Locating tip-of-the-tongue in lexical retrieval

by

Angus Ball

A Thesis submitted to the Faculty of Graduate and Postdoctoral Studies of

The University of Manitoba

In partial fulfilment of the requirements of the degree of

MASTER OF ARTS

Department of Psychology

University of Manitoba

Winnipeg

Abstract

Tip-of-the-tongue states (TOTs) commonly occur when a word cannot be retrieved, despite the conviction that it is known and will emerge imminently. Past research has probed access to target information during TOTs, but these tasks typically use definitions, common knowledge questions, or pictures to cue recall. This mimics naturalistic language use, but this semantic input creates challenges for observing the retrieval of semantic information. In Experiment 1a (N = 85), participants studied a list of uncommon words before completing a cued-recall task using vague clues to elicit TOTs thus limiting semantic input. When recall failed, participants rated their mental state before attempting to pick a semantic associate in a 4AFC task. In Experiment 1b (N = 88) recall was cued by presenting one item from studied word pairs, providing no direct information about the targets. The proportion of associate choice was significantly higher after reporting TOTs compared to forgetting in Experiment 1a but not 1b. The latter design was replicated in Experiments 2 (N = 90) and 3 (N = 86) using phonological and orthographic associate words, respectively. Finally, Experiment 4 (N = 86) included a semantic, phonological, and orthographic associate in the 4AFC task to place all three modalities in mutual competition. The results showed that TOTs benefited only phonology in isolation, but orthography was preferred when more lexical information was accessible. The experimental data provided a basis to evaluate a computational model of lexical retrieval, with the aim of simulating the observed recall and TOT phenomena. The model's lexicon included vector representations for the semantic, phonological, and orthographic relationships between 52,176 English words. Lexical retrieval required an initial semantic step which acted as a gate to the phonological and orthographic features of the word. Simulations of the empirical results using the same experimental stimuli showed promise, but the model struggled to capture the recall and 4AFC

outcomes with a single set of parameters. Adaptations are proposed to better capture the progress of TOTs over time and the shift in attention from semantics towards phonology and orthography as retrieval unfolds.

Keywords: Lexical retrieval, tip-of-the-tongue, recall, metacognition

Acknowledgements

I would like to thank my advisor, Dr. Randy Jamieson, for his mentorship and support. I would also like to thank my thesis committee, Dr. Launa Leboe-McGowan and Dr. Steven Greening, for sharing their perspectives, advice, and interest in this work. Thank you to Dr. Nick Reid for sharing his vector representations, and to Chelsea Capellán for assisting with data collection. Thank you as well to my former lab mate, Dr. Jackie Spear, for being an excellent mentor and role model, and for showing me the ropes in many ways.

I would also like to thank my parents, Joe and Caley, my sister Emma, my partner Simran, and my roommate Connell for all their encouragement and patience, and for listening whenever I needed to talk my way through the many ideas, decisions, and obstacles that arose along the way. Thank you to all my wonderful family and friends!

Table of Contents

Introduction

Tip-of-the-tongue states (TOTs) are highly contradictory experiences in which one feels certain they possess information yet are unable to report it. When people experience a TOT, they typically feel as though a breakthrough is imminent while also being vividly aware of the momentary lapse dragging on. Diary studies have shown that TOTs occur on at least a weekly basis, increasing with both age and stress (A. S. Brown, 2011, pp. 31-34). Yet, the vivid experience of a TOT does not feel like routine cognitive processing. It has become a convention within the TOT literature to quote the description of William James who wrote, “The state of our consciousness is peculiar. There is a gap therein; but no mere gap. It is a gap that is intensely active” (James, 1983, p. 152). TOTs sound much like feeling-of-knowing judgements (FOKs) on a surface level, and they are likely to be related phenomena, however TOTs are set apart as a singular experience - even described as “mild torment” (Brown & McNeill, 1966, p. 326). Research into TOTs has been growing since 1966, and the opportunity they provide to investigate the inner workings of memory has great potential to advance our understanding of memorial, linguistic, and metacognitive processes.

TOT Prospecting

The most prevalent approach to studying TOTs was established by Brown and McNeill (1966), who recognized the major shortcomings of naturalistic observations and diary studies. It was possible to take note of TOTs occurring in daily life and record guesses about target features or similar-feeling words that occur while searching for a target. However the rate of data collection was slow and the record of impressionistic observations proved difficult to quantify. Data were commonly lost when a TOT could not be immediately documented or the target word could not be identified, leaving the TOT unresolved. Without knowledge of the target, any

recorded data were difficult to interpret. Without control over the target words, multiple TOTs could not be observed for the same targets. This limited interpretation beyond identifying any matching features or correct guesses of target information, some of which would be expected to arise from random chance. To allow for deeper study, they introduced a method of *prospecting* for TOTs by reading dictionary definitions of uncommon words aloud. When TOTs arose, participants could be invited to report guesses about the number of syllables, first letter, words with similar sounds, and words with similar meaning to the target. TOTs that could be resolved were termed positive TOTs (+TOTs). These included cases where participants were able to independently report the word they were trying to remember, whether it was the target or not, as well as cases where they felt certain that the target was the word they had been trying to remember once it was revealed by the experimenter. Negative TOTs (-TOTs) were those which could not be resolved through recall or by recognizing the target, suggesting that individuals may be stranded farther from the brink of resolution. The key findings of this study offered empirical support to the common intuitions about TOTs, as participants had a significant chance of producing accurate guesses about their unrecalled targets. These guesses included identifying the correct number of syllables as well as producing similar sounding words with matching syllabic stress patterns and shared initial or final letters. This was taken as evidence for a form of generic recall where the recalled information is not of sufficient detail to disambiguate between possible items in memory, resulting in either abstract forms or partial details being recalled instead. This seminal study led to the direct-access view of TOTs, which states that TOTs arise when information about the target is retrieved in an incomplete or unclear manner. This view has been supported by many studies demonstrating access to semantic, phonological, and syntactic information (A. S. Brown, 2011, Chapter 5; see also Kumar et al., 2019).

Gollan and Brown (2006) explained positive and negative TOTs with consideration to leading two-stage models of speech production from the psycholinguistics literature (e.g., Levelt, 1992, 1993) in which retrieval is divided into a semantic stage and a phonological stage. The specific item being targeted for retrieval is identified in the semantic stage, and the information needed to produce that item is retrieved during the phonological stage. They noted that the object of retrieval can only be known in positive TOTs and successful recall, showing that the first step was completed at the time of the TOT. In the case of negative TOTs where the elusive target is never recalled or recognized, there is no indication that the first step was ever completed. Furthermore, separating TOTs into first and second step retrieval failures made it possible to observe differences between younger and older adults as well as monolinguals and bilinguals that were missed when both categories of TOTs were conflated. Another line of research showed that participants who repeatedly performed a prospecting task were more likely to experience repeated TOTs for the same items if they were not self-resolved on the first test, due to implicit learning of the ineffective retrieval state (D'Angelo & Humphreys, 2015; Warriner & Humphreys, 2008). The same phonological interlopers were also repeated when participants were instructed to think aloud, suggesting that these repeated TOTs occurred between the retrieval of semantic and phonological information (Oliver & Humphreys, 2019).

Paired Associate Learning, TOTs, and FOKs

Ryan et al. (1982) used a paired associate learning task to study the relationship between the strength of a memory trace and the retrieval effort during TOTs. The strength of memory traces was manipulated in the study phase, as each cue word was presented for 8 seconds and the target was presented at a delay to allow 3, 5, or 7 seconds of study time. They measured how TOTs, correct recalls, and Don't Know (DK) responses on the cued recall task impacted

performance on a concurrent task. Unlike recall rates, the probability of TOTs was unaffected by presentation time. This suggests that TOT is a complex phenomenon and not merely a middle ground state between recall and forgetting or an FOK judgement, as those explanations would predict an impact of presentation time. TOT trials were associated with significantly diminished performance on the concurrent task in comparison to DK trials and practice trials, whereas both DK trials and correct recall trials were not significantly different from practice trials. This provides evidence that TOTs involve active and ongoing memorial processes that tax attention, with additional support from neuroimaging work showing that TOTs are distinguished from FOKs by activity in the anterior singular cortex, right dorsolateral prefrontal cortex, right inferior frontal cortex, and anterior prefrontal cortex (Maril et al., 2005).

Others have criticized this study for not doing enough to ensure that it produced valid TOTs, and not merely strong FOKs. Participants in this study indicated on a questionnaire that their phenomenological experience was identical to naturally occurring TOTs, but there was no validation for their TOT claims. Words must be retrieved based on an intended meaning in the classic prospecting method, so individuals experiencing TOTs invariably have some access to semantic information but this was not verified in the above experiment (Valentine et al., 1996). It may be that the new and artificial cue – target association is insufficient for target retrieval, setting this phenomenon aside from true TOTs where the cue is sufficient for retrieval but the difficulty appears to lie in accessing the target itself (A. S. Brown, 2011, p. 47). Despite these critiques, similar results have been obtained from other research using a picture-naming method to elicit TOTs (Schwartz, 2008). Retrieving words based on a novel associative connection is also a relatively common requirement in natural English use, because many words have multiple meanings (polysemy) which must be learned as they are encountered or as language evolves over

time. Given the situation, it is premature to reject the subjective report of participants when it is possible to test for access to semantic information during TOTs elicited in this manner.

Heuristic – Metacognitive Account

Partial access to target information has been well documented in TOTs and is commonly taken as support for the direct-access view, however some argue that TOTs are better explained as metacognitive states that arise from judgements about our knowledge. In this view, TOTs do not occur because of partial retrieval but rather due to heuristic cues used to judge whether retrieval is likely to succeed. If the context would typically be sufficient for retrieval or the individual has reason to believe they ought to be able to access the word, then they will experience TOTs when successful retrieval does not occur. Strong evidence in support of the heuristic account of TOTs comes from Metcalfe et al. (1993) who employed a paired associate learning task in which the strength of cue – target pairs were manipulated at study using the interference theory paradigm. In specific, the association strength for a cue – target pair (A-B) could be supported with a repeated presentation (A-B, A-B), weakly supported with synonym pairing (A-B', A-B), not strengthened at all (C-D, A-B), or weakened with an alternative pairing (A-D, A-B). Across a series of four experiments, TOT occurrences were consistently driven by a cue familiarity heuristic (i.e., the number of presentations of A) and not target memorability (i.e., the associative connection strength between A and B).

The heuristic account also provides an alternate explanation for negative TOTs; in some cases there may be a TOT for an unintended target which is difficult to identify, but in other cases there may not be an actual target and the judgement is due to the cue alone. Support for this explanation comes from illusory TOTs that can be induced in a control condition when participants attempt to answer unanswerable questions (e.g., identifying the currency of a

nonexistent country). Since there is no possible target to recall, these illusory TOTs can only arise from demand characteristics, cue familiarity, or retrieval of non-target information related to the cue (Schwartz, 1998). Huebert et al. (2023) showed that partial target information produced during TOTs was no more accurate than information produced on non-TOT trials. Participants produced more guesses about phonology during TOTs and were slightly more accurate only when guessing first letters of targets, but their improved guesses may be better explained by a tendency to guess high-frequency first letters more often during TOTs than non-TOTs. These findings suggest that the TOT experience has less to do with partial access to target information and more to do with a change in search strategies, like relying more on first-letter frequency bias to direct retrieval efforts. When recall unexpectedly fails, TOTs could encourage the production of more probable high-frequency guesses (not partially retrieved information) with which to extend retrieval efforts. In this case, the increased cognitive load of TOTs over FOKs (Ryan et al., 1982; Schwartz, 2008) can be explained by a shift in search strategies and redoubling of search efforts that is compelled by TOT judgements.

It is important to note that TOTs are not always illusory, so there is a need to reconcile the compelling evidence of partial access to target information with these findings that TOTs can occur without retrieving partial information or even without having a target to retrieve. One such synthesis involves applying the heuristic-metacognitive perspective to the phenomenological aspects of TOTs and the direct-access perspective to the retrieval processes themselves (Schwartz & Metcalfe, 2011). According to this theory, when an initial retrieval attempt fails, the retrieved information is assessed by a metacognitive monitor mechanism to determine if the information available is sufficient to indicate that recall is possible. Sub-threshold retrieval results in giving up, but surpassing the threshold triggers the TOT state along with repeated search attempts. By

taking the strongest elements of each viewpoint, this theory provides a compelling explanation of findings spanning the breadth of TOT research without neglecting any one area, though more work is needed to obtain a detailed understanding of both the cognitive and metacognitive processes at play as well as how they interact.

Current Project

The goal of this research project is to test a jointly memorial and metacognitive theory of TOTs, closely related to the framework postulated by Schwartz and Metcalfe (2011). A series of behavioural experiments are paired with a modelling investigation aiming to simulate TOT processing with consideration to two-stage theories of lexical access. These theories state that lexical access begins with retrieving semantic (and often syntactic) information which contributes to the retrieval of the phonological and/or orthographic form, although the connections between these representations and their exact nature is uncertain (Caramazza & Miozzo, 1997, 1998; Roelofs et al., 1998). Models of TOT states, on the other hand, tend to remain productively ambiguous as to the stages of retrieval and how different sources of information contribute to TOT phenomena, instead treating retrieved information as an integrated mass (e.g., Gollan & Brown, 2006; Schwartz & Metcalfe, 2011). There may be more to learn by testing for partial access to these modalities independently, or by placing them in competition with one another to determine which modalities are relied upon while attempting to resolve TOTs.

In most TOT research, whether TOTs are induced through prospecting, picture-naming, or general knowledge questions, detailed semantic information about the target tends to be provided to participants. An exception to this comes from paired-associate learning tasks, which have been criticized for failing to demonstrate partial retrieval of target information. Since the

early role of semantic information is so crucial to theories of lexical access, the primary critique of paired associate learning for inducing TOTs may be a strength of this method. If partial access to semantic information can be demonstrated for TOTs produced using this task, that would suggest that this method can effectively induce non-illusory TOTs. Furthermore, this would show correspondence to the lexical access literature by confirming that semantic information is accessed early enough in retrieval to be present at the time of a TOT. This form of cued recall may differ from day-to-day word retrieval as we do not typically rely on novel associative connections for word retrieval, but the demands of this task are not dissimilar to recalling newly learned words or acquiring new meanings for familiar words.

The aim of this series of experiments is to induce TOTs and probe access to a variety of lexical information. An initial experiment was conducted using a cued recall task which was effective for inducing TOTs, despite providing cues with reduced semantic information compared to prospecting and picture-naming methods. By selecting a semantic associate on a forced choice task, participants showed that they had access to semantic information about unrecalled target words during TOTs which they could not demonstrate during normal forgetting (Experiment 1a). I follow up that effort using a paired associate learning task in which I evaluate people's access to semantic (Experiment 1b), phonological (Experiment 2), and orthographic information (Experiment 3) using the same method. To gain a better understanding of lexical retrieval and processing during TOTs, I conducted a final experiment examining these types of lexical information together. I placed semantic, phonological, and orthographic information in competition to test how these three dimensions of word knowledge are prioritized in the information retrieved when people find themselves in a TOT state (Experiment 4).

The results of these experiments provide data about the retrieval of lexical information under controlled conditions, as well as providing evidence about their retrieval and how they contribute to triggering and resolving TOTs. This data is then used to evaluate a computational model that will attempt to instantiate a two-stage theory of lexical access and a jointly meta-cognitive and memorial theory of TOTs. The formal computational theory is outlined following the experimental work.

Methods

Experiment 1a: Semantic retrieval

To investigate the role of semantic information in lexical access, I conducted an online experiment to test another approach for inducing and probing TOT states while withholding the detailed semantic information that is provided to participants in studies that prospect for TOTs using definitions and pictures. To accomplish this, participants were invited to study lists of single words (e.g., decanter, linen, allocate) and recall them from a short clue containing only a vague hint of the target's meaning (e.g., holds drinks). I predicted that this would encourage a higher rate of positive TOTs (relative to negative TOTs) due to recent exposure to the target, as well as allow me to manipulate performance by adjusting the length of the study list, presentation time, and precision of the clues. Using this cued recall framework also provided several metrics useful in assessing whether participants were paying attention and engaging in the task such as their rates of correct recall and intrusions (both intralist and extralist). Beyond assessing whether the method would be successful in inducing TOTs, I also aimed to obtain further evidence to support established theories that semantic information is retrieved early in word retrieval (Levelt, 1992, 1993) and contributes to TOTs (e.g., Meyer & Bock, 1992; Schwartz & Metcalfe, 2011).

Beyond supporting previously established findings, the data also served as the basis for my initial computational modelling.

Method

Participants: One hundred and seven undergraduate students (27 male, M_{age} = 19.50 years, *SD* = 2.82) were recruited through the University of Manitoba SONA psychology participant pool. Participants completed the study online and received one research participation credit towards their introduction to psychology course.

Materials: The complete materials for this experiment can be found in Appendix A. There were two lists of six target words. Some targets were drawn from stimuli used by Kumar et al. (2019) and others were selected from the SUBTLEX_US database (Brysbaert & New, 2009) with the aim of identifying words that were moderately infrequent but common enough that participants would be aware of their meaning. Low-frequency words were desirable because they are more difficult to recall in pure list situations (MacLeod & Kampe, 1996). For each target, a brief clue was written to provide a hint towards its meaning with less specificity than a definition (e.g., "holds drinks" for the target word decanter). One semantic associate and three unrelated foil words were selected for each clue. These were then checked using the TASA semantic vectors from Günther et al. (2015) to show that, according to the vectors, target words had higher cosine similarity to associates (*M* = 0.25, *SD* = 0.24, range = 0.01 – 0.78) than foils (*M* = 0.02, *SD* = 0.09, range = -0.14 – 0.19). For the remainder of this document, any analysis of word similarities using the cosine similarity of their vector representations will simply be termed "vector analysis" and will make use of semantic vectors from LSA (Landauer & Dumais, 1997), phonological vectors from Pincelate (Parrish, 2017), and orthographic vectors from SERIOL2 (Reid et al., 2023; Whitney & Marton, 2013).

Procedure: After providing informed consent to participate in the study, participants completed two study/test cycles. Before the study phase, participants were instructed, "You will be presented with some words one by one to study for only a moment each. Pay attention because you will be asked to recall the words afterwards." During the study phase, a list of six targets were presented in lowercase 21pt Arial font one at a time. Each target word was displayed for 1250ms in a randomized order, with a fixation cross during the 1000ms interstimulus interval.

