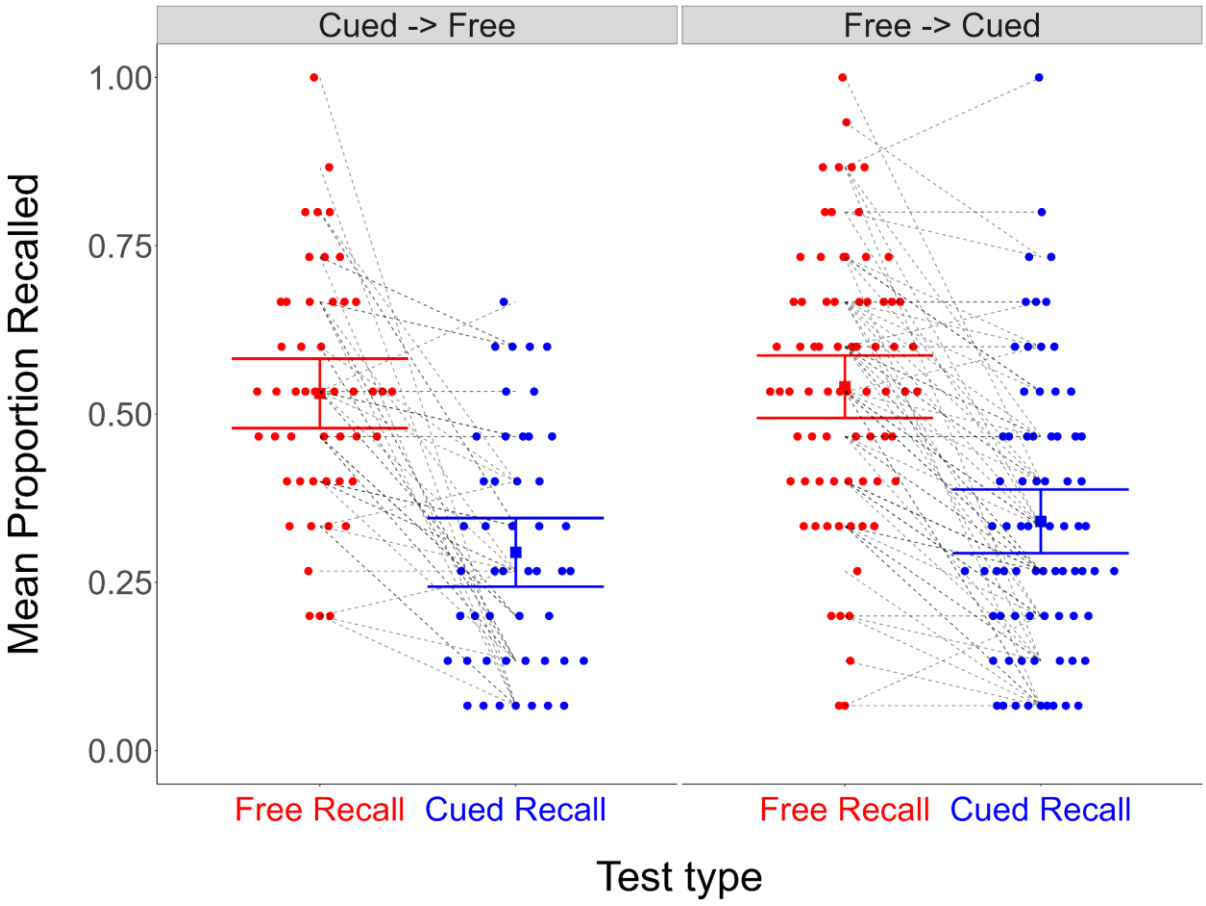
These models resulted in the following estimates:

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.