After the study phase, participants completed a cued-recall test for the studied targets. Before the test phase, participants were presented with the following instructions:

> In this next phase you will attempt to recall one of the words you just viewed based on a short clue. When you recall it, you will have a chance to type the word before proceeding to the next trial. When you do not recall, you will indicate whether you feel it is on the tip of your tongue or truly forgotten. You will then try to choose the most similar-feeling word from a short list.

The test phase involved presenting one clue to prompt recall of a target from the studied list (e.g., "holds drinks" was the clue to recall "decanter"). On each trial, the instructions stated, "If you can recall which word from the list matches the following clue, type it below. If you cannot remember the word, click continue." Clues were presented in bold 19pt Arial font above a 6cm text box. Participants could type any word they recalled into the text box before clicking continue. If a word was reported, clicking continue or pressing the enter key would proceed to the next recall trial. If they did not recall a target, participants indicated this by clicking continue or pressing enter without typing a word in the text box. If no word was reported, they were invited to indicate the degree to which they felt they had fully forgotten the target or were experiencing a TOT. To collect those ratings, the clue remained visible on the page, and below it

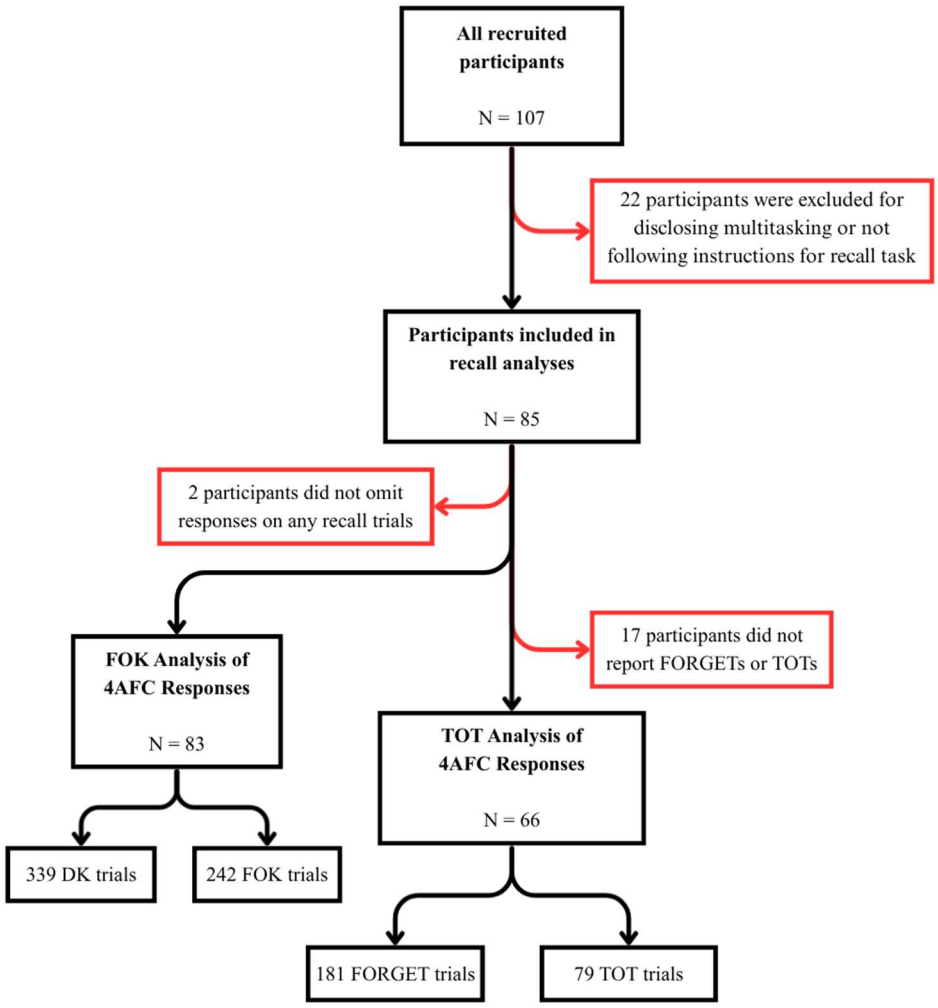
there was a 16cm horizontal sliding scale. The left end was labelled “I do not know it at all,” the right end was labelled “I am certain I know it, but it is on the tip of my tongue,” and the sliding indicator recorded responses on a hidden 100-point scale. Participants were required to move the indicator away from the exact center of the scale to proceed, placing the slider along the line to indicate their judgement and clicking the continue button to record that judgement. After providing their metacognitive rating, participants were presented with four unstudied words including a semantic associate (e.g., “bottle”) and three other unrelated words. They clicked on one word amongst the four alternatives which they believed best matched the target they were unable to report. These words were presented simultaneously in a randomized order, written in lowercase 21pt Arial font. After choosing one of the words, they proceeded to the next cued-recall trial. Once all six clues were presented, the study and test phases were repeated a second time with another list of targets and clues. After completing both study and test phases, participants were thanked for their participation and debriefed. All text was displayed in black on a white background.

Results

Of the 107 recruited participants, data from 22 were removed from the analysis for disclosing multitasking and distractions while participating or failing to follow the task instructions during the recall phase of the experiment (such as reporting nonwords, empty spaces, or typing sentences during the recall task). This left a sample of 85 participants for the following analyses. Figure 1 displays a flowchart of data processing, including the number of participants or trials contributing to each section of the analysis.

Figure 1

Data processing flowchart.



Note. From the full sample of 107 participants, 22 participants were excluded from the analysis for failing to follow instructions or disclosing multitasking and distractions while completing the experiment. Recall performance was analyzed for the remaining 85 participants. The FOK analysis included all 4AFC trials from participants who did not report a word on every recall trial, with 83 participants contributing 339 DKs and 242 FOKs. The TOT analysis included only trials where participants indicated a strong sense of forgetting or TOT state, with 66 participants contributing 181 FORGETs and 79 TOTs.

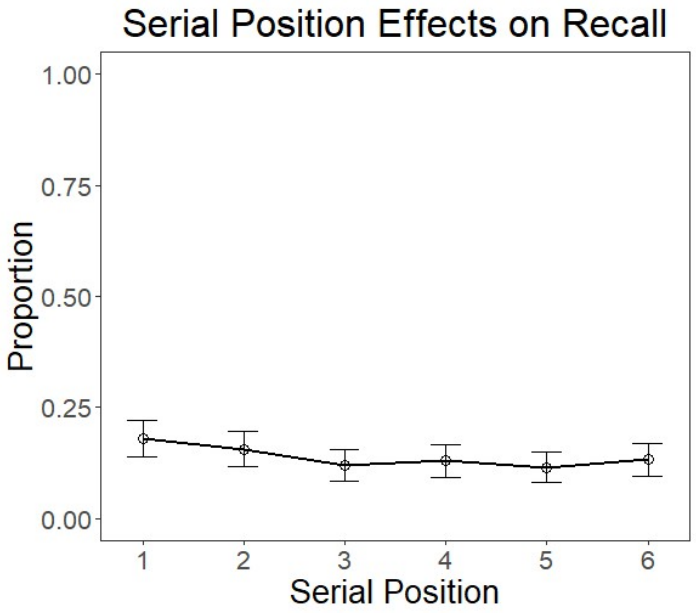
Recall performance was scored using the R stringdist package (v0.9.12; Loo, 2014), and the produced items were scored as recalls if the Levenshtein Distance between the response and the target was two or less. This ensured that typos (e.g., “traanquiliser” instead of “tranquilizer”),

changes in tense (e.g., “allocated” instead of “allocate”), and changes in plurality (e.g., “javelins” instead of “javelin”) could be identified as correct recalls, but the most similar targets “allocate” and “abdicate” would not be treated the same. I then manually checked the response scoring to confirm that each hit with a Levenshtein Distance of one or two was not a distinct word from the target.

The serial position curve for recall performance is displayed in Figure 2.

Figure 2

Recall performance.



Note. The proportion of correct recalls by serial position in Experiment 1a. Error bars represent one standard error above and below the observed proportion.

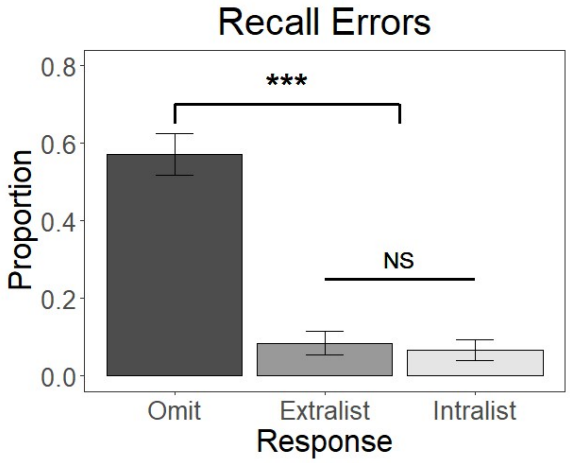
The median and interquartile range are presented for recall performance due to skewed distributions of response types. The mean proportion of omitted responses was highest (*Mdn* = 0.58, *IQR* = 0.42 – 0.75), followed by the proportion of recalls (*Mdn* = 0.25, *IQR* = 0.08 – 0.41), intralist intrusions (*M*dn = 0, *IQR* = 0 – 0.08), and extralist intrusions (*Mdn* = 0, *IQR* = 0 - 0.08).

A one-way within-subjects ANOVA showed that there was a significant main effect of serial position on recall performance, $F(5, 420) = 2.39$, $p = .038$, such that items near the beginning of the list were recalled more often.

Figure 3 displays the mean error proportions for the recall task.

Figure 3

Recall errors.



Note. The proportion of omits, extralist intrusions, and intralist intrusions in Experiment 1a. Error bars represent one standard error above and below the observed proportions.

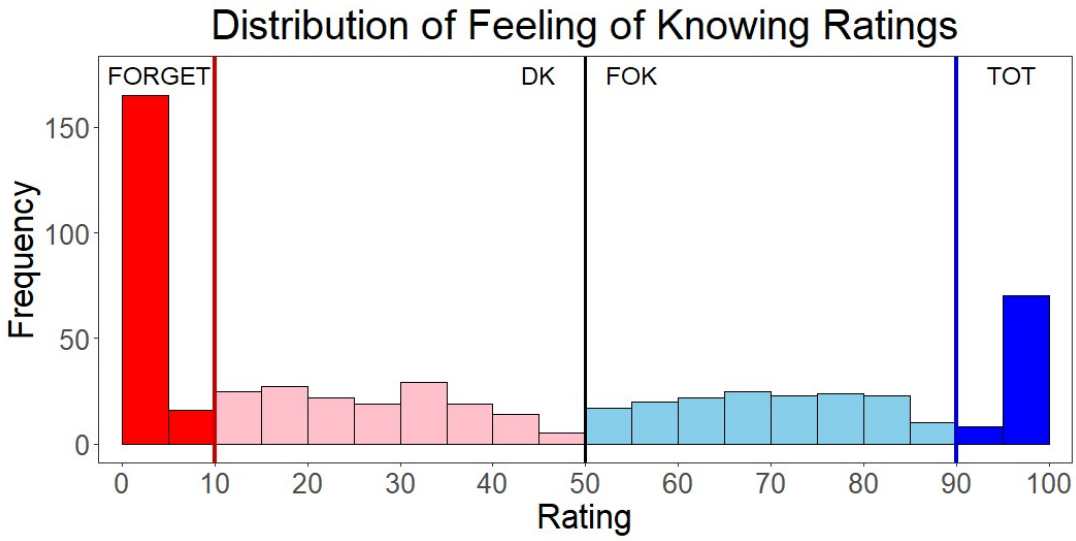
The overall error proportions were analyzed using a one-way within-subjects ANOVA, and there was a significant difference between the three error types, $F(2, 168) = 220.80$, $p < .001$. Planned contrast analysis showed that there were significantly more omits than intrusions, $F(1, 84) = 1197.59$, $p < .001$, and no significant difference between the proportions of extralist and intralist intrusions, $F(1, 84) = 1.14$, $p = .288$.

TOT Analysis

The distribution of sliding-scale FOK ratings following omits, and the thresholds defining FORGET and TOT trials are found in Figure 4.

Figure 4

Sliding-scale responses.



Note. Distribution of FOK ratings on a scale from 0 – "I do not know it at all", to 100 – "I am certain I know it, but it is on the tip of my tongue". Ratings of 10 or below (dark red region) were categorized as FORGET trials and ratings of 90 or above (dark blue region) were categorized as TOT trials. For the analysis of FOKs, all ratings below 50 were categorized as DKs (full red region) and all ratings above 50 were categorized as FOKs (full blue region).

To define the mental states of interest, FORGETs and TOTs, I examined the sliding-scale TOT ratings obtained when participants were unable to produce a target word. Out of 583 trials where a recall response was omitted, the minimum rating of zero was selected 130 times and the maximum rating of 100 was selected 57 times. Since most participants tended to characterize themselves using the extreme ends of the scale and indicated feelings between FORGETs and TOTs more rarely, I used thresholds to categorize TOT trials and FORGET trials. To accommodate imprecise use of the slider, FORGET trials were defined by responses of 10 or lower and TOT trials were defined by responses of 90 or higher. Using these cutoffs, 181 FORGET trials were obtained from 50 participants and 79 TOT trials were collected from 26

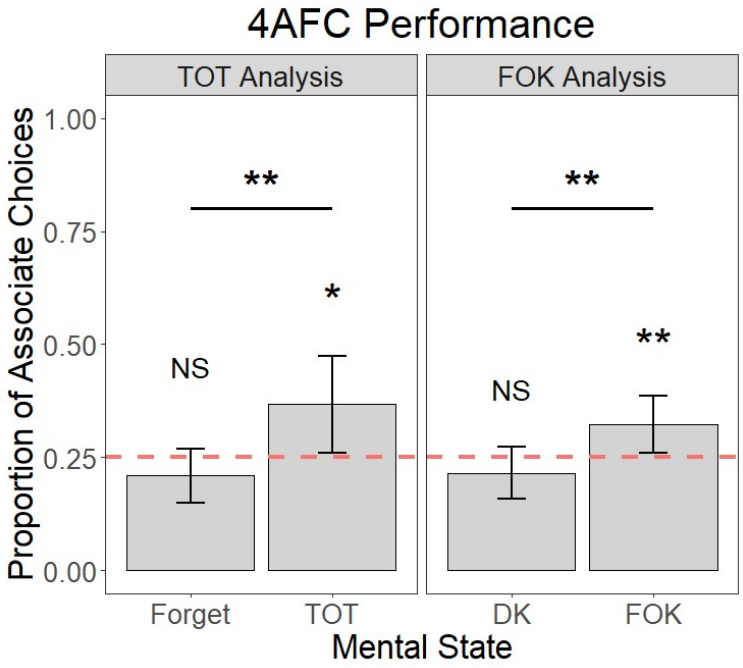
participants. Across both types of trials, data from 66 participants contributed to the following analyses.

4AFC:

Figure 5 contains the associate choice proportions during FORGET and TOT trials.

Figure 5

Performance on the 4AFC task.



Note. Associate choices occurred in significantly greater proportions during TOT trials compared to FORGET trials and during FOK trials compared to DK trials. The proportion of associate choices during TOT and FOK trials, but not FORGET or DK trials, was significantly above random chance. Error bars indicate the 95% CI for the observed proportions.

Two-sided tests of proportions were used to compare the proportions of associate choice during FORGET and TOT trials, as the number of observations for each mental state could not be directly controlled or balanced. A two-sided test of proportions showed that participants performed differently on the 4AFC task during TOT trials than during FORGET trials, $z = 2.66$,

p = .008. Two-sided tests of proportions were also used to determine whether 4AFC performance during TOT trials and FORGET trials differed from guessing, given a probability of 0.25 to choose the associate at random. The proportion of associate choices during FORGET trials (0.21) was not significantly different from chance, z = 1.25, p = .213, whereas the proportion of associate choices during TOT trials (0.37) was significantly above chance, z = 2.40, p = .016. This pattern supports the conclusion that participants experiencing TOTs had retrieved more semantic information about the target, thus improving their ability to recognize the most semantically related item during 4AFC.

FOK Analysis

Although participants tended to use the extreme ends of the slider when defining their mental states, the full length of the slider was used to some degree (see Figure 4). The responses in between FORGETs and TOTs (range = 11 – 89) may still provide insight into lexical retrieval and allow for a greater number of observations than relying solely on extreme ratings corresponding to TOTs. Although the exact relationship between FOKs and TOTs is debated, for my purposes TOTs are trials where participants were compelled to move their slider to the upper extreme (ratings of 90 to 100) whereas FOKs are trials in which they moved their slider to the right by any magnitude (51 to 100). Likewise, FORGETs are trials in which participants moved their slider to the lower extreme (ratings of 0 to 10) and DKs are trials in which they moved their slider to the left by any magnitude (ratings of 0 to 49). This follows the findings of Bahrick et al. (2011) who directly compared TOTs and FOKS using recall and recognition tasks to show that performance in TOTs was consistent with dichotomized judgements on the continuous FOK confidence scale. The analysis of TOTs and FORGETs will be limited to trials from both extremes of the scale and the analysis of FOKs and DKs will use a binary split for ratings above

or below the midpoint of the scale (i.e., 50). Differentiating TOTs in this manner is more like nominal TOT scales (e.g., Umanath et al., 2025), and also reduces the impact of individual differences in TOT criterion points that Bahrick et al. (2011) identified in their FOK confidence scale (i.e., FOK ratings of 4/6 corresponded with TOT-like performance for conservative raters and no-TOT-like performance for liberal raters).

See Figure 4 for the distribution of FOK ratings and criterion point for DK and FOK trials. The following analyses compare performance between DK trials (FOK ratings below 50) and FOK trials (FOK ratings above 50). There were a total of 339 DK trials and 242 FOK trials, and two excluded trials where participants moved the slider before returning it to exactly 50.

4AFC:

See Figure 5 for associate choice proportions during DK and FOK trials. A two-sided test of proportions showed that the proportion of associates chosen during the 4AFC task was significantly higher for FOK trials than DK trials, $z = 2.90$, $p = .004$. Given a 0.25 probability of guessing the associate at random, two-sided tests of proportions also showed that the proportion of associate choices during FOK trials (0.32) was significantly above chance, $z = 2.60$, $p = .009$, whereas the proportion during DK trials (0.21) did not significantly differ from chance, $z = 1.47$, $p = .141$. This pattern further supports the metacognitive monitoring of semantic retrieval, as higher FOKs were associated with increased ability to recognize the most semantically related item during 4AFC.