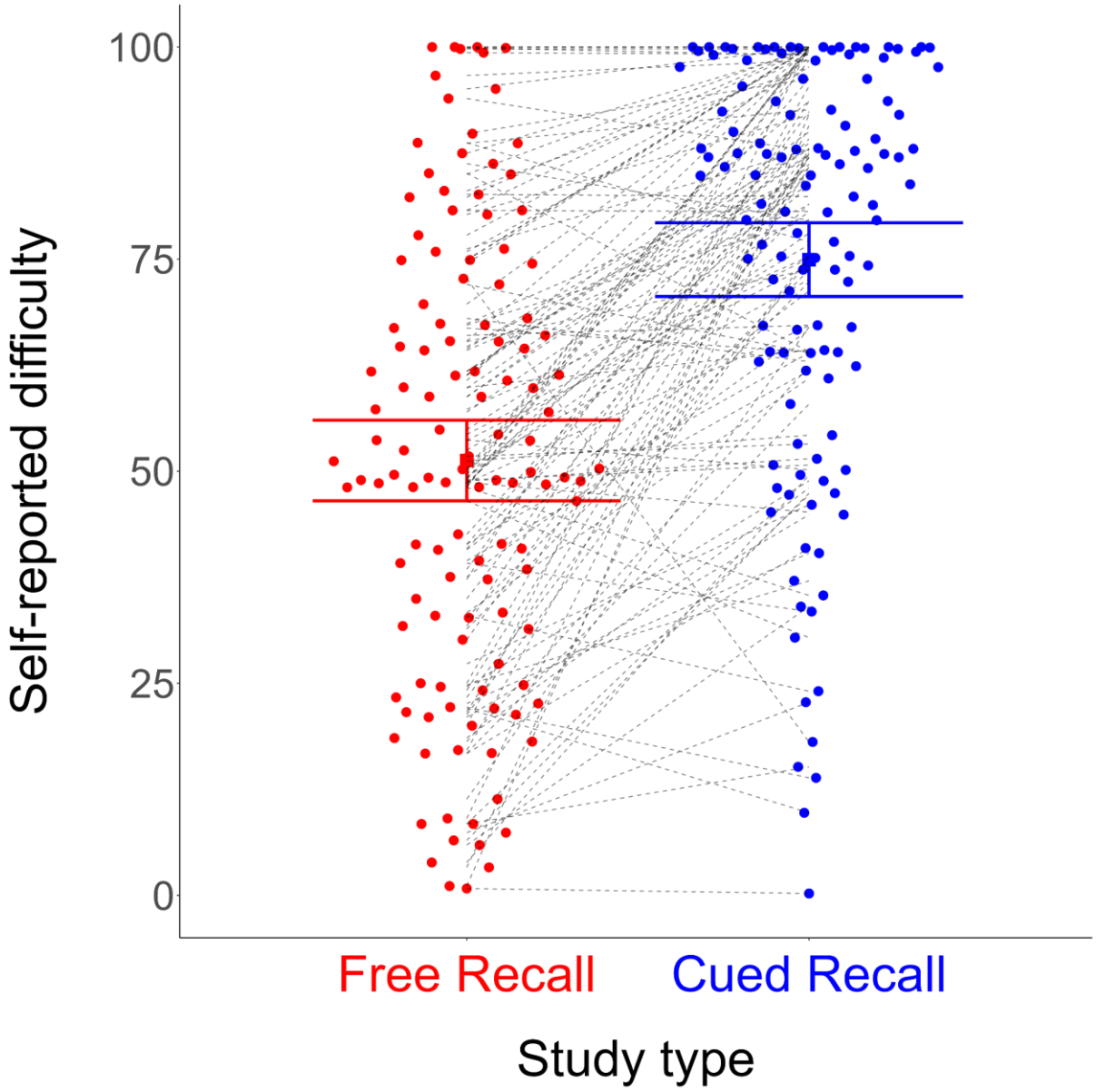
c. Order effects. 48 participants completed CR before FR, and 74 completed FR before

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:



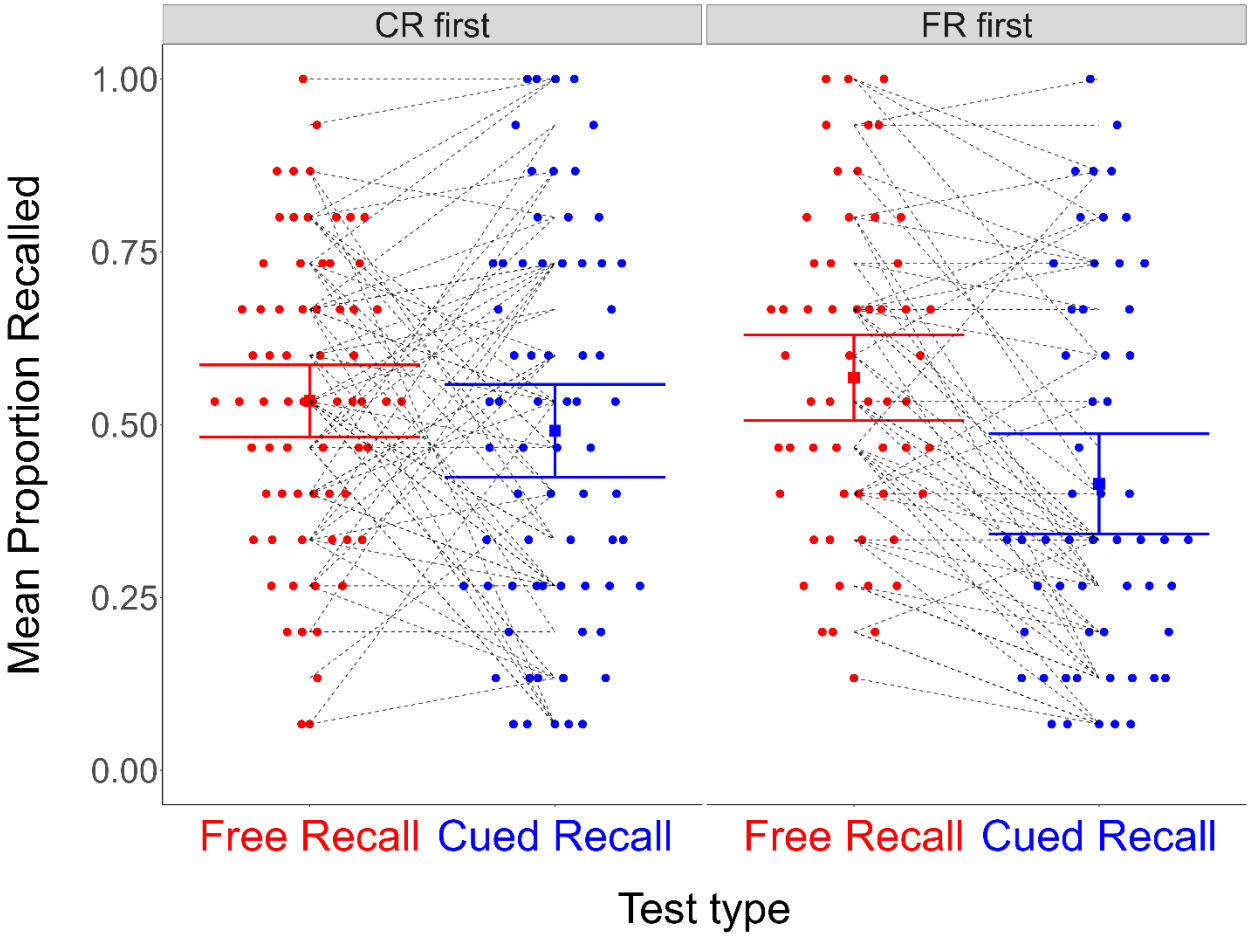
Note. Error bars = 95% CIs (within-subjects). Points jittered horizontally based on relative frequency. Dashed lines connect FR and CR difficulty ratings for individual participants.