Discussion

This experiment showed that a cued recall method can effectively induce TOTs despite limiting the semantic information provided, as approximately one third of the participants

produced at least one TOT trial and the overall rate amounted to an average of 1.06 TOT trials per participant. Performance on the 4AFC task supported the primacy of retrieving semantic information during word recall, and its contribution to the experience of TOTs. While experiencing a TOT, participants were more likely to choose the associate for their unrecalled target than random chance would dictate. This supports prior research implicating incomplete semantic retrieval or failure to map semantic information to the studied target as potential underlying causes of TOTs (R. Brown & McNeill, 1966; Caramazza & Miozzo, 1997; Hanly & Vandenberg, 2010; Meyer & Bock, 1992; Roelofs et al., 1998). When participants had no sense of a TOT state, their ability to choose the associate was not significantly different from random guessing. This suggests that the performance during TOT trials is unlikely to be explained by an issue with the relationships between the clues, associates, and foils. Instead, the distinction between performance during TOT trials and FORGET trials supports a change in the mental state experienced by participants during TOT trials, one which they could deliberately and accurately report.

Experiment 1b: Semantic retrieval

A shortcoming of Experiment 1a was the difficulty of assembling stimuli such that the clue for each target could be associated with the target and all four alternative words, while also ensuring that the associate is more closely related to the target than the foils. Experiment 1b will address this by using a paired associate learning task in which participants study lists of cue-target pairs, after which they are invited to recall the targets in response to a cue (e.g., Kahana, 2002). The semantic relationships linking the clue to the target, associate, and foils in Experiment 1a will be replaced by the acquired cue-target pairing, thus the relationships to control are only the semantic similarity between the target, associate, and foils. This will provide

additional data to corroborate the results from Experiment 1a, and comparing the results of both experiments will give evidence for how effective the paired associate learning task is for inducing TOTs.

Method

Participants: A two-tailed 80% power analysis was conducted in G*Power 3.1 (Faul et al., 2009). Using the proportions of associate choices made on TOT trials and FORGET trials in Experiment 1a, with the observed rates of TOTs and FORGETs within that sample, it was found that a minimum of 82 FORGET trials and 46 TOT trials must be collected to achieve power equal to 0.8. Each participant in Experiment 1a contributed an average of 1.06 TOT trials and 1.86 FORGET trials, which would indicate that at least 82 participants should be recruited. Since Experiment 1b followed a different procedure which may impact performance, I recruited 100 undergraduate students (34 male, M_{age} = 19.38 years, SD = 4.11) from the University of Manitoba SONA psychology participant pool. Fifty participants completed the study online and 50 completed the study in computer labs at the University of Manitoba under researcher supervision. Students participating in-person were tested in groups of up to 20 participants. All participants received one research participation credit for their introduction to psychology course. To be eligible, participants must not have completed Experiment 1a. Because the study involved reading and typing English words, only participants who indicated that they are comfortable reading and writing in English on the departmental pre-screening questionnaire were eligible to participate.

Materials: The materials include a list of 24 cue – target pairs, as well as a semantic associate of the target and three foils which are not closely related to the target by semantics, phonology, or orthography. A vector analysis using semantic representations from LSA (Landauer & Dumais,

1997) verified the greater semantic similarity for target-associate pairs (M = 0.40, SD = 0.23) than target-foil pairs (M = 0.06, SD = 0.06). Comparing representations from Pincelate (Parrish, 2017) showed that phonological similarity was low for both target-associate pairs (M = 0.02, SD = 0.08) and target-foil pairs (M = 0.02, SD = 0.09). Finally, comparing representations from SERIOL2 (Reid et al., 2023; Whitney & Marton, 2013) showed that orthographic similarity was low for both target-associate pairs (M = 0.12, SD = 0.14) and target-foil pairs (M = 0.09, SD = 0.11). Target words and their associates were selected from cue-target pairs in the University of South Florida free association norms (Nelson et al., 2004), and the foil and cue words were randomly selected words from the same set of norms. The stimuli for Experiment 1b can be found in Appendix B.

Procedure: The procedure is similar to Experiment 1a, with changes to the stimuli and recall task. All text was presented in black on a white background, with all instructions printed in 18pt Arial font and all stimuli presented in lowercase 21pt Arial font.

After providing informed consent, participants completed four study/test cycles. During each study phase, participants viewed six cue – target pairs (e.g., cork – drive). These pairs were presented one pair at a time for 2000ms, followed by a 1000ms interstimulus interval with a fixation cross. After studying six pairs, participants completed a cued recall test for the studied targets. On each test trial, one of the studied cues was presented on the screen (e.g., cork) and participants reported the paired target (i.e., drive) by typing it in a textbox and pressing a button to proceed to the next trial. When unable to recall the target, they proceeded without typing anything, then indicated on a sliding scale the extent to which they felt they had forgotten the target or had it on the tip of their tongue. The slider recorded responses about the participant's sense of TOT on an invisible 100-point scale, with 0 labelled "I do not know it at all" and 100

labelled “I am certain I know it, but it is on the tip of my tongue.” After indicating their metacognitive judgement by moving the slider in either direction and clicking continue to proceed, participants were presented with four words including an unstudied semantic associate of the target (e.g., car) and three unrelated words. They chose one word amongst the four alternatives that they believed best matched the target they were unable to report. All four choice alternatives were presented simultaneously in a randomized order in a single row across the screen. After selecting one of the four words by clicking on it, they had a final opportunity to recall and type the target before proceeding to the next trial. Once all six cue words were presented, the study/test cycle repeated three more times, each time with a new list of six cue – target pairs. After completing all four study and test phases participants who completed the study online or in the lab were asked if they had been distracted or had been multi-tasking while completing the experiment, after which they were debriefed and thanked for their participation.

Experiment 1b continued to use six-item study lists as this list length proved effective in Experiment 1a. Participants in Experiment 1a tended to move through the test phases more quickly than anticipated so the total number of trials was increased to 24 (from 12 in Experiment 1a) to improve the precision of point estimates of performance, and to provide more opportunities for TOTs to arise on a participant-to-participant basis.

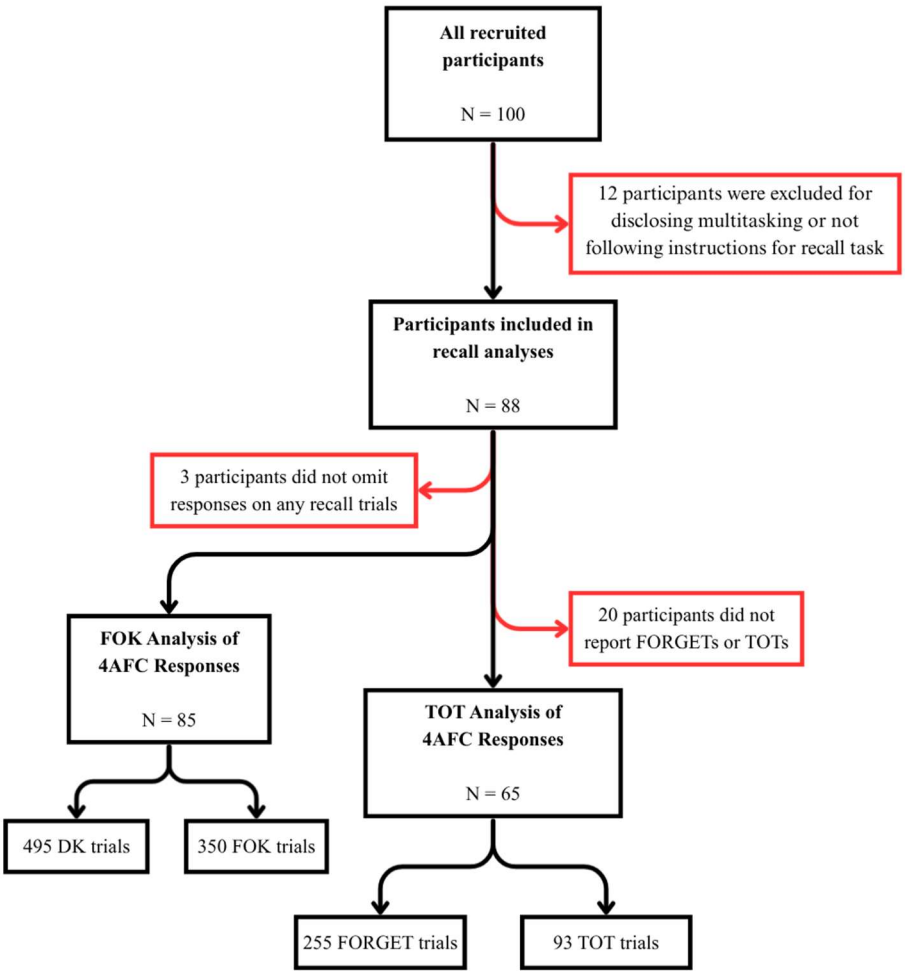
Results

Of the 100 participants, 12 were excluded from the analysis for disclosing distractions and multitasking such as having conversations, using their phones, or watching television while completing the tasks. One of these participants completed the experiment in the computer lab and the other 11 completed the experiment online. This left a sample size of 88 for the following analyses (49 in-person and 39 online participants). A comparison of results from the online and

in-person versions of the experiment is found in Appendix F. Figure 6 displays a flowchart of data processing including the number of participants or trials contributing to each section of the analysis.

Figure 6

Data processing flowchart.



Note. From the full sample of 100 participants, 12 were excluded for failing to follow instructions or disclosing multitasking and distractions while completing the experiment. Recall performance was analyzed for the remaining 88 participants. The FOK analysis included all 4AFC trials from participants who did not report a word on every recall trial, with 85 participants contributing 495 DKs and 350 FOKs. The TOT analysis included only trials where participants

indicated a strong sense of forgetting or TOT state, with 65 participants contributing 255 FORGETs and 93 TOTs.

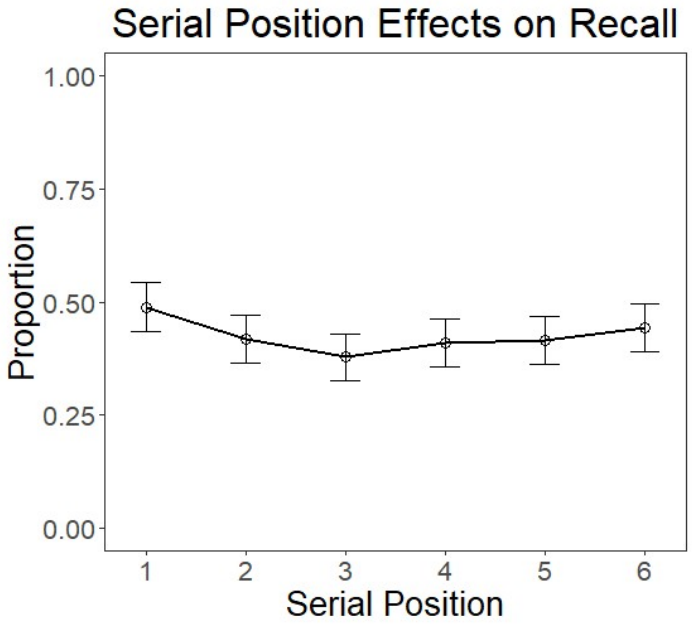
Recall:

Responses on the recall task were scored using the R stringdist package (v0.9.12; Loo, 2014) and responses with a Levenshtein distance of one from the target were scored as correct to allow small typos (e.g., “drvie” instead of “drive”) and changes in plurality (e.g., “flowers” instead of “flower”). These items were manually checked by the author to ensure that any responses with a distinct word were scored as incorrect even if they were within that window (e.g., “moth” instead of “mouth”).

Recall performance is presented in Figure 7.

Figure 7

Recall performance.



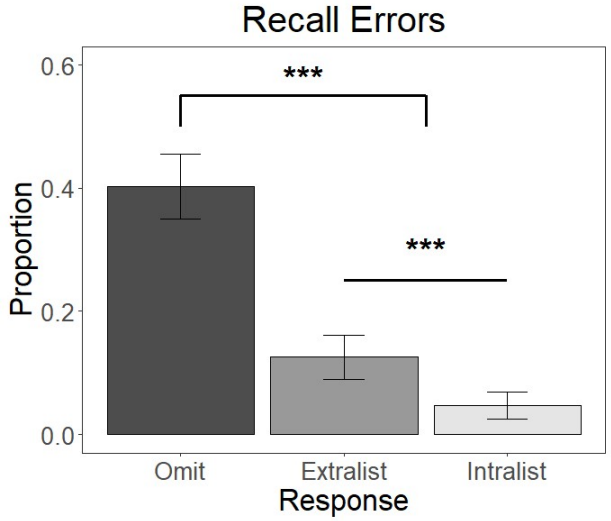
Note. The proportion of correct recalls by serial position in Experiment 1b. Error bars represent one standard error above and below the observed proportions.

The median and interquartile range for recall performance are presented due to skewed distributions of intrusion errors. The proportion of omitted responses was greatest (*Mdn* = 0.42, *IQR* = 0.29), followed closely by recalls (*Mdn* = 0.38, *IQR* = 0.29), extralist intrusions (*Mdn* = 0.08, *IQR* = 0.13), and intralist intrusions (*Mdn* = 0.04, *IQR* = 0.08). A one-way within-subjects ANOVA showed that there was a significant main effect of serial position on recall performance, $F(5, 435) = 2.60$, $p = .025$, with recall better for targets studied in early and late serial positions.

The error proportions are presented in Figure 8.

Figure 8

Recall errors.



Note. The proportion of omits, extralist intrusions, and intralist intrusions in Experiment 1b. Error bars represent one standard error above and below the observed proportions.

The overall error proportions were analyzed using a one-way within-subjects ANOVA, and as shown in Figure 8 there was a significant difference between the three error types, $F(1, 174) = 119.60$, $p < .001$. Planned contrast analysis showed that there were significantly more omits than intrusions, $F(1, 87) = 853.27$, $p < .001$, and significantly more extralist than intralist intrusions, $F(1, 87) = 39.52$, $p < .001$.

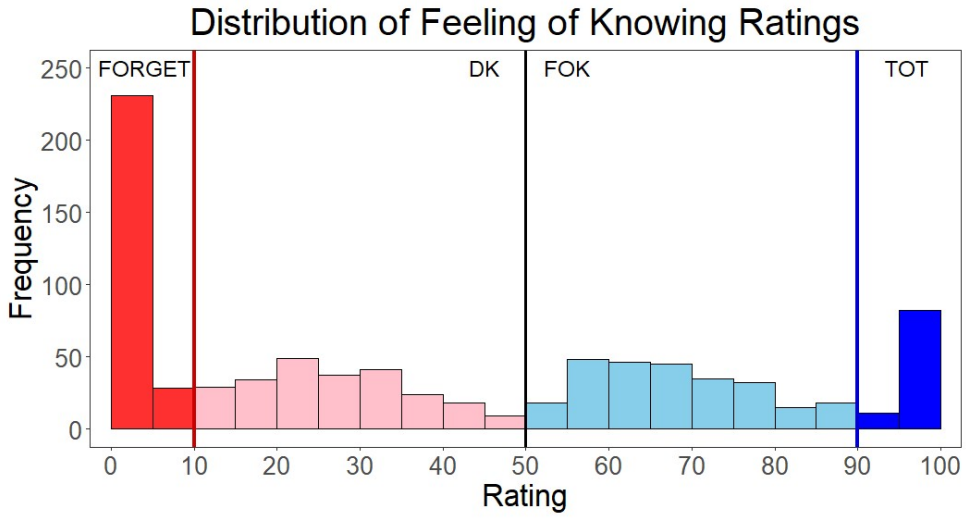
The extralist intrusions were analyzed for their similarity to the correct target, using vector analysis. Excluding nonword responses (e.g., “drin”), the cosine similarity between an extralist response and the correct target was highest for orthography ($M = 0.136$, $SD = 0.158$), followed by semantics ($M = 0.127$, $SD = 0.129$), and finally phonology ($M = 0.044$, $SD = 0.151$).

TOT Analysis

The distribution of sliding-scale FOK ratings following omits, and the thresholds defining FORGET and TOT trials are found in Figure 9.

Figure 9

Sliding-scale responses.



Note. Distribution of TOT ratings on a scale from 0 – “I do not know it at all”, to 100 – “I am certain I know it, but it is on the tip of my tongue”. Ratings of 10 or below (dark red region) were categorized as FORGET trials and ratings of 90 or above (dark blue region) were categorized as TOT trials. For the analysis of FOKs, all ratings below 50 were categorized as DKs (full red region) and all ratings above 50 were categorized as FOKs (full blue region).

Eighty-five participants contributed a total of 850 omits when a target could not be recalled, with sliding scale TOT ratings. As in Experiment 1a, participants used the center of the sliding scale more moderately and produced peaks at both extremes indicating they commonly

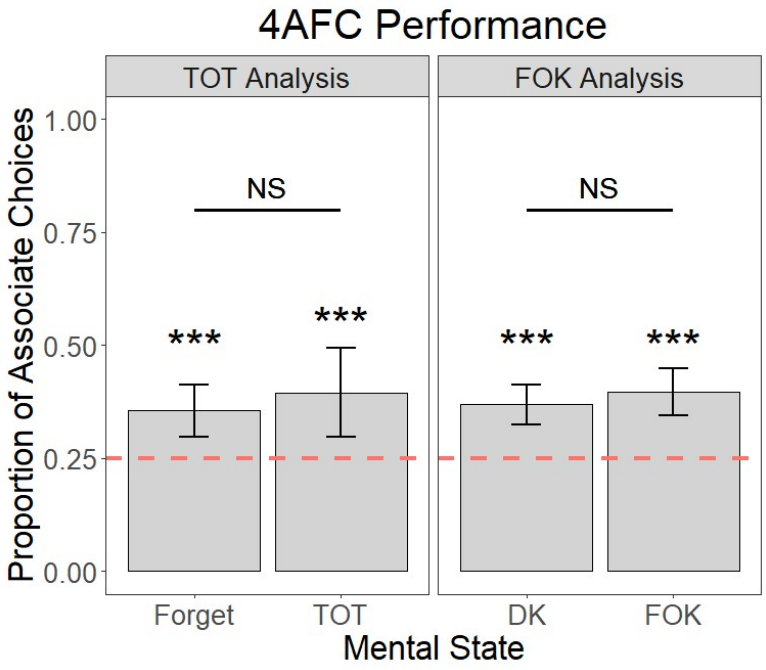
felt a strong sense of either FORGET or TOT. The same thresholds from Experiment 1a were used to define the mental states of interest, with responses of 10 or below constituting FORGET trials and responses of 90 or above constituting TOT trials. Using these cutoffs, there were 255 FORGET trials (174 responses of exactly zero) from 57 participants, and 93 TOT trials (69 responses of exactly 100) from 33 participants. Across both types of trials, responses from 65 of the total 88 participants contributed to the following analyses.