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

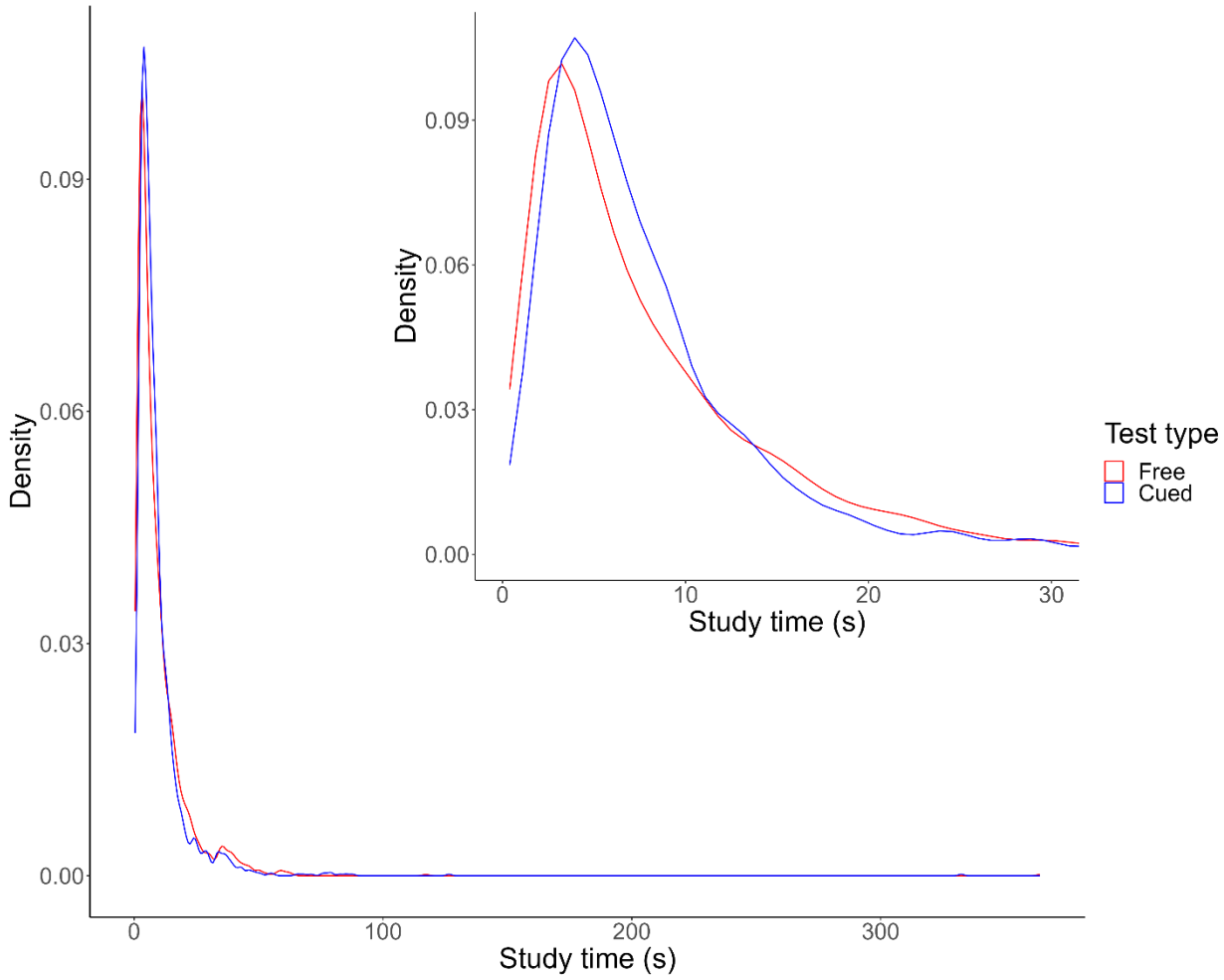
Bootstrapped CR:FR variance ratio = 1.29 [95% CI: 1.07, 1.56]

b. ***FR first***

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

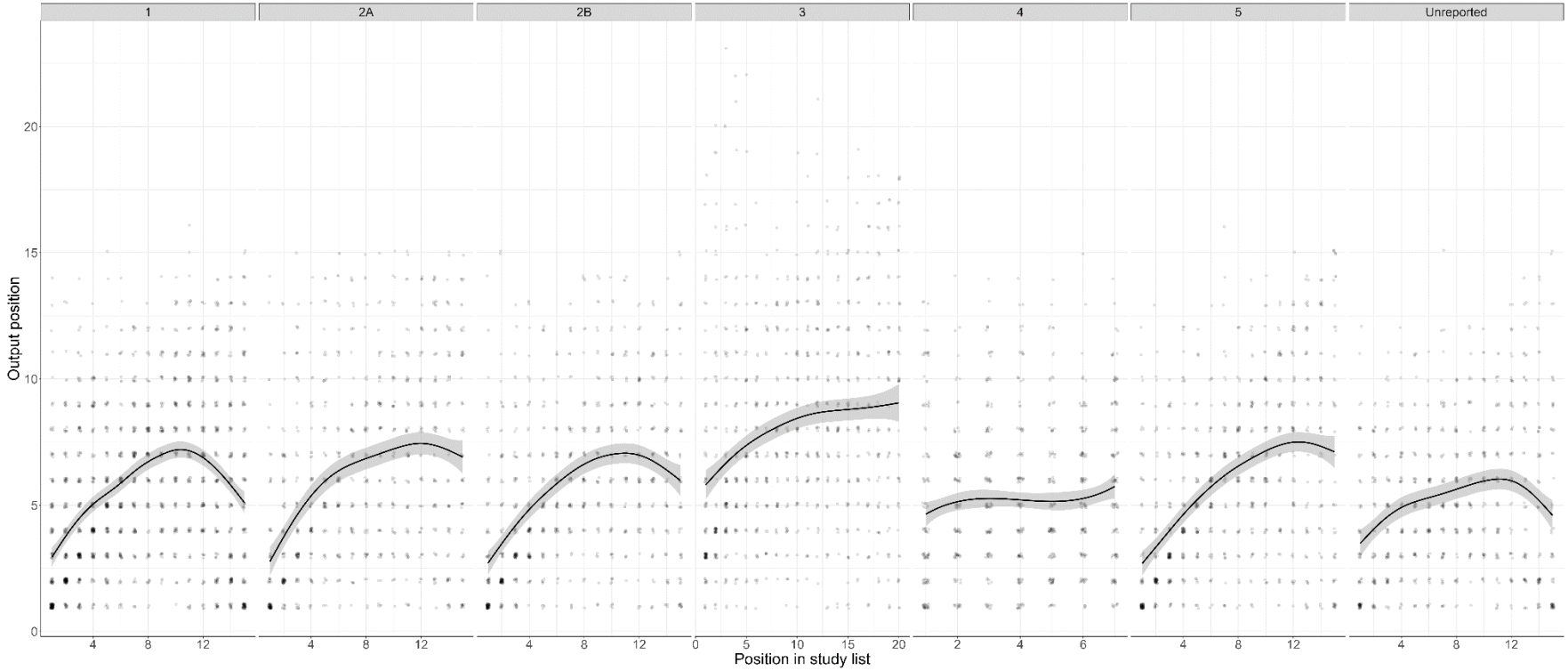
Bootstrapped CR:FR variance ratio = 1.18 [95% CI: .96, 1.43]