4AFC:

Figure 10 contains bar graphs depicting associate choice proportions during TOT and FORGET trials.

Figure 10

Performance on the 4AFC task.



Note. The proportion of associate choices was significantly above the random chance level of 0.25 (red line) for all trial types. This proportion did not significantly differ between FORGET

and TOT trials or DK and FOK trials. Error bars indicate the 95% CI for the observed proportions.

A two-sided test of proportions showed that participants did not behave significantly differently on the 4AFC task during TOT trials than during FORGET trials, $z = 0.71$, $p = .480$. Given a 0.25 probability of guessing the associate at random, two-sided tests of proportions also showed that the proportion of associate choices during TOT trials (0.40) was significantly above chance, $z = 3.30$, $p < .001$ as was the proportion during FORGET trials (0.36), $z = 3.30$, $p < .001$. Appendix G contains an analysis of 4AFC performance with additional exclusion criteria removing participants who performed very poorly on the recall task.

Final recall:

There were 313 out of 850 trials (37%) where a target was not initially recalled, but was reported following completion of the 4AFC task, and the proportion of correct recalls at this stage was 0.26 overall. Contrasting by mental state, targets were reported following 38 out of 96 TOT trials (40%) with a correct recall rate of 0.34, and 76 out of 259 FORGET trials (29%) with a correct recall rate of 0.16. A statistical test confirmed that the proportion of recalls was significantly higher for targets reported following TOT trials than FORGET trials, $z = 2.24$, $p = .025$. Comparing by 4AFC choice, targets were reported following 155 out of 324 associate choices (48%) with a correct recall rate of 0.46, and 158 out of 526 foil choices (30%) with a correct recall rate of 0.06. A statistical test confirmed that the proportion of recalls was significantly higher for responses after participants chose an associate over a foil, $z = 8.13$, $p < .001$. Of the 197 extralist errors reported during the final recall task, 53 came from FORGET trials and 21 came from TOT trials. Participants reported the semantic associate from the 4AFC task 52 times and a foil word 86 times. Unlike initial recall attempts, vector analysis showed that

extralist intrusions at this stage were most like the target in meaning regardless of whether the response followed a TOT or FORGET trial. This semantic dominance was driven by participants reporting the associate word after viewing it on the 4AFC task. Excluding trials where the associate was reported left 145 trials overall, with 36 from FORGET trials and 19 from TOT trials. FORGET trials were like initial recall with greater orthographic similarity followed by semantic and phonological. Across all trials, orthographic and semantic similarity was very close, with phonology trailing. On TOT trials, the similarity of all three modalities increased, and semantic similarity was slightly greater than orthographic similarity. The cosine similarities between extralist responses and targets, excluding associate reports, are found in Table 1. Excluding both associate and foil responses resulted in nearly identical results as excluding only associate responses.

Table 1

Mean cosine similarity of extralist responses and targets on the final recall task.

	Semantic		Phonological		Orthographic	
	M	*SD*	*M*	*SD*	*M*	*SD*
All Trials	0.107	0.125	0.031	0.129	0.113	0.141
TOT	0.157	0.197	0.084	0.187	0.151	0.212
FORGET	0.072	0.062	0.033	0.122	0.111	0.120

Note. Extralist errors on FORGET trials were closest to the target's orthography, followed by semantics and phonology. Over all trials, phonology and orthography were similar to FORGET trials, whereas semantics increased to be similar to orthography. Finally, all three modalities increased on TOT trials with semantics becoming slightly greater than orthography and phonology continuing to trail.

FOK Analysis

Figure 9 displays the categories, the criterion point, and the distribution of ratings for DK and FOK trials. The following analyses compare performance between DK trials (TOT ratings

below 50) and FOK trials (TOT ratings above 50). There were a total of 350 FOK trials, 495 DK trials, and five excluded trials where participants moved the slider and returned it to exactly 50.

4AFC:

Figure 10 displays the associate choice proportions for FOK and DK trials. A two-sided test of proportions showed that there was no significant difference in the proportion of trials in which participants chose the target's associate during FOK and DK trials, $z = 0.80$, $p = .424$. Given a 0.25 probability of guessing the associate at random, two-sided tests of proportions also showed that the proportion during FOK trials (0.40) was significantly above chance, $z = 6.36$, $p < .001$, as was the proportion during DK trials (0.37), $z = 6.15$, $p < .001$.

Final Recall:

A target was reported following 161 out of 350 FOK trials (46%) with a correct recall rate of 0.29 and following 150 out of 495 DK trials (30%) with a correct recall rate of 0.22. There was no significant difference in the proportion of recalls between FOK and DK trials, $z = 1.45$, $p = 0.147$.

Discussion

Performance on the recall task was slightly improved compared to Experiment 1a, with a higher proportion of recalls, a lower proportion of omits, and a slight increase in extralist errors. The prevalence of TOTs was similar to Experiment 1a despite the change to a paired associate recall task, with 38% of participants reporting at least one TOT and an overall rate of 1.06 TOT trials per participant. Performance on the 4AFC task, however, showed several differences when compared to Experiment 1a. Firstly, participants were not significantly more likely to select the semantic associate on TOT trials than FORGET trials. Secondly, participants did select the

associate significantly above chance during FORGET trials as well as TOT trials. These differences may be due to replacing the semantic clues from Experiment 1a with a cue-word presentation in Experiment 1b. Participants could have relied more on the retrieval of non-semantic lexical information during the task because they were not restricted by a clue that only revealed semantic target information. Furthermore, the increase in associate guesses during FORGET trials likely occurred due to the inclusion of random foil words. In Experiment 1a, the foil words had to be plausible responses to the semantic clue thus all responses had some shared meaning. Participants in Experiment 1b would need to retrieve less semantic information to rule out foil words. Thus the changes in 4AFC performance between Experiment 1a and 1b may be explained by changes in retrieval strategy and increased discriminability of foils on FORGET trials.

Although there was no difference between TOT and FORGET trials in the 4AFC task, there were significant differences on the final recall task where TOT trials were associated with an 11% higher probability of reporting a word and greater accuracy of those reports. It is impossible to know how many TOTs may have been resolved after the trial (+TOTs) or not (-TOTs), but the presence of +TOTs for the correct target during the final recall task supports the efficacy of this procedure for eliciting TOTs. Choosing the associate word rather than a foil on the 4AFC task was also associated with an 18% higher probability of reporting a target and increased accuracy. The associate seemed to act as a secondary cue, providing additional semantic information to aid retrieval of the unrecalled target as if through triangulation. Coupled with the results of Experiment 1a this result supports the conclusion that semantics play a role in TOTs, but some semantic input may be necessary for individuals to shift their focus towards semantic information. Extralist errors on the final recall task overall had similar relationships with the

target word to extralist errors on initial recall attempts, but guesses on TOT trials were closer to the target for all modalities and semantic similarity narrowly surpassed orthographic similarity. Since there was an increased hit rate following associate choice and similarities increased on TOT trials for all modalities even when associate reports were excluded, this is likely due to additional support from the associate word. The associate commonly felt "close enough," leading participants to report it as the target despite knowing they must be different words, which contributed to semantic dominance in extralist errors when those associate reports were included. Orthographic similarity was high for extralist errors on initial recall trials as well as all final recall trial types, likely because the procedure used only written stimuli and typed responses.

The analysis of FOKs and DKs supported the findings from TOTs and FORGETs. In the 4AFC task, there was still no significant difference in associate choice rates. The final recall responses showed the same relationship of increased recall attempts and response accuracy during FOK trials but with a smaller difference than the TOT-FORGET comparisons, as expected from the less extreme mental state. In both cases, their confidence in recalling the target was validated by their final recall performance but not linked to their ability to identify a semantic associate. This contrasts Experiment 1a in which high ratings corresponded with 4AFC performance. Retrieving semantic information of a target may trigger TOTs only when the task requires retrieval through a semantic clue, or when the semantic information available is great enough to distinguish an associate from amongst four items that are similarly plausible for that clue as opposed to distinguishing an associate from random foils.

Experiment 2: Phonological retrieval

The initial experiments served to gather data supporting and replicating established findings that semantic information plays an important and early role in lexical retrieval, while

building up an experimental framework for investigating TOTs which occur in the midst of retrieval. Taking the lessons learned from Experiments 1a and 1b, I now turn to a different kind of lexical information which has been intimately linked to TOT states – phonology. The relationship between phonology and TOTs has been a central focus since seminal work of Brown and McNeill (1966) who observed that similar sounding words were more commonly produced than similar meaning words during TOTs, that the similar sounding words shared first and last letters with the target word at a high rate, and that participants appeared capable of identifying syllabic stress patterns for unrecalled targets.

Further research into the role of phonology in lexical retrieval has repeatedly shown that acquiring phonological information is more beneficial to recall than increasing access to semantic information. Meyer and Bock (1992) found that presenting participants with phonological cue words (e.g., cueing "malevolent" with "molecular") in a prospecting study improved lexical retrieval more than providing them with semantic cues (e.g., cueing "malevolent" with "hostile"). This makes intuitive sense, as semantic cues would tend to be redundant alongside the semantic information in the definition provided but the phonological cue provides novel details and gets to the core problem of TOTs – failure to report the form of the word itself. Kumar et al. (2019) presented definitions to cue participants' recall of a target word, followed immediately by a phonological prime, semantic prime, or prime for both modalities at once. Phonological primes were more likely to aid in correct recall of the target compared to semantic primes, while primes for both semantics and phonology fell in between. In primes for both modalities, the semantic overlap between the target and prime was inverse to target retrieval accuracy so it appears that priming with semantic information diminished participants' ability to make use of the phonological information in the prime. The research above, as with most

research into TOTs, involved presenting a definition for the target word before the cues or primes were presented to ensure that nearly all semantic information about the target was available before presenting the cues or primes. This study aims to test access to phonological information without presenting either target definitions or semantic and phonological primes. Because associates are identified in a forced choice task rather than being freely reported, this experiment will provide a different measure of phonological retrieval than past studies. The shared phonology of the associate must be recognized rather than recalled, and illusory phonological interlopers will not impact the number of guesses as observed in past studies (e.g., Huebert et al., 2023) because participants are limited to one choice.

Methods

Participants: I recruited 102 undergraduate students (35 male, M_{age} = 20.35 years, *SD* = 5.47) from the University of Manitoba SONA psychology participant pool. Fifty participants completed the study online and 52 completed the study in computer labs at the University of Manitoba under researcher supervision. Students participating in-person were tested in groups of up to 20 participants. All participants received one research participation credit for their introduction to psychology course. To be eligible, participants must not have completed Experiments 1a or 1b. Because the study involved reading and typing English words, only participants who indicated that they are comfortable reading and writing in English on the departmental pre-screening questionnaire were eligible to participate.

Materials: The materials consist of 24 target words and phonological associates which are pronounced similarly to, but spelled differently from, the targets (i.e., homophones). As in Experiment 1b, each target had a paired cue word, one related associate, and three foil words that were not related to the target by semantics, orthography, or phonology. The targets and

phonological associates include homophone pairs taken from Pexman, Lupker, and Jared (2001) and near-homophone rhyming pairs taken from Polich, McCarthy, Wang, and Donchin (1983). Vector analysis confirmed that target-associate pairs had a greater phonological similarity (M = 0.63, SD = 0.39) than orthographic (M = 0.38, SD = 0.19) or semantic similarity (M = 0.11, SD = 0.11), and that foil-target pairs were not highly similar by semantics (M = 0.11, SD = 0.13), phonology (M = 0, SD = 0.02), or orthography (M = 0.07, SD = 0.09). The complete set of stimuli for this experiment can be found in Appendix C.

Procedure: The procedure follows that of Experiment 1b, differing only by the stimuli used and the relationship between the targets and associate words. After providing informed consent, participants completed four study/test cycles. During the study phase, participants viewed six cue – target pairs (e.g., quantity – aloud). These pairs were presented, one pair at a time, for 2000ms followed by a 1000ms interstimulus interval with a fixation cross. After studying six pairs, participants completed a cued recall test for the studied targets. On each test trial, one of the studied cues was presented (e.g., quantity) and participants reported the paired target (i.e., aloud) by typing it and pressing a button to proceed to the next trial. When unable to recall the target, they proceeded without typing anything, then indicated on a sliding scale the extent to which they felt the target was forgotten or on the tip of their tongue. The slider recorded responses about the participant's sense of TOT on an invisible 100-point scale, with 0 labelled "I do not know it at all" and 100 labelled "I am certain I know it, but it is on the tip of my tongue." After indicating their metacognitive judgement by moving the slider in either direction, participants were presented with four words including an unstudied phonological associate (i.e., a homophone) of the target (e.g., allowed) and three unrelated words. They chose one word amongst the four alternatives which they felt best fit the target they were unable to report. These

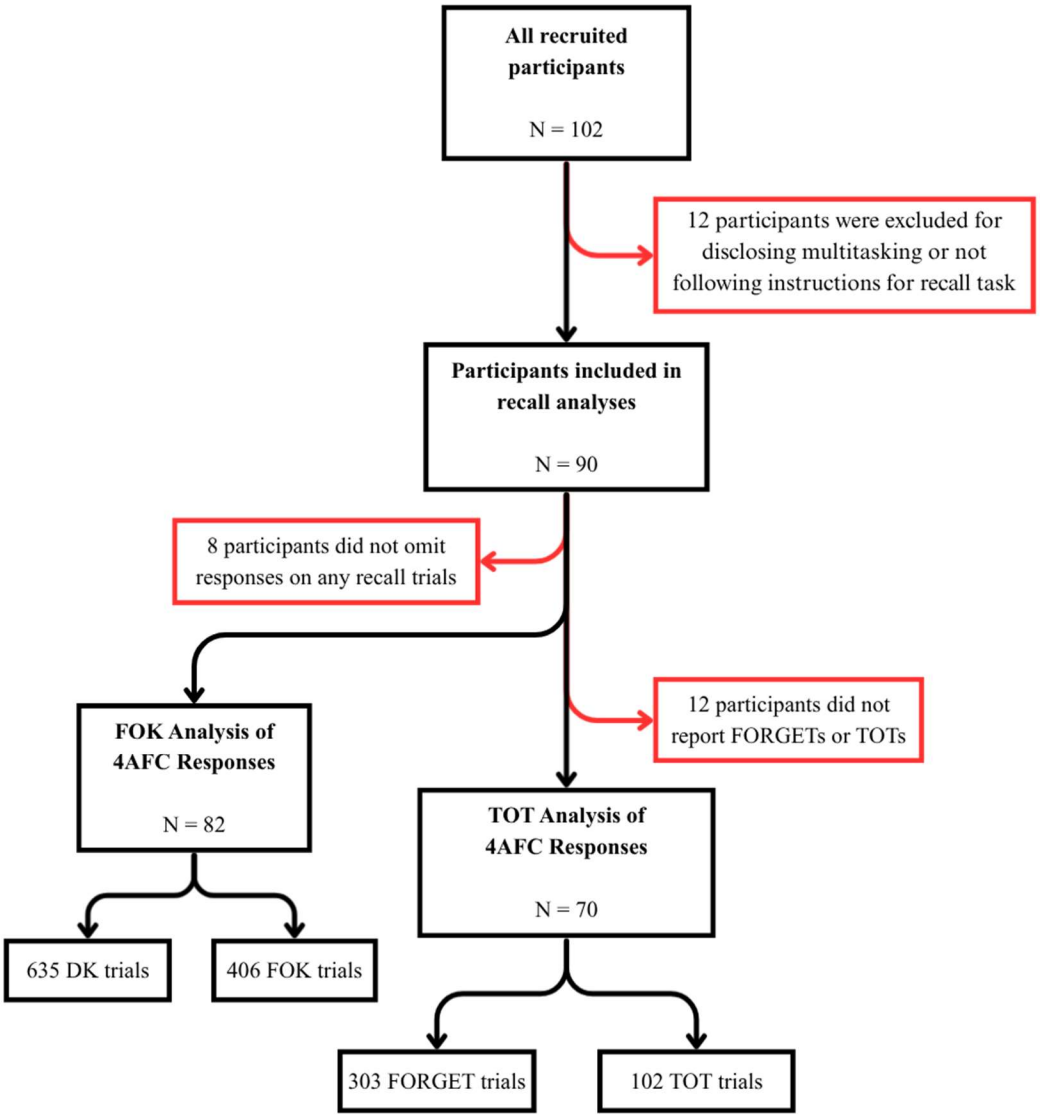
words were presented simultaneously in a randomized order. After selecting one of the four words by clicking on it, they had a final opportunity to recall and type the target before proceeding to the next trial. Once all six cue words were presented, the study/test cycle repeated three more times with new lists of six cue – target pairs. After completing all four study and test phases, participants were asked if they had been distracted or multi-tasking while completing the experiment, after which they were debriefed and thanked for their participation.

Results

Of the 102 participants, 12 were excluded from the analysis for disclosing distractions and multitasking. One of these participants completed the experiment in the computer lab and the other 11 completed the experiment online. This left a sample size of 90 for the following analyses (51 in-person and 39 online participants). A comparison of results from the online and in-person versions of the experiment is found in Appendix F. Figure 11 displays a flowchart of data processing, including the number of participants or trials contributing to each section of the analysis.

Figure 11

Data processing flowchart.



Note. From the full sample of 102 participants, 12 were excluded from the analysis for failing to follow instructions or disclosing multitasking and distractions while completing the experiment. Recall performance was analyzed for the remaining 85 participants. The FOK analysis included all 4AFC trials from participants who did not report a word on every recall trial, with 82 participants contributing 635 DKs and 406 FOKs. The TOT analysis included only trials where participants indicated a strong sense of forgetting or TOT state, with 70 participants contributing 303 FORGETs and 102 TOTs.