B. Study time



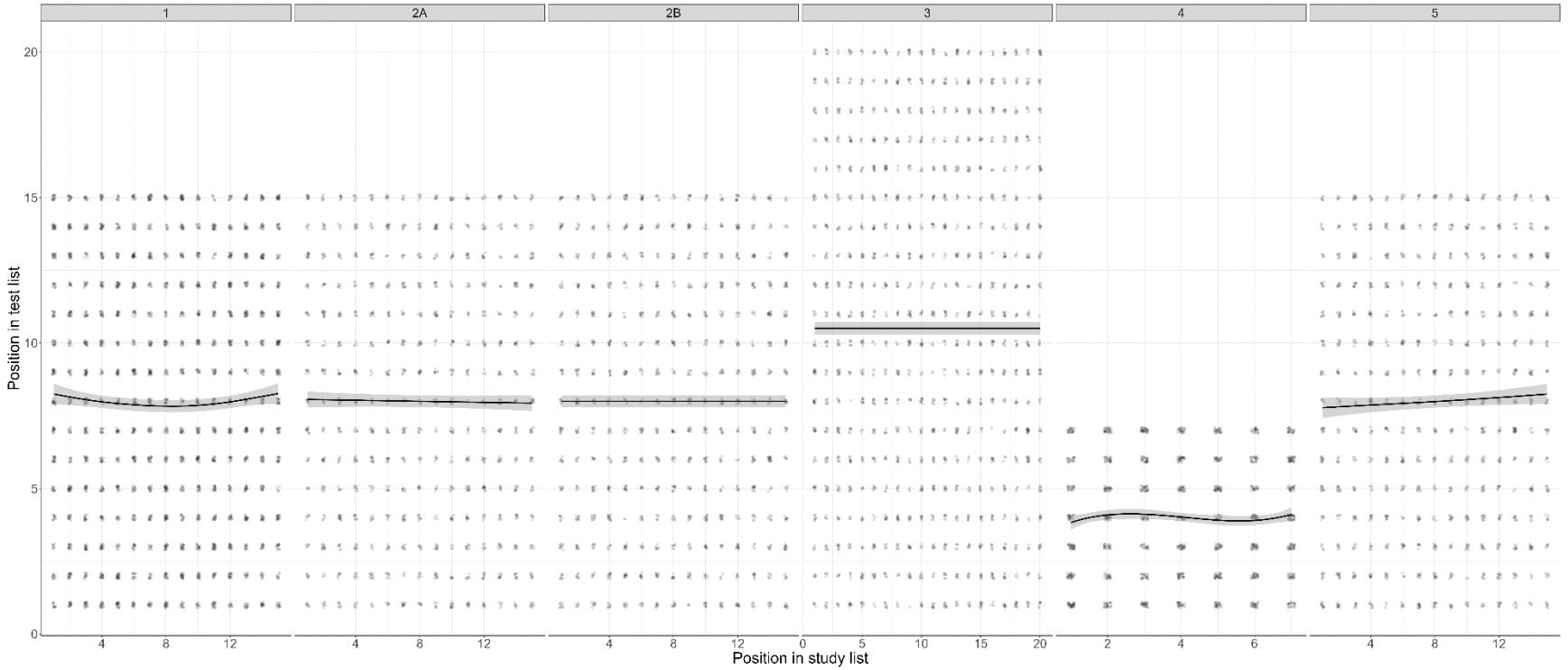
8. Output order

A. Free recall



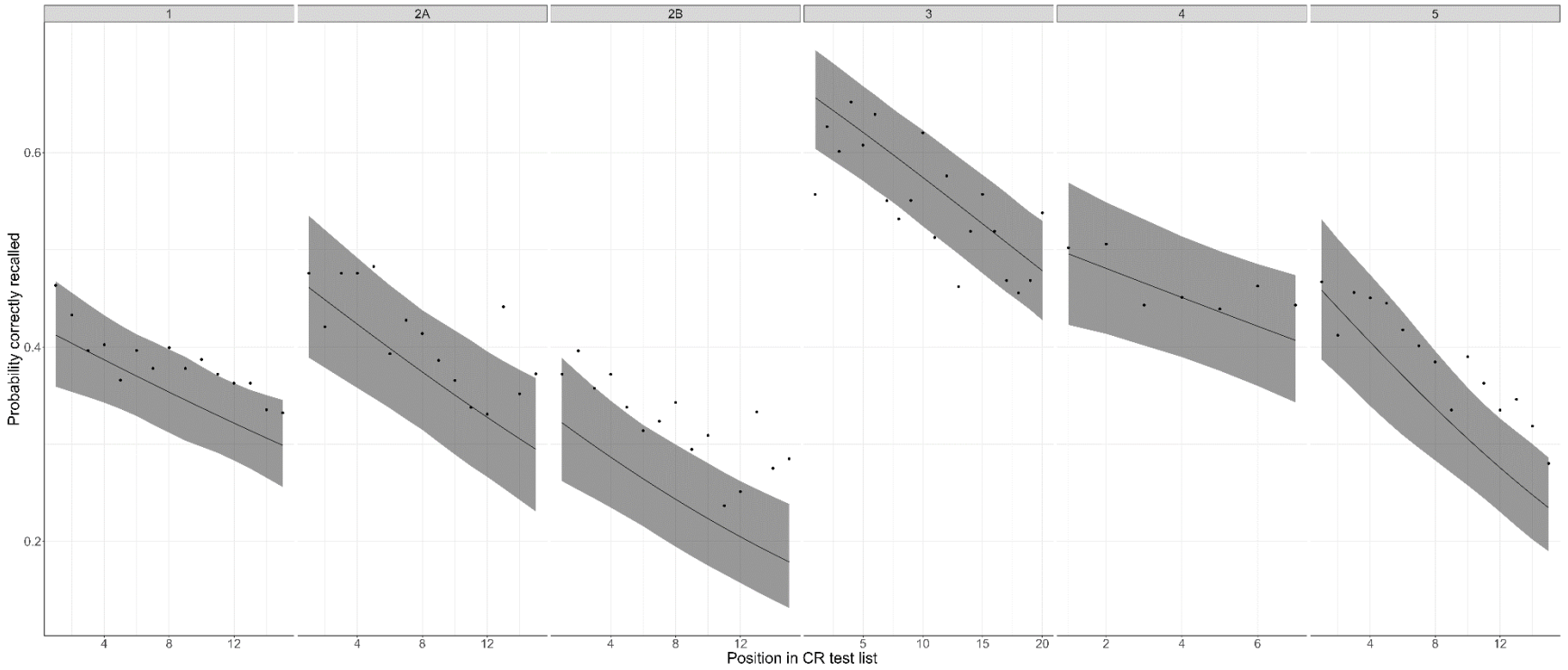
The figure above depicts (correct) output position for free recall as a function of position in study list. Jittered points represent individual words, and lines and ribbons represent regression lines (local polynomial regression fitting/LOESS) with 95% confidence intervals. The data here show a pattern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