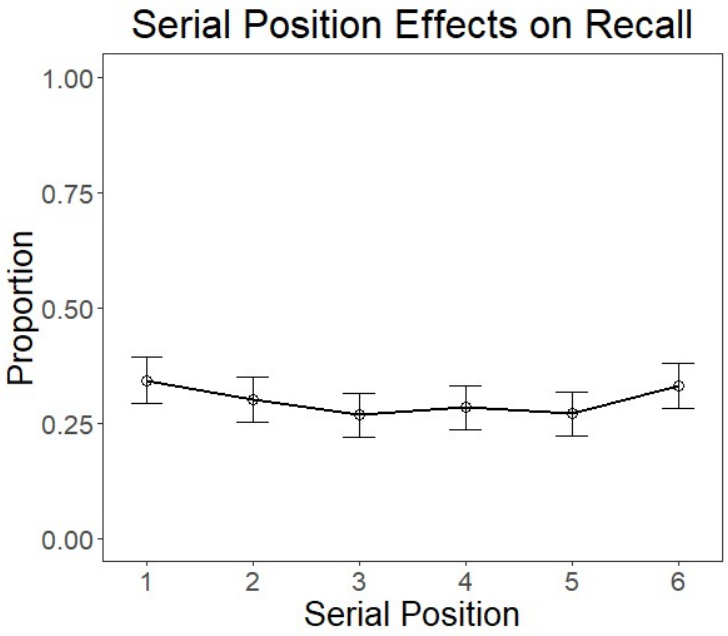
Recall:

I used the R stringdist package (v0.9.12; Loo, 2014) to score recalled words with a Levenshtein distance of one from the target as correct to allow small typos (e.g., “wether” instead of “whether”) and changes in plurality (e.g., “jewels” instead of “jewel”). These items were manually checked by the author to ensure that any responses with a distinct word were scored as intrusions even if they were within that window (e.g., “goose” instead of “moose”).

Recall performance is shown in Figure 12.

Figure 12

Recall performance.



Note. The proportion of correct recalls by serial position in Experiment 2. Error bars represent one standard error above and below the observed proportions.

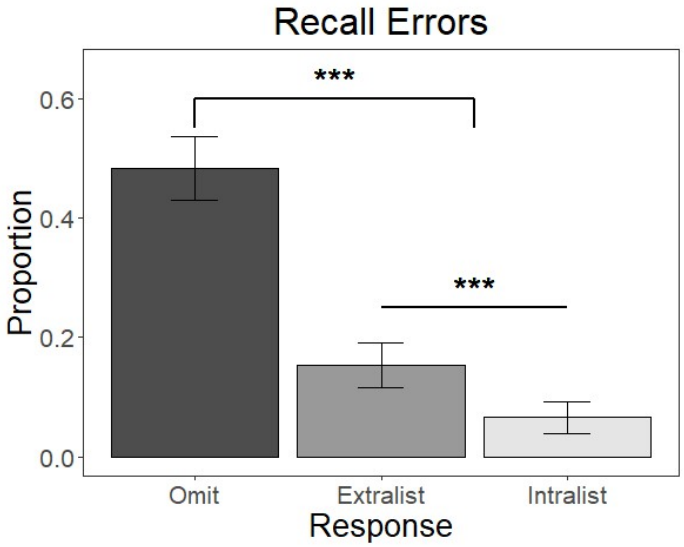
The median and interquartile range for the recall responses are presented due to the presence of skewed distributions. The proportion of omitted responses was greatest (*Mdn* = 0.50, *IQR* = 0.33), followed by recalls (*Mdn* = 0.25, *IQR* = 0.21), extralist intrusions (*Mdn* = 0.08, *IQR* = 0.17), and intralist intrusions (*Mdn* = 0.04, *IQR* = 0.08). A one-way within-subjects ANOVA

showed that there was no significant main effect of serial position on recall performance, $F(5, 445) = 2.09$, $p = .066$.

Figure 13 displays the mean error proportions.

Figure 13

Recall errors.



Note. The proportion of omits, extralist intrusions, and intralist intrusions in Experiment 2. Error bars represent one standard error above and below the observed proportions.

The overall error proportions were analyzed using a one-way within-subjects ANOVA, and there was a significant between the three error types, $F(2, 178) = 113.70$, $p < .001$. Planned contrast analysis showed that there were significantly more omits than intrusions, $F(1, 89) = 930.27$, $p < .001$, and significantly more extralist than intralist intrusions, $F(1, 89) = 38.13$, $p < .001$.

A vector analysis compared the similarity of extralist intrusions to the correct target. Excluding nonword responses (e.g., “mert”), the mean cosine similarity between an extralist

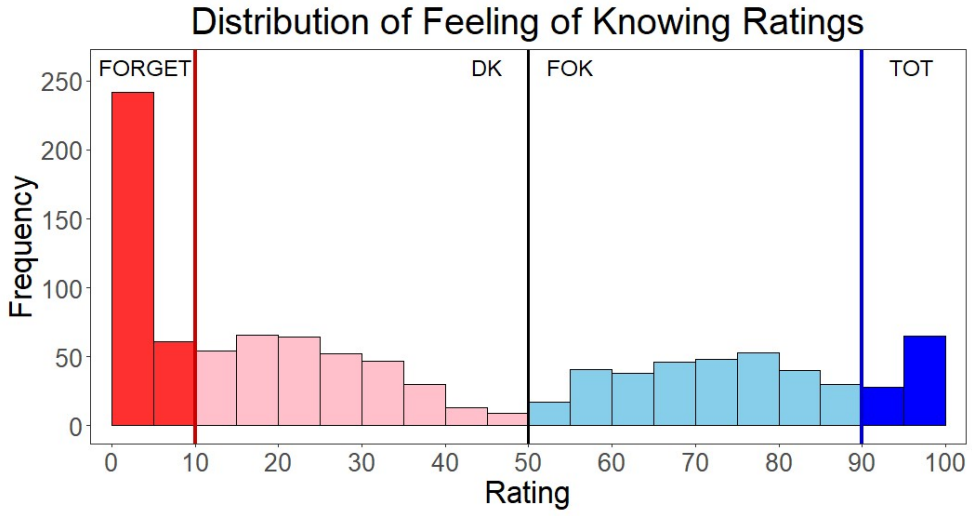
response and the correct target was highest for orthography ($M = 0.214$, $SD = 0.224$) followed by phonology ($M = 0.174$, $SD = 0.334$) and finally semantics ($M = 0.117$, $SD = 0.113$).

TOT Analysis

The distribution of sliding-scale FOK ratings following omits, and the thresholds defining FORGET and TOT trials are found in Figure 14.

Figure 14

Sliding-scale responses.



Note. Distribution of TOT ratings on a scale from 0 – "I do not know it at all", to 100 – "I am certain I know it, but it is on the tip of my tongue". Ratings of 10 or below (dark red region) were categorized as FORGET trials and ratings of 90 or above (dark blue region) were categorized as TOT trials. For the analysis of FOKs, all ratings below 50 were categorized as DKs (full red region) and all ratings above 50 were categorized as FOKs (full blue region).

Eighty-two participants contributed a total of 1044 sliding scale ratings on trials where a target was not recalled. As in previous experiments, participants used the extreme ends of the scale to produce a larger peak at 0 and a smaller peak at 100. Trials with ratings of 10 or below were taken for FORGETs and ratings of 90 or above were taken for TOTs. Using these thresholds, there were 303 FORGET trials (178 responses of exactly zero) from 68 participants,

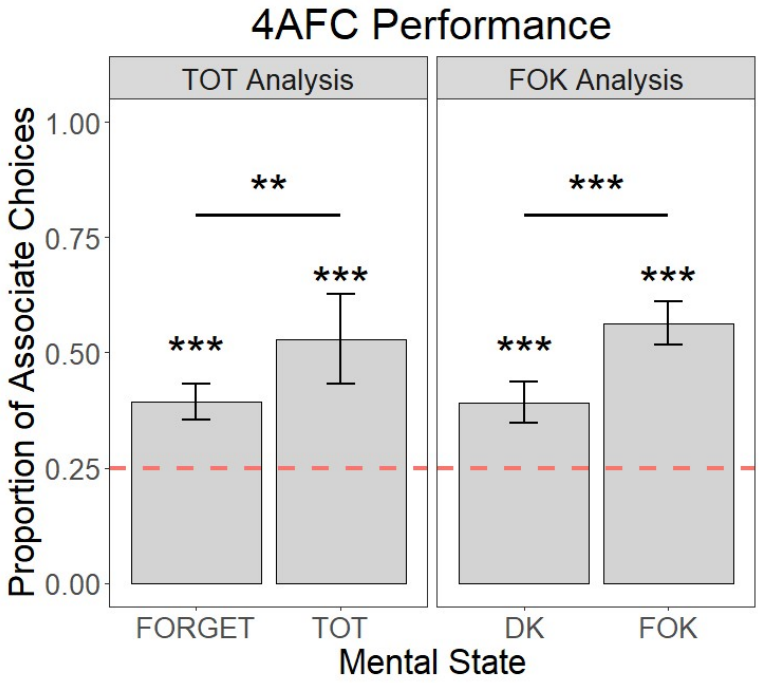
and 102 TOT trials (46 responses of exactly 100) from 34 participants. Across both types of trials, responses from 70 of the total 90 participants contributed to the following analyses.

4AFC:

Figure 15 contains bar graphs depicting associate choice proportions during TOT and FORGET trials.

Figure 15

Performance on the 4AFC task.



Note. Associate choices occurred significantly above the random chance level of 0.25 (red line) for all trial types. The proportion of associate choices was significantly higher during TOT trials compared to FORGET trials as well as FOK trials compared to DK trials. Error bars indicate the 95% CI for the observed proportions.

A two-sided test of proportions showed that participants chose the phonological associate on the 4AFC task significantly more often during TOT trials than during FORGET trials, $z = 3.08$, $p = .002$. Given a 0.25 probability of guessing the associate at random, two-sided tests of

proportions also showed that the proportion of associate choices during TOT trials (0.53) was significantly above chance, $z = 6.52$, $p < .001$, as was the proportion during FORGET trials (0.36), $z = 4.29$, $p < .001$. Across all 4AFC trials, a homophone associate was presented 473 times and a near-homophone associate was presented 568 times. The proportion of associate choices was higher for the homophones (0.57) than near-homophones (0.37), $z = 6.47$, $p < .001$. Homophones were chosen significantly more often than near-homophones whether examining responses during FORGETs, TOTs, or all trials. Comparing by mental state, homophone associates were chosen significantly more often during TOTs (0.70) than FORGETs (0.49), $z = 2.66$, $p = .008$. There was no significant difference in the proportion of near-homophone associate choices during TOTs (0.38) and FORGETs (0.31), $z = 1.04$, $p = .299$. Appendix G contains an analysis of 4AFC performance with additional exclusion criteria removing participants who performed very poorly on the recall task.

Final recall:

Participants reported a target at the end of 380 out of the 1044 trials (36%) where a response was omitted on the initial recall task, with a correct recall rate of 0.15. Targets were reported following 62 out of 102 TOT trials (61%) with a correct recall rate of 0.18, compared to 87 out of 303 FORGET trials (29%) with a correct recall rate of 0.09. There was no significant difference in the proportion of recalls following TOT trials and FORGET trials, $z = 1.54$, $p = 0.123$. Targets were reported following 230 out of 480 associate choices (48%) with a correct recall rate of 0.24, compared to 150 out of 564 foil choices (27%) with a correct recall rate of 0.01. The proportion of recalls was significantly higher following associate choices than foil choices, $z = 6.09$, $p < .001$. Of the 283 extralist errors reported during the final recall task, 150 came from FORGET trials and 41 came from TOT trials. Participants reported the phonological

associate from the 4AFC task 116 times and a foil word 108 times. Vector analysis showed that extralist intrusions at this stage were most similar to the target by phonology regardless of whether the response followed a TOT or FORGET trial. This phonological dominance was driven by participants reporting the associate word after viewing it on the 4AFC task, and removing those trials resulted in orthographic similarity being highest across all trial types. The magnitude of similarities differed depending on whether associate and foil reports or only associate reports were filtered out. In both cases, however, orthography was highest across all trial types, with semantics decreasing and phonology increasing between FORGET and TOT trials. The cosine similarities between extralist responses and targets, excluding associate reports, are found in Table 2. Excluding the associate responses left 167 extralist intrusions in total, with 97 from FORGET trials and 24 from TOT trials.

Table 2

Mean cosine similarity of extralist responses and targets on the final recall task.

	Semantic		Phonological		Orthographic	
	M	*SD*	*M*	*SD*	*M*	*SD*
All Trials	0.093	0.087	0.081	0.218	0.131	0.175
TOT	0.076	0.061	0.081	0.186	0.135	0.201
FORGET	0.089	0.080	0.046	0.173	0.123	0.148

Note. Extralist errors were most similar to the target's orthography, regardless of trial type. Phonological similarity increased between FORGET and TOT trials, whereas semantic similarity decreased.

FOK Analysis

Figure 14 displays the distribution of ratings with the categories and criterion point for DK and FOK trials. The following analyses compare performance between DK trials (TOT

ratings below 50) and FOK trials (TOT ratings above 50). There were a total of 635 DK trials, 406 FOK trials, and three excluded trials where participants returned the slider to exactly 50.

4AFC:

See Figure 15 for bar graphs displaying the associate choice proportions for FOK and DK trials. A two-sided test of proportions showed that the proportion of associate choices was significantly greater during FOK than during DK trials, $z = 5.43$, $p < .001$. Given a 0.25 probability of guessing the associate at random, two-sided tests of proportions also showed that the proportion of associate choices during FOK trials (0.56) was significantly above chance, $z = 14.61$, $p < .001$, as was the proportion during DK trials (0.39), $z = 8.27$, $p < .001$.

Final Recall:

A target was reported following 197 out of 406 FOK trials (49%) with a correct recall rate of 0.19 and following 181 out of 635 DK trials (29%) with a correct recall rate of 0.11. The recall rate was significantly higher for FOK trials than for DK trials, $z = 2.22$, $p = .026$.

Discussion

Recall performance was slightly worse compared to Experiment 1b, with a lower proportion of recalls and more omitted responses. The decrease in recall performance came with an increase in TOTs, as 76% of participants reported at least one TOT and the overall rate was 1.13 TOT trials per participant. 4AFC performance displayed the hypothesized difference, with a significantly higher proportion of associate choices on TOT trials than FORGET trials. As in Experiment 1b, participants selected the associate above chance levels during both FORGET and TOT trials. The rate of associate choice during FORGET trials was very close to Experiment 1b, which supports the explanation that the randomly selected foils are more discriminable when less

information is retrieved compared to the carefully chosen foils from Experiment 1a. Unlike Experiment 1b, participants experiencing TOTs responded particularly strongly to phonological associates which produced a significant increase in performance compared to FORGET trials. This supports the direct-access view that TOTs occur with the retrieval of lexical information from memory and do not solely arise from beliefs that a target must be available or strategic guessing of probable features. Participants were more likely to correctly identify homophone associates than near-homophones that did not share a first letter with the target. TOTs were associated with a significant increase in identifying homophone associates, but not near-homophone associates, which supports previous findings that first letters are more frequent and reliable phonological interlopers during TOTs (Huebert et al., 2023).

As in Experiment 1b, participants were more likely to report a word following the 4AFC task if they experienced a TOT or chose the associate word. Associate choice was associated with a significant increase in hit rate, while TOTs showed an increase that was not significant. There were some differences in extralist intrusions between initial and final recall. Extralist intrusions from initial recall trials were closest to the target by orthography followed by phonology and then semantics, but at final recall phonology overtook orthography. As in Experiment 1b this was driven by participants reporting the phonological associate during the final recall task, and removing associate responses resulted in orthographic similarity being most prevalent. This supports the conclusion that the associate words can act as a secondary cue that aids participants in attempting to triangulate the target using the new information along with what they had retrieved from the initial cue. Considering the increased hit rate following associate choices, it appears that the more similar extralist intrusions are due to both cueing from the associate and cases where participants reported the associate itself as an intrusion.

The analysis of FOKs and DKs supported the general findings from TOTs and FORGETs. The difference in associate choice proportions for FOKs and DKs was consistent with the difference for TOTs and FORGETs, and both trial types were significantly above random chance once again. FOKs were associated with increased attempts to recall the target following the 4AFC task, and the hit rate was significantly higher for FOKs than TOTs.

Experiment 3: Orthographic retrieval

Orthography is another kind of lexical information that is often overlooked in TOT studies, likely due to being closely related to (and difficult to isolate from) phonology. Although phonology and orthography frequently correspond, they can be differentiated as evidenced by homophones (words with the same pronunciation but different spelling) and heteronyms (words with the same spelling but different pronunciation). There are fewer letters than sounds in English, and few words with multiple common spellings (e.g., British English labour versus American English labor) so orthography tends to be very stable unlike phonology where pronunciation varies between accents, speakers, and even slightly with each utterance. TOTs are not limited to speech production, but occur during writing and typing, so it would be folly to neglect retrieval of orthographic information, even if it is less immediate during TOTs.

The two-stage model of lexical retrieval is typically applied to speech production, however there has been debate about how orthographic information may be accessed in this model and others (Caramazza & Miozzo, 1997; Roelofs et al., 1998). If partial access to orthographic information cannot be demonstrated during TOTs, this would suggest that orthographic retrieval is slower or depends on access to other modalities such as phonology. This may be related to the fact that people almost always learn spoken (or signed) language before written language, thus phonology is privileged in how people know words. If orthographic

information is partially accessible during TOTs, then it is more likely that it is retrieved in parallel with other lexical information and somewhat independently. If they are partially accessible as predicted, it may be that they are not perceived as a component of TOT because they are less useful for judging whether target information has been retrieved. There is less variation across orthography due to the constraint of 26 letters in the English alphabet, whereas several phonemes can arise from an individual letter as well as ordered sequences of letters due to coarticulation, so orthographic information should be less helpful. Furthermore, recent work has shown that a variety of orthographic representations can be incorporated to improve performance for memory models at the item level (Osth & Zhang, 2024; Reid et al., 2023). Unless a separate and specialized retrieval process for writing and typing is assumed, orthographic retrieval must occur during word retrieval and therefore should be studied alongside semantics and phonology.

Methods

Participants: I recruited 99 undergraduate students (40 males, M_{age} = 19.18, SD = 2.98) from the University of Manitoba SONA psychology participant pool. Fifty participants completed the study online and 49 completed the study in computer labs at the University of Manitoba under researcher supervision. Students participating in-person were tested in groups of up to 20 participants. All participants received one research participation credit for their introduction to psychology course. To be eligible, participants must not have completed Experiments 1a, 1b, or 2. Because the study involved reading and typing English words, only participants who indicated that they are comfortable reading and writing in English on the departmental pre-screening questionnaire were eligible to participate.