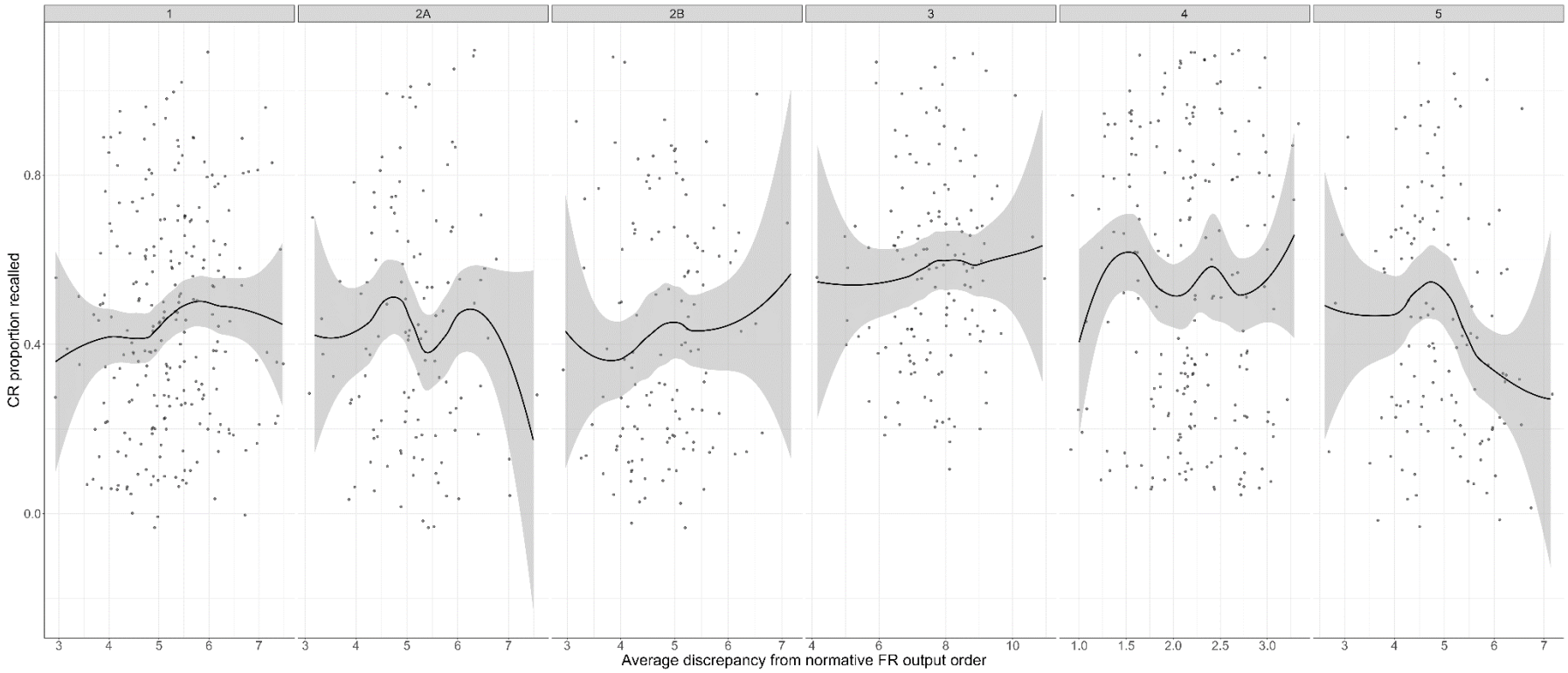
Similar figure for CR showing that, as expected, there was no relationship between position in the study list and likelihood of recall at a particular test position.