Materials: The materials include 24 cue – target word pairs to be studied. For each target, there is an orthographic associate which is spelled similarly to, but pronounced differently from the target (e.g., "rough" and "dough") as well as three foil words which are not closely related to the target by semantics, phonology, or orthography. A vector analysis verified that the orthographic similarity of target – associate pairs (M = 0.68, SD = 0.03) was greater than their phonological (M = 0.19, SD = 0.13) or semantic similarity (M = 0.11, SD = 0.11). Target – foil pairs also had low orthographic (M = 0.07, SD = 0.07), phonological (M = 0, SD = 0.06), and semantic similarities (M = 0.12, SD = 0.11). The targets and associates are taken from Polich et al. (1983), and these stimuli can be found in Appendix D.

Procedure: The procedure is identical to the format of Experiments 1b and 2. The only difference is the stimuli used and the relationship between the targets and associate words. All text will be presented in black on a white background, with all instructions printed in 18pt Arial font and all stimuli presented in lowercase 21pt Arial font.

After providing informed consent, participants completed four study/test cycles. During the study phase, participants viewed six cue – target pairs (e.g., wake – comb). These pairs were presented, one pair at a time, for 2000ms followed by a 1000ms interstimulus interval with a fixation cross. After studying six pairs, participants completed a cued recall test for the studied targets. On each test trial, one of the studied cues was presented (e.g., wake) and participants reported the paired target (i.e., comb) by typing it and pressing a button to proceed to the next trial. When they were unable to recall the target, they proceeded without typing anything, then indicated on a sliding scale the extent to which they felt they had forgotten the target or had it on the tip of their tongue. The slider recorded responses about the participant's sense of TOT on an invisible 100-point scale, with 0 labelled "I do not know it at all" and 100 labelled "I am certain I

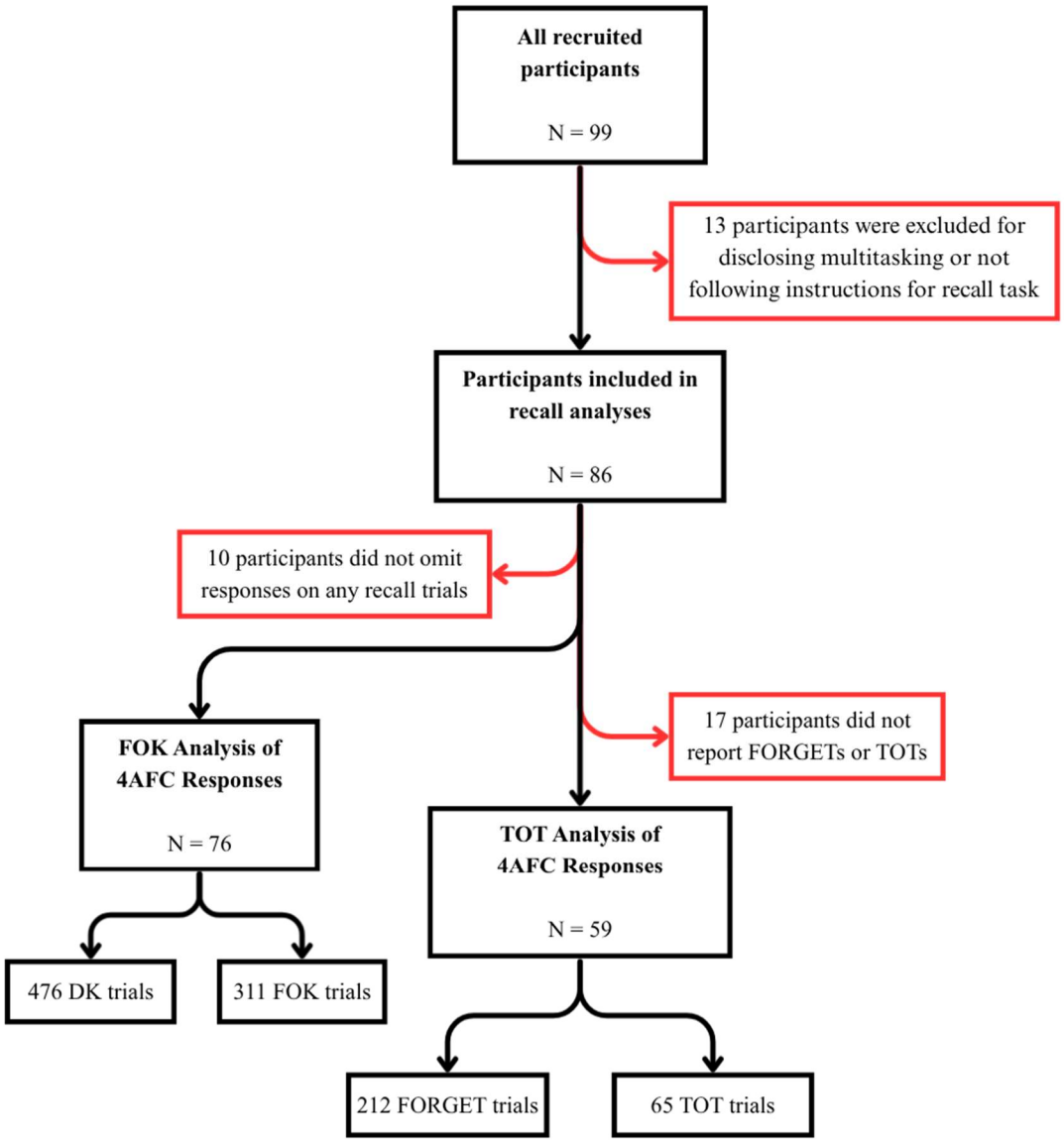
know it, but it is on the tip of my tongue." After indicating their metacognitive judgement by moving the slider in either direction, participants were presented with four words including an unstudied orthographic associate of the target (e.g., tomb) and three unrelated words. They chose one word amongst the four alternatives that they believed best matched the target they were unable to report. These words were presented simultaneously in a randomized order in a single row across the screen. After selecting one of the four words by clicking on it, they had a final opportunity to type the target before proceeding to the next trial. Once all six cue words are presented, the study/test cycle repeated three times, each time with a new list of six cue – target pairs. After completing all four study and test phases, participants were asked if they had been distracted or multi-tasking while completing the experiment, after which they were debriefed and thanked for their participation.

Results

Of the 99 participants, 13 were excluded from the analysis for disclosing distraction and multitasking during the experiment. One of these participants completed the experiment in the computer lab and the other 12 completed the experiment online. This left a sample size of 86 for the following analyses (48 in-person and 38 online participants). A comparison of results from the online and in-person versions of the experiment is found in Appendix F. Figure 16 displays a flowchart of data processing, including the number of participants or trials contributing to each section of the analysis.

Figure 16

Data processing flowchart.



Note. From the full sample of 99 participants, 13 were excluded from the analysis for failing to follow instructions or disclosing multitasking and distractions while completing the experiment. Recall performance was analyzed for the remaining 86 participants. The FOK analysis included all 4AFC trials from participants who did not report a word on every recall trial, with 76 participants contributing 476 DKs and 311 FOKs. The TOT analysis included only trials where participants indicated a strong sense of forgetting or TOT state, with 59 participants contributing 212 FORGETs and 65 TOTs.

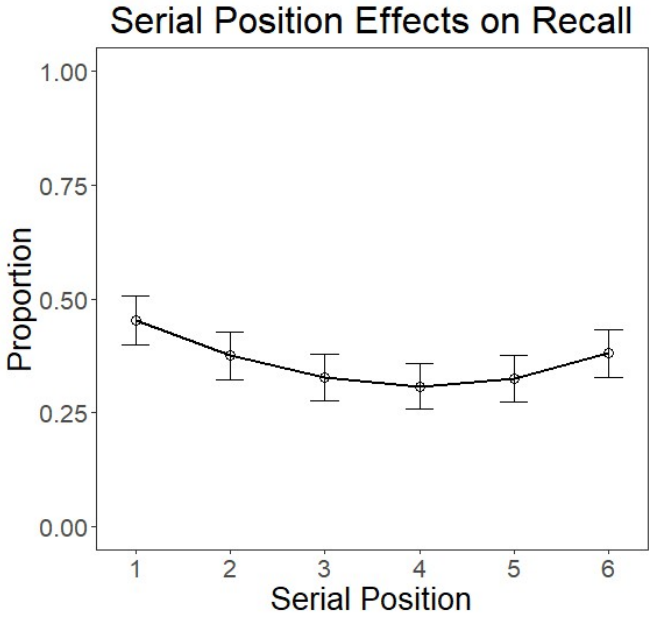
Recall:

Responses on the recall task were scored using the R stringdist package (v09.12; Loo, 2014) and responses with a Levenshtein distance of one away from the target were scored as correct to allow small typos (e.g, “wanf” instead of “wand”) and changes in plurality (e.g., “bones” instead of “bone”). These items were manually checked by the author to ensure that any responses with a distinct word were scored as incorrect even if they were within that window (e.g., “drone” instead of “drove”).

Recall performance is displayed in Figure 17.

Figure 17

Recall performance.



Note. Proportions of recalls by serial position in Experiment 3. Error bars represent one standard error above and below the observed proportions.

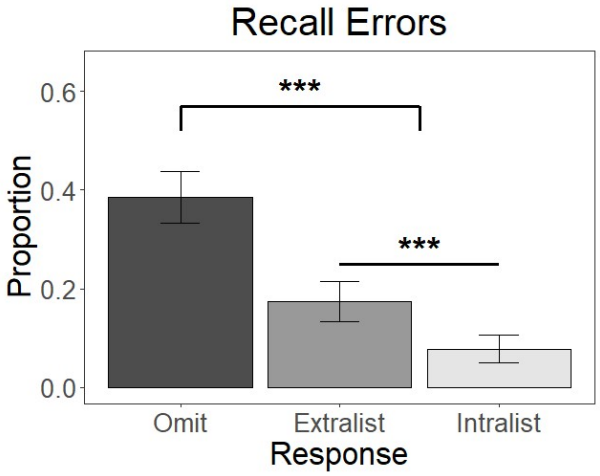
The median and interquartile range for recall responses are presented due to the presence of skewed distributions. The proportion of omitted responses was greatest (*Mdn* = 0.38, *IQR* = 0.32), followed by recalls (*Mdn* = 0.33, *IQR* = 0.32), extralist intrusions (*Mdn* = 0.13, *IQR* = 0.24), and intralist intrusions (*Mdn* = 0.04, *IQR* = 0.08). A one-way within-subjects ANOVA

showed that there was a significant main effect of serial position on recall performance, $F(5, 425) = 4.84$, $p < .001$, with items studied in earlier and later serial positions being recalled better than those in the middle of study lists.

Figure 18 displays the mean error proportions.

Figure 18

Recall errors.



Note. The proportion of omits, extralist intrusions, and intralist intrusions in Experiment 3. Error bars represent one standard error above and below the observed proportions.

The overall error proportions were analyzed using a one-way within-subjects ANOVA, and there was a significant difference between the three error types, $F(2, 170) = 51.88$, $p < .001$. Planned contrast analysis showed that there were significantly more omits than intrusions, $F(1, 85) = 403.68$, $p < .001$, and significantly more extralist than intralist intrusions, $F(1, 85) = 41.16$, $p < .001$.

A vector analysis compared the similarity of extralist intrusions to the correct target. Excluding nonword responses (e.g., “stull”), the mean cosine similarity between extralist

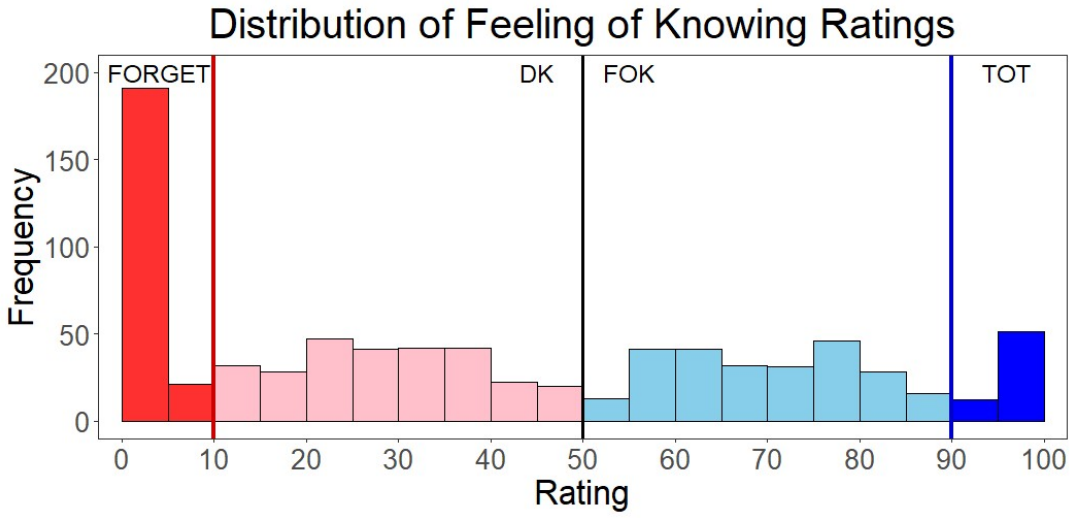
responses and the correct target was highest for semantics (M = 0.147, SD = 0.126), followed by orthography (M = 0.119, SD =0.126), and finally phonology (M = 0.039, SD = 0.124).

TOT Analysis

The histogram of sliding-scale FOK ratings and thresholds for FORGET and TOT trials are found in Figure 19 below.

Figure 19

Sliding-scale responses.



Note. Distribution of TOT ratings on a scale from 0 – "I do not know it at all", to 100 – "I am certain I know it, but it is on the tip of my tongue". Ratings of 10 or below (dark red region) were categorized as FORGET trials and ratings above 90 (dark blue region) were categorized as TOT trials. For the analysis of FOKs, all ratings below 50 were categorized as DKs (full red region) and all ratings above 50 were categorized as FOKs (full blue region).

Seventy-six participants contributed a total of 797 omits where a target could not be recalled, with sliding scale TOT ratings. As in previous experiments, participants used the extreme ends of the scale to produce a large peak at 0 and a smaller peak at 100. Trials with ratings of 10 or below were taken for FORGETs and ratings of 90 or above were taken for TOTs. Using these thresholds, there were 212 FORGET trials (140 responses of exactly zero) from 49

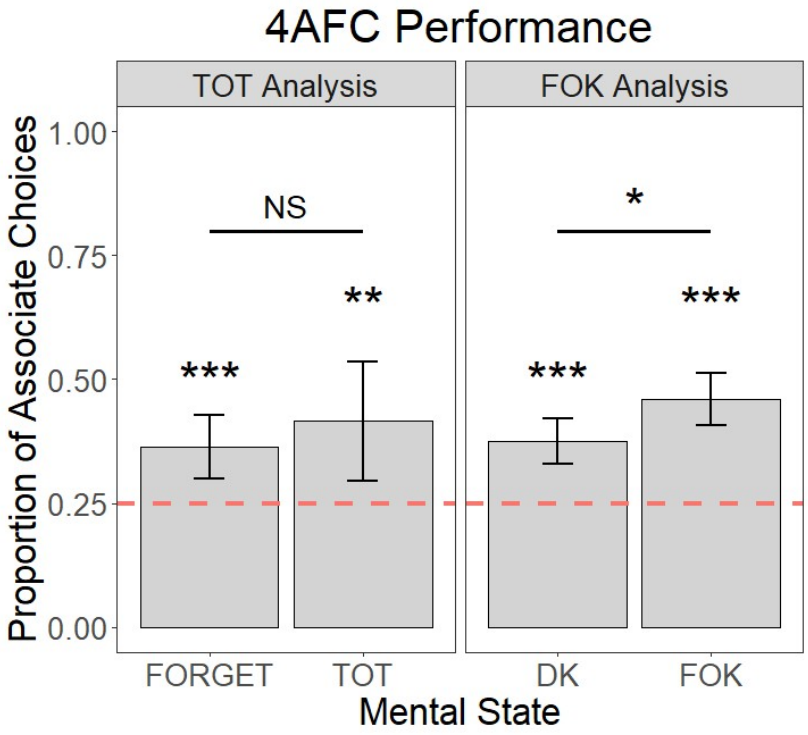
participants, and 65 TOT trials (35 responses of exactly 100) from 28 participants. Across both types of trials, responses from 59 of the total 86 participants contributed to the following analyses.

4AFC:

Figure 20 contains a bar graph depicting associate choice proportions during TOT and FORGET trials.

Figure 20

Performance on the 4AFC task.



Note. Associate choices occurred significantly above the random chance level of 0.25 (red line) for all trial types. The proportion of associate choices was not significantly higher during TOT trials compared to FORGET trials, but was significantly higher during FOK trials compared to DK trials. Error bars indicate the 95% CI for the observed proportions.

A two-sided test of proportions showed that participants did not perform significantly better on the 4AFC task during TOT trials than FORGET trials, $z = 0.760$, $p = .447$. Given a 0.25 probability of guessing the associate at random, two-sided tests of proportions also showed that the proportion of associate choices during TOT trials (0.42) was significantly above chance, $z = 3.07$, $p = .002$, as was the proportion during FORGET trials (0.36), $z = 3.81$, $p < .001$. Appendix G contains an analysis of 4AFC performance with additional exclusion criteria removing participants who performed very poorly on the recall task.

Final recall:

Participants reported a target at the end of 259 out of 797 trials (32%) where a response was omitted on the initial recall task, with a hit rate of 0.18. Targets were reported following 18 out of 65 TOT trials (28%) with a hit rate of 0.22, and 60 out of 212 FORGET trials (28%) with a hit rate of 0.18. Of the 165 extralist errors reported during final recall, 43 came from FORGET trials and 12 came from TOT trials. Because there were only 18 final recall attempts following TOT trials, the relative proportions of correct recalls are not compared statistically. Targets were reported following 131 out of 325 associate choices (40%) with a correct recall rate of 0.29, compared to 128 out of 472 foil choices (27%) with a correct recall rate of 0.07. The proportion of recalls was significantly higher following associate choices than foil choices, $z = 4.59$, $p < .001$. Of the 165 extralist errors during the final recall task, the orthographic associate was reported as a target 22 times and a foil was reported 30 times. Vector analysis showed that extralist intrusions were most similar to the target by phonology regardless of whether the responses followed a TOT or FORGET trial. Unlike prior experiments, this was not driven by participants reporting the associate word. Removing trials where the associate word from the 4AFC task was reported (or removing both associate and foil reports) had no impact on the

relative similarities for each modality, only decreasing orthographic similarity by approximately 0.05 across all trial types. The cosine similarities between extralist responses and targets are found in Table 3 below. The table includes trials with associate and foil reports because removing them did not change any relationships between trial types or modalities, and to avoid reducing the number of final recall observations from TOT trials.