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).

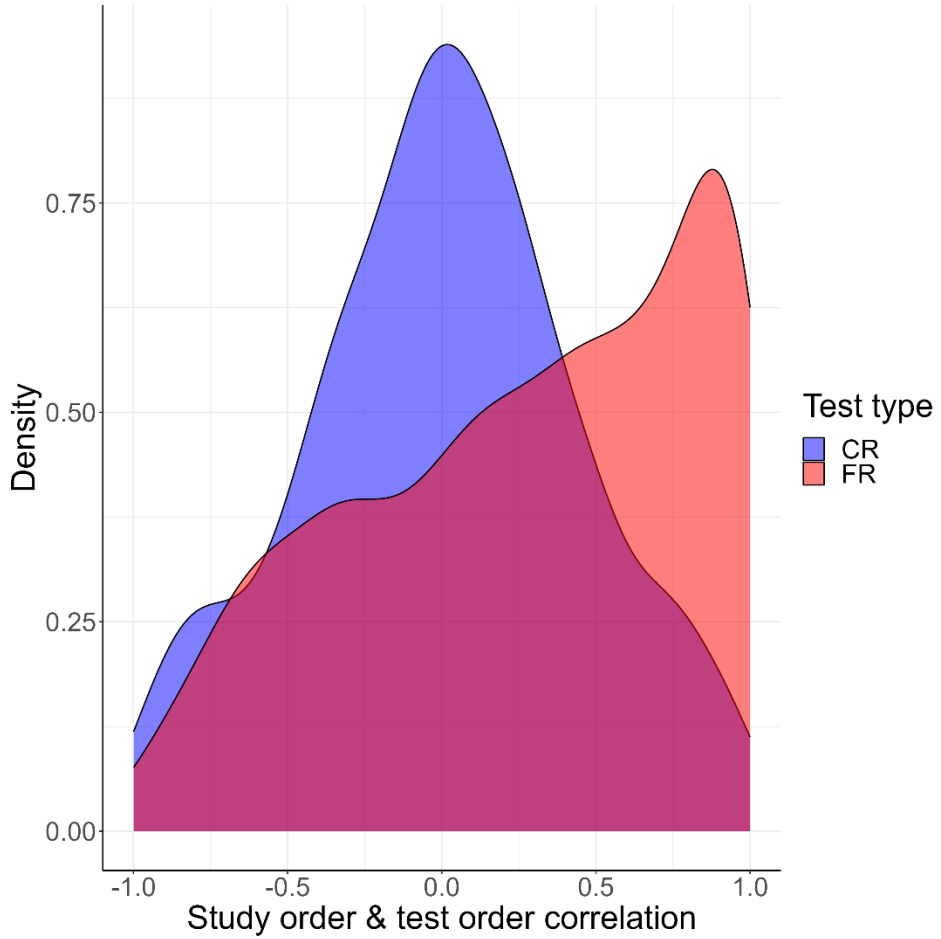
b. Concordance with 'normative' recall order



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.