Table 3

Mean cosine similarity of extralist responses and targets on the final recall task.

	Semantic		Phonological		Orthographic	
	M	*SD*	*M*	*SD*	*M*	*SD*
All Trials	0.12	0.11	0.11	0.18	0.34	0.30
TOT	0.13	0.11	0.14	0.23	0.48	0.29
FORGET	0.09	0.09	0.06	0.14	0.24	0.28

Note. Extralist errors were most similar to the target's orthography regardless of trial type. Orthographic similarity also increased the most between FORGET and TOT trials, followed by phonological and semantic similarity.

FOK Analysis

Figure 19 displays the distribution of ratings with the categories and the criterion point for DK and FOK trials. The following analyses compare performance between DK trials (TOT ratings below 50) and FOK trials (TOT ratings above 50). There were a total of 476 DK trials, 311 FOK trials, and 10 excluded trials where participants returned the slider to exactly 50.

4AFC:

See Figure 20 for a bar graph displaying the associate choice proportions for FOK and DK trials. A two-sided test of proportions showed that the proportion of associate choices was significantly greater during FOK than during DK trials, $z = 2.336$, $p = .019$. Given a 0.25

probability of guessing the associate at random, two-sided tests of proportions also showed that the proportion of associate choices during FOK trials (0.46) was significantly above chance, $z = 8.545$, $p < .001$, as was the proportion during DK trials (0.38), $z = 6.351$, $p < .001$.

Final recall:

A target was reported following 119 out of 311 FOK trials (38%) with a correct recall rate of 0.24, and following 130 out of 476 DK trials (27%) with a correct recall rate of 0.15. A two-sided test of proportions showed that the proportion of recalls was not significantly higher for FOK trials than for DK trials, $z = 1.80$, $p = .073$.

Discussion

Performance on the recall task was consistent with the previous experiments with omits occurring significantly more often than intrusions, and extralist intrusions being more common than intralist intrusions. At least one TOT was reported by 33% of participants which is comparable to the 38% from Experiment 1b, but participants reported fewer TOTs in general with an overall rate of 0.76 TOT trials per participant. The rate of associate choice was significantly above chance levels for both FORGET and TOT trials, and there was no significant difference between the two trial types. This outcome was very similar to the semantic case in Experiment 1b, standing in contrast to the strong effect of phonology in Experiment 2. Although there was no significant difference in associate choice proportions during TOTs and FORGETs, they were significantly more likely to choose the associate during FOKs than DKs.

There were some differences in final recall performance relative to previous experiments. Participants were equally likely to report a target following a TOT trial or a FORGET trial, and the accuracy of these reports were similar. The percentage of final recall attempts following

FORGET trials was consistent with previous experiments, so the difference appears to come from diminished attempts during TOT trials. Associate choices were positively correlated with increased final recall attempts and significantly higher accuracy compared to foil choices, which is consistent with the previous experiments. Another consistent finding was the prevalence of orthographic similarity between extralist intrusions and targets. In Experiments 1b and 2, participants reporting the semantic or phonological associate from the 4AFC task as an extralist intrusion led to similarity of the tested modality overtaking orthography. In both cases, removing those trials uncovered a bias towards orthography in extralist intrusions. Thus, it is unsurprising that the orthographic dominance in Experiment 3 was not caused by the reporting of the orthographic associate from the 4AFC task. The fact that semantic similarity often fell between orthographic and phonological similarity for extralist intrusions in Experiments 1b, 2, and 3 supports the successful manipulation of the stimuli to separate phonological and orthographic information (i.e., orthography and phonology of extralist errors correlated less with one another than with semantics).

The bias towards orthography in extralist intrusions was consistent across three experiments, which makes it more surprising that orthographic associates were not chosen significantly more often during TOTs than FORGETs. It is possible that participants are unconcerned with orthography when attempting to recognize which word is related to the unrecalled target, yet give orthography significant weight when recalling words. Identifying the target during 4AFC was associated with a significant increase in performance on the final recall task which shows that an orthographic cue was effective. The TOT and FOK analyses corresponded closely in Experiments 1b and 2, and the associate choice proportions from the TOT and FOK analyses were very similar in Experiment 3 as well. Considering all of these

factors, it seems likely that the lower incidence of TOTs in this experiment simply left too few trials to detect a significant effect for orthography during TOTs. The tendency to report orthographically close extralist intrusions was unexpected but the mixed results on the 4AFC task are less surprising. It appears that participants had some access to orthographic information (and possibly greater access during TOTs compared to FORGETs) but did not focus on orthography when making their choice on the 4AFC task. This makes sense when considering that orthography has lower total variability than phonology, and the two are frequently redundant.

Experiment 4: Competing forced choice

I systematically examined participants' access to different modalities of lexical information while experiencing TOTs in Experiments 1a – 3. In Experiment 4, I aimed to broaden the scope while maintaining close control of the possible responses as in Experiments 1a – 3. This experiment placed these modalities in competition with one another, looking beyond their individual influence at the time of a TOT occurrence to test how they may interact during lexical retrieval. Suppose that some phonological and some orthographic information has each been retrieved at the time of a TOT occurring. If a participant chooses to neglect available orthographic information and select a homophone associate, this would indicate that phonological information is prioritized over orthographic information when resolving TOTs. By including semantic, orthographic, and phonological associates which are distinct from the target in each of the other modalities, the 4AFC task will require participants to select the closest fit relying on a subset of lexical information over information from the other two modalities. Experiments 1a – 3 provided data on the retrieval of semantic, phonological, and orthographic information and how those modalities influenced choices during TOTs. Experiment 4 extended

these results to test which information has priority during TOTs. This will help provide a better understanding of how retrieval unfolds when the retrieved information is insufficient to report a candidate, or when multiple modalities support conflicting candidates.

Methods

Participants: I recruited 99 undergraduate students (42 males, M_{age} = 19.03, SD = 2.29) from the University of Manitoba SONA psychology participant pool. Fifty participants completed the study online and 49 completed the study in computer labs at the University of Manitoba under researcher supervision. Students participating in-person were tested in groups of up to 20 participants. All participants received one research participation credit for their introduction to psychology course. To be eligible for participation, participants must not have completed Experiments 1a, 1b, 2, 3, or 4. Because the study involves reading and typing English words, only participants who indicated that they are comfortable reading and writing in English on the departmental pre-screening questionnaire were eligible to participate.

Materials: The materials consist of 24 target words, each with a cue, a semantic associate, a phonological associate, an orthographic associate, and a foil. The cue and foil words are not related to the target by semantics, phonology, or orthography, and the similarity between each associate and the target is limited to a singular modality of lexical information. A vector analysis confirmed that target – foil pairs had low cosine similarity for all three modalities, and that each of the three associate types were highly similar to the target in its respective modality but not the others. The phonological associates did show some orthographic similarity to the targets, but to a lesser degree than the orthographic associates. The cosine similarities from this vector analysis are displayed in Table 4.

Table 4

Mean cosine similarities for target – foil and target – associate pairs.

	Semantic		Phonological		Orthographic	
	M	*SD*	*M*	*SD*	*M*	*SD*
Foil	0.12	0.12	0.02	0.08	0.11	0.13
Semantic Associate	**0.36**	**0.19**	0.00	0.04	0.09	0.09
Phonological Associate	0.13	0.10	**0.64**	**0.28**	0.36	0.17
Orthographic Associate	0.14	0.13	0.13	0.15	**0.71**	**0.06**

Note. Foils were not highly similar to the target in any modality, and associates were highly similar to the target in their respective modality but not the others. The phonological associates were moderately similar to the target in orthography.

The targets and associates were assembled from Pexman et al. (2001) and Polich et al. (1983) by identifying homophones and near-homophone rhyming pairs for which an orthographic associate could be generated, or orthographic associates for which a phonological associate could be generated. The complete set of stimuli for this experiment can be found in Appendix E below.

Procedure: The procedure is identical to the format of Experiments 1b - 3. The only difference is the stimuli used and the relationship between the targets and associate words. All text was presented in black on a white background, with all instructions printed in 18pt Arial font and all stimuli presented in lowercase 21pt Arial font.

After providing informed consent, participants completed four study/test cycles. During the study phase, participants viewed six cue – target pairs (e.g., hurry – flower). These pairs were presented, one pair at a time, for 2000ms followed by a 1000ms interstimulus interval with a fixation cross. After studying six pairs, participants completed a cued recall test for the studied targets. On each test trial, one of the studied cues was presented (e.g., hurry) and participants

reported the paired target (i.e., flower) by typing it and pressing a button to proceed to the next trial. When unable to recall the target, they proceeded without typing anything, then indicated on a sliding scale the extent to which they felt they had forgotten the target or had it on the tip of their tongue. The slider recorded responses about the participant's sense of TOT on an invisible 100-point scale, with 0 labelled "I do not know it at all" and 100 labelled "I am certain I know it, but it is on the tip of my tongue." After indicating their metacognitive judgement by moving the slider in either direction, participants were presented with four words including three unstudied associates of the target and one unrelated word. The associates were related to the target through semantics (e.g., petal), phonology (e.g., flour), and orthography (e.g., lower). They chose one word amongst the four alternatives which they felt best matched the target they were unable to report. These words were presented simultaneously in a single row with a randomized order. After selecting one of the four words by clicking on it, they had a final opportunity to type the target before proceeding to the next trial. Once all six cue words were presented, the study/test cycle repeated three times with new lists of six cue – target pairs. After completing all four study and test phases, participants were asked if they had been distracted or multi-tasking while completing the experiment, after which they were debriefed and thanked for their participation.

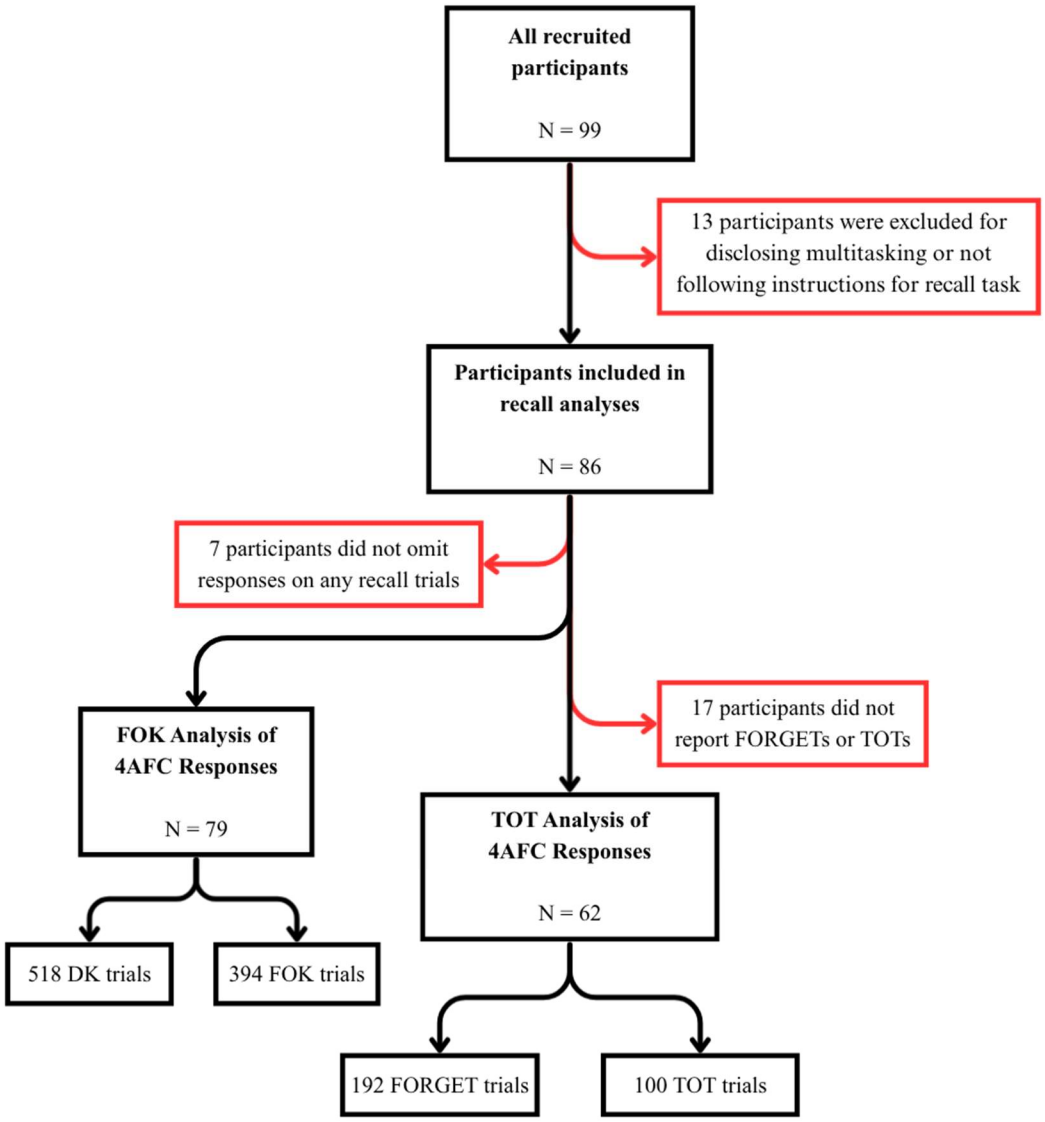
Results

Of the 99 participants, 13 were excluded from the analysis for disclosing distractions and multitasking. One of these participants completed the experiment in the computer lab and the other 12 completed the experiment online. This left a sample size of 86 for the following analyses (48 in-person and 38 online participants). A comparison of results from the online and in-person versions of the experiment is found in Appendix F. Figure 21 displays a flowchart of

data processing, including the number of participants or trials contributing to each section of the analysis.

Figure 21

Data processing flowchart.



Note. From the full sample of 99 participants, 13 were excluded from the analysis for failing to follow instructions or disclosing multitasking and distractions while completing the experiment. Recall performance was analyzed for the remaining 86 participants. The FOK analysis included all 4AFC trials from participants who did not report a word on every recall trial, with 79

participants contributing 518 DKs and 394 FOKs. The TOT analysis included only trials where participants indicated a strong sense of forgetting or TOT state, with 62 participants contributing 192 FORGETs and 100 TOTs.

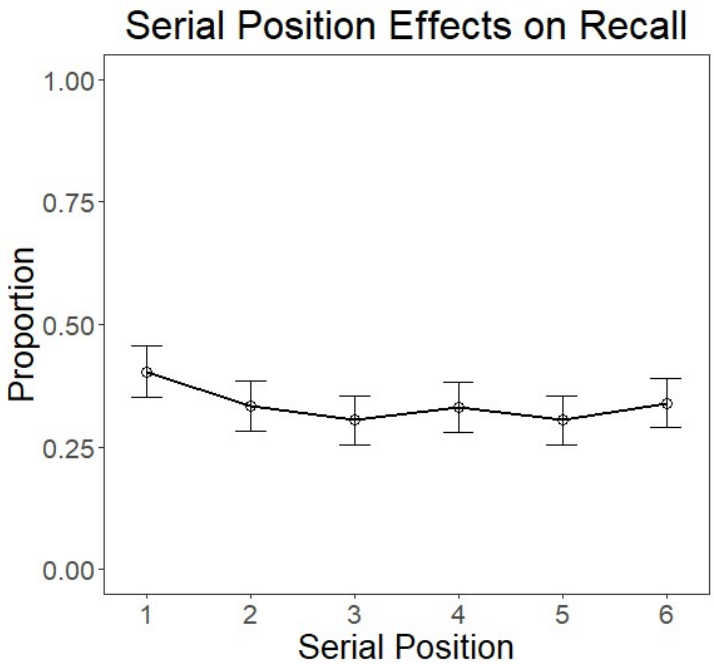
Recall:

Responses on the recall task were scored using the R stringdist package (v0.9.12; Loo, 2014) and responses with a Levenshtein distance of one from the target were scored as correct to allow small typos (e.g., “boygh” instead of “bough”) and changes in plurality (e.g., “flowers” instead of “flower”). These items were manually checked by the author to ensure that any responses with a distinct word were scored as intrusions even if they were within that window (e.g., “tear” instead of “wear”).

Recall performance is displayed in Figure 22 below.

Figure 22

Recall performance.



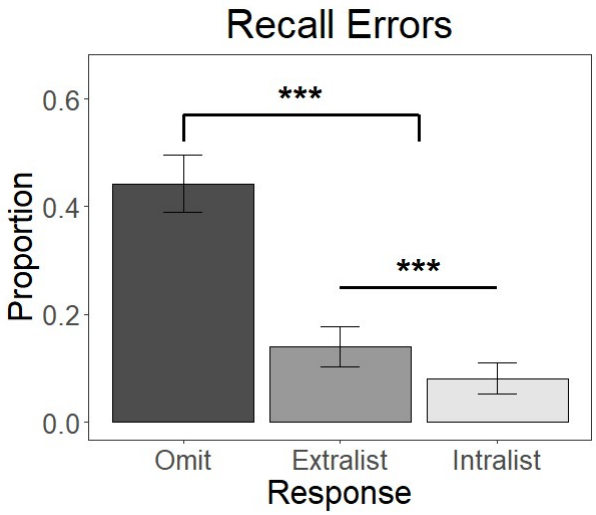
Note. The proportion of correct recalls by serial position in Experiment 4. Error bars represent one standard error above and below the observed proportions.

The median and interquartile range for recall responses are presented due to skewed distributions of intrusion errors. The proportion of omitted responses was greatest (*Mdn* = 0.46, *IQR* = 0.28), followed by recalls (*Mdn* = 0.29, *IQR* = 0.25), extralist intrusions (*Mdn* = 0.08, *IQR* = 0.13), and intralist intrusions (*Mdn* = 0.04, *IQR* = 0.13). A one-way within-subjects ANOVA showed that there was a significant main effect of serial position on recall performance, $F(5, 425) = 2.50$, $p = .030$, with targets studied in early serial positions being recalled better than those in the middle and end of the list.

Figure 23 displays the mean error proportions.

Figure 23

Recall errors.



Note. The proportion of omits, extralist intrusions, and intralist intrusions in Experiment 4. Error bars represent one standard error above and below the observed proportions.

The overall error proportions were analyzed using a one-way within-subjects ANOVA, and there was a significant difference between the three error types, $F(2, 170) = 87.83$, $p < .001$. Planned contrast analysis showed that there were significantly more omits than intrusions, $F(1,$

85) = 742.84, $p < .001$, and significantly more extralist than intralist intrusions, $F(1, 85) = 17.67$, $p < .001$.

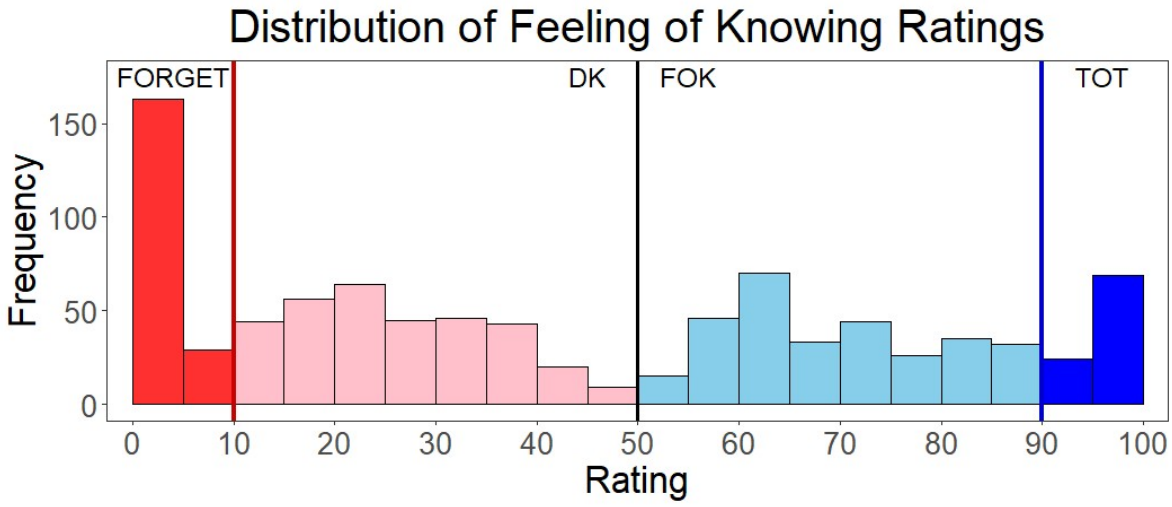
Vector analysis was used to determine the similarity of extralist intrusions and correct targets. Excluding nonword responses (e.g., “tatse”), the cosine similarities between extralist responses and targets was greatest for orthography ($M = 0.163$, $SD = 0.182$), followed by semantics ($M = 0.134$, $SD = 0.105$) and phonology ($M = 0.084$, $SD = 0.219$).

TOT Analysis

The histogram of sliding-scale FOK ratings and thresholds for FORGET and TOT trials are displayed in Figure 24.

Figure 24

Sliding-scale responses.



Note. Distribution of TOT ratings on a scale from 0 – “I do not know it at all”, to 100 = “I am certain I know it, but it is on the tip of my tongue”. Ratings of 10 or below (dark red region) were categorized as FORGET trials and ratings of 90 or above (dark blue region) were categorized as TOT trials. For the analysis of FOKs, all ratings below 50 were categorized as DKs (full red region) and all ratings above 50 were categorized as FOKs (full blue region).

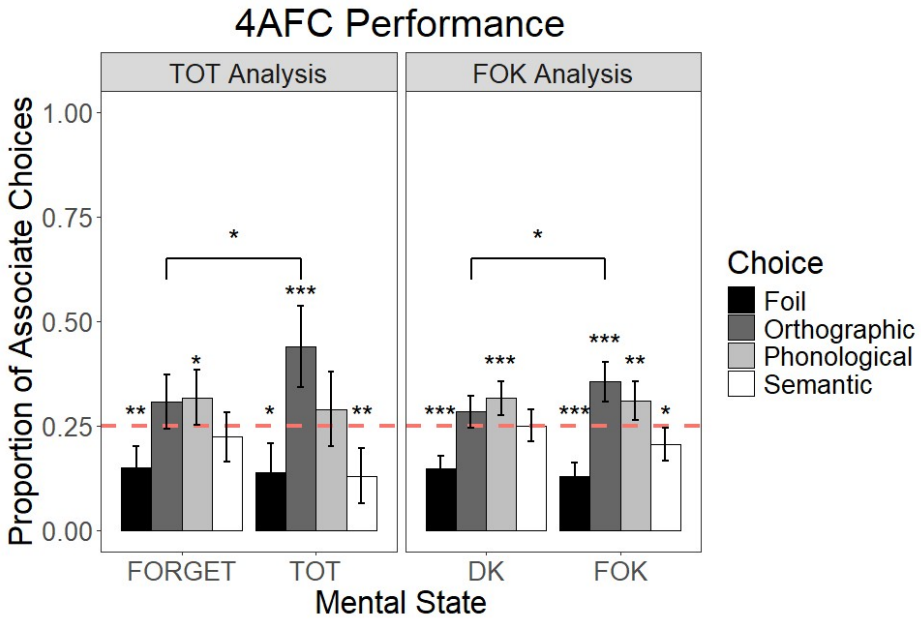
Seventy-nine participants contributed a total of 913 omits when a target could not be recalled, with sliding scale TOT ratings. As in the previous experiments, participants produced a larger peak at 0 and a smaller peak at 100. Trials with ratings of 10 or below were taken for FORGETs and ratings of 90 or above were taken for TOTs. Using these thresholds, there were 192 FORGET trials (120 responses of exactly zero) from 51 participants, and 100 TOT trials (52 responses of exactly 100) from 31 participants. Across both trial types, responses from 62 of the total 86 participants contributed to the following analyses.

4AFC:

Figure 25 displays the proportions of associate and foil choices during FORGET and TOT trials.

Figure 25

Performance on the 4AFC task.



Note. Orthographic associate choices occurred significantly more often during TOT trials compared to FORGET trials, and during FOK trials compared to DK trials. Error bars indicate the 95% CI for the observed proportions.

A two-sided test of proportions showed that participants were more likely to select the orthographic associate on the 4AFC task during TOT trials (0.44) than FORGET trials (0.31), z = 2.25, p = .024. There was no significant difference between TOT and FORGET trials in the proportion of foil choices, z = 0.253, p = .801, phonological associate choices, z = 0.487, p = .627, or semantic associate choices, z = 1.94, p = .053. Given a 0.25 probability of randomly choosing one of the four alternatives, the proportion of orthographic associate choices was not significantly above chance during FORGET trials, z = 1.83, p = .067, but was above chance during TOT trials, z = 4.39, p < .001. The proportion of phonological associate choices was significantly above chance during FORGET trials (0.32), z = 2.17, p = .030, but not during TOT trials (0.29), z = 0.92, p = .356. The proportion of semantic associate choices was not significantly different from chance during both FORGET trials (0.22), z = 0.83, p = .405, but was significantly below chance during TOT trials (0.13), z = 2.77, p = .006. Finally, the proportion of foil choices was significantly below chance level during both FORGET trials (0.15), z = 3.17, p = .002, and TOT trials (0.14), z = 2.54, p = .011. Appendix G contains an analysis of 4AFC performance with additional exclusion criteria removing participants who performed very poorly on the recall task.

Final recall:

There were 384 out of 913 trials (42%) where a target was not initially recalled, but was reported following completion of the 4AFC task, and the proportion of correct recalls at this stage was 0.43 overall. Contrasting by mental state, targets were reported following 67 out of 100 TOT trials (67%) with a correct recall rate of 0.42, and 72 out of 192 FORGET trials (38%) with a correct recall rate of 0.40. The proportion of recalls was not significantly higher for targets reported following TOT trials and FORGET trials, z = 0.18, p = .856. Targets were reported

following 31 out of 384 foil choices, but there were zero correct recalls. A chi-square test of independence showed that there was no significant difference in the proportions of correct recalls following semantic (0.47), phonological (0.51), and orthographic associate choices (0.44), χ^2 (2, N = 353) = 1.15, p = .563.

Table 5 contains the cosine similarities between all 187 extralist responses and their targets, including reported associates and foils.

Table 5

Mean cosine similarity of extralist responses and targets on the final recall task.

	Semantic		Phonological		Orthographic	
	M	*SD*	*M*	*SD*	*M*	*SD*
All Trials	0.18	0.15	0.26	0.34	0.40	0.28
TOT	0.17	0.14	0.17	0.31	0.45	0.30
FORGET	0.21	0.14	0.20	0.33	0.33	0.29

Note. Extralist errors were most similar to the target's orthography regardless of trial type. Orthographic similarity increased between FORGET and TOT trials, whereas semantic and phonological similarity slightly decreased.

Of the 187 extralist errors reported during the final recall task, 38 came from FORGET trials and 34 came from TOT trials. Vector analysis showed that extralist intrusions were most similar to the target by orthography regardless of whether the responses followed a TOT or FORGET trial. Participants reported their choice from the 4AFC task as a recalled target 149 times, and this consisted of 21 foils, 36 semantic associates, 42 phonological associates, and 50 orthographic associates. This left only 38 extralist errors that were not items from the 4AFC task, including six from FORGET trials and four from TOT trials. Like Experiment 3, removing associate reports (or associate and foil reports) from the vector analysis did not change any

relationships between the modalities, but reduced the cosine similarities relatively evenly across all trial types.

FOK Analysis

Figure 24 displays the distribution of ratings with the criterion point for DK and FOK trials. The following analyses compare performance between DK trials (TOT ratings below 50) and FOK trials (TOT ratings above 50). There were a total of 518 DK trials, 394 FOK trials, and 1 excluded trial where a participant returned the slider to exactly 50.

4AFC:

See Figure 25 for a bar graph displaying the 4AFC proportions for FOK and DK trials. A two-sided test of proportions showed that participants were more likely to select the orthographic associate on the 4AFC task during FOK trials (0.36) than DK trials (0.28), $z = 2.30$, $p = .021$. There was no significant difference between DK and FOK trials in the proportion of foil choices, $z = 0.83$, $p = .408$, phonological associate choices, $z = 0.22$, $p = .823$, or semantic associate choices, $z = 1.61$, $p = .107$. During DK trials, the proportion of foil choices (0.15) was significantly below chance, $z = 5.33$, $p < .001$, and the proportion of phonological associate choices (0.32) was significantly above chance, $z = 3.50$, $p < .001$. During FOK trials, the proportion of foil choices (0.13) was significantly below chance, $z = 5.53$, $p < .001$, as was the proportion of semantic associate choices (0.21), $z = 2.04$, $p = .042$. Finally, the proportion of orthographic associate choices (0.36) was significantly above chance, $z = 4.83$, $p < .001$, as was the proportion of phonological associate choices (0.31), $z = 2.73$, $p = .006$.

Final recall:

A target was reported following 209 out of 394 FOK trials (53%) with a correct recall rate of 0.44, and following 175 out of 518 DK trials (34%) with a correct recall rate of 0.43. A two-sided test of proportions showed that the proportion of recalls was not significantly higher for FOK trials than for DK trials, $z = 0.02$, $p = .982$.

Discussion

Recall performance was consistent with the preceding experiments, with omits occurring more than intrusions, and extralist intrusions occurring more than intralist intrusions. At least one TOT was reported by 31% of participants, with an overall rate of 1.01 TOT trials per participant. During FORGET trials, only the phonological associate was chosen above chance levels. During TOT trials, only the orthographic associate was chosen above chance levels, and more often than it was selected during FORGET trials. These outcomes were further supported by the FOK analysis, and the only difference was that the phonological associate was chosen above chance levels during both DK and FOK trials. Choosing any of the three associates was associated with a similar benefit in the final recall task and, like the previous experiments, extralist errors during the recall and final recall tasks were closest to the target by orthography. Final recall performance improved above the preceding experiments, with a higher rate of recall attempts and a much higher accuracy. This offers further evidence that associate words during the 4AFC task act as additional cues for target information, and viewing three cues for different modalities of information is the best explanation for the stark increase in final recall accuracy.

The 4AFC task revealed that semantic similarity was neglected for phonological and orthographic similarity across all mental states. The results of Experiments 1a and 1b show that this is not due to a lack of access to target semantics, so there must be a preference for the communicable form of a word that shapes the guessing behaviour. The TOT and FOK analyses

both showed that only orthographic associate choices differentiated behaviour between low and high recollection states. The sliding scale ratings appear to have accurately reflected how close participants were to recalling the target exactly as presented at study (i.e., in written form). Thus higher ratings were most associated with choices that reflect their proximity to the specific form of the word that was presented. The consistently above-chance associate choice proportions in Experiments 1b, 2, and 3 show that participants must be accessing all three modalities, but they are likely processing orthography more deeply due to directly perceiving it at study. Future studies could promote deeper processing of other modalities by having participants produce their own semantic or phonological associates (i.e., perform free association or identify rhyming words) at study. If these tasks could push 4AFC performance towards the semantic or phonological associate, it would support the conclusion that participants prefer the richest information in memory.

Overall this confirms what Experiments 1b, 2 and 3 show, which is that people rarely have meaning on the tip of their tongues unless they are cued with semantic information as in Experiment 1a. Participants reliably demonstrated some level of access to semantic, phonological, and orthographic information about an unrecalled word even when they felt it was completely forgotten. Participants may have attended less to semantics when studying unrelated word pairs than when studying individual words, which could explain the differences in associate choice between Experiment 1a compared to Experiments 1b and 4. Likewise, the differences between Experiment 2 and Experiment 4 could be explained by orthography being encoded better than phonology, or by a preference for an associate that is closest to the target in the form it was studied which would benefit from increased triangulation in Experiment 4. The relationships between orthography and phonology across Experiments 2 – 4 were likely shaped

by reading processes, for example phonological encoding of orthographic information through the inner voice (see Baddeley & Lewis, 2017 for a discussion of articulatory, acoustic, and visual coding during reading).

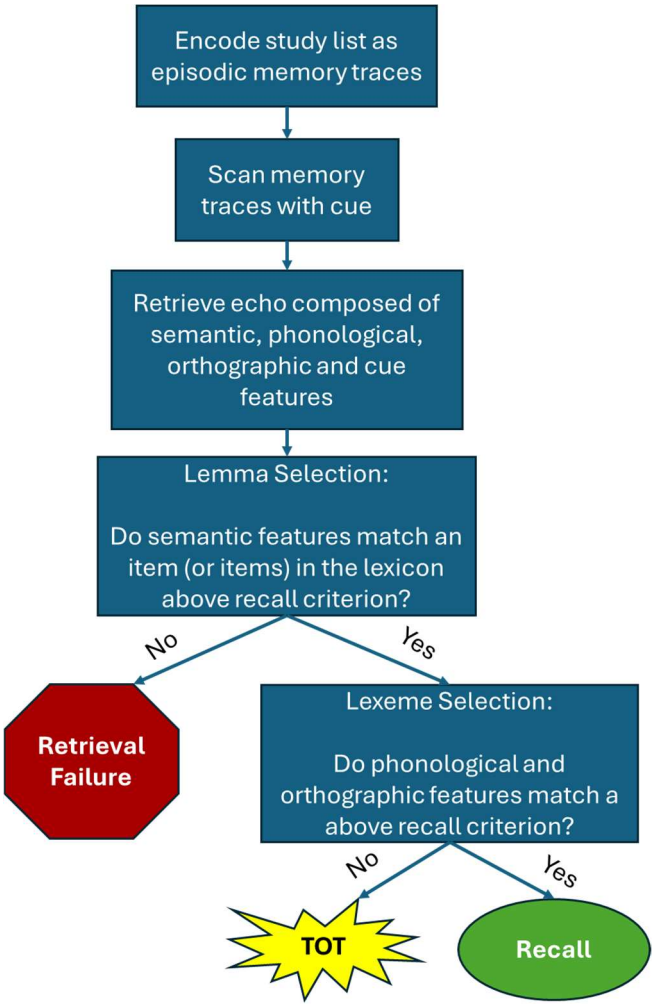
Theoretical Work

The experiments above establish a body of facts about behaviour during TOTs. However, my ultimate goal is to provide an articulate account of performance that accommodates these facts. In particular, my model for simulating TOT performance explores the connections between semantic, phonological, and orthographic information, how these three modalities contribute to heuristic judgements of the probability of successful recall, and how they may be brought into the searching procedure to aid resolution of TOTs. To gain traction, I use an instance model similar to MINERVA 2 and others (Hintzman, 1984; Nosofsky, 1986; Shiffrin & Steyvers, 1997) in which study is simulated by encoding items in episodic memory as traces with some noise. When a retrieval probe is presented to memory, it elicits an echo of memory contents that contains information from each trace, with the contribution of each trace weighted according to its similarity to the probe (Hintzman, 1986). The model possesses a large lexicon of 52,176 English words, aiming to approximate the vocabulary size of L1 English speakers (Miralpeix, 2019). Traditionally, random representations have been used for studied items. To make trial-level predictions using real word stimuli, more sophisticated representations are needed to capture the multi-faceted relationships between words in natural language. Each word in the vocabulary has vector-based representations for semantics from Latent Semantic Analysis (Landauer et al., 1998; Landauer & Dumais, 1997), phonology from Pincelate (Parrish, 2017), and orthography from SERIOL2 (Reid et al., 2023; Whitney & Marton, 2013).

To simulate TOTs in the context of a paired associate learning task, study is simulated by encoding the summed multi-modal representations for cue – target pairs with some forgetting to produce a single representation of their association in episodic memory. Probing memory with the representation for a cue word elicits an echo of the studied associate pairs (i.e., a weighted sum of activated traces) which is not identical to the representation of any one word or pairing but can be used to select the best fitting word within the lexicon. The echo is composed of information of three distinct modalities (semantic, phonological, and orthographic) which is retrieved from memory, in keeping with direct-access theories. The processing of these features to produce a word occurs in two stages supporting ideation (lemma selection) and then word production (lexeme selection). To perform lemma selection, the retrieved semantic information in the echo is accessed first to identify a set of candidate words. If the semantic features are sufficient to select a lemma (or multiple lemmas), indicating that an existing target word is known, only then will phonological and orthographic information retrieved in the echo be compared to the corresponding lexeme(s). Lexeme selection requires matching the elicited phonological and orthographic information in the echo to the phonological and orthographic information associated with a lemma selected in the previous stage. An outline of this theory can be found in Figure 26.

Figure 26

Flowchart of the model's processes for recall, TOTs, and retrieval failure.



Note. An outlined theory of TOTs which coordinates two-stage lexical access, partial feature retrieval, and metacognitive monitoring of lexical retrieval.

To explore the relationships between modalities, simulated results will be compared to behavioural data from each of the above experiments. Experiments 1b, 2, and 3 are concerned with the accessibility of semantic, phonological, and orthographic information, respectively. The model's learning rate (*L)* and response threshold (*T)* parameters will be fit to the empirical data for each experiment individually so that comparisons can be made between each set of results and a version of the model fit to those specific results.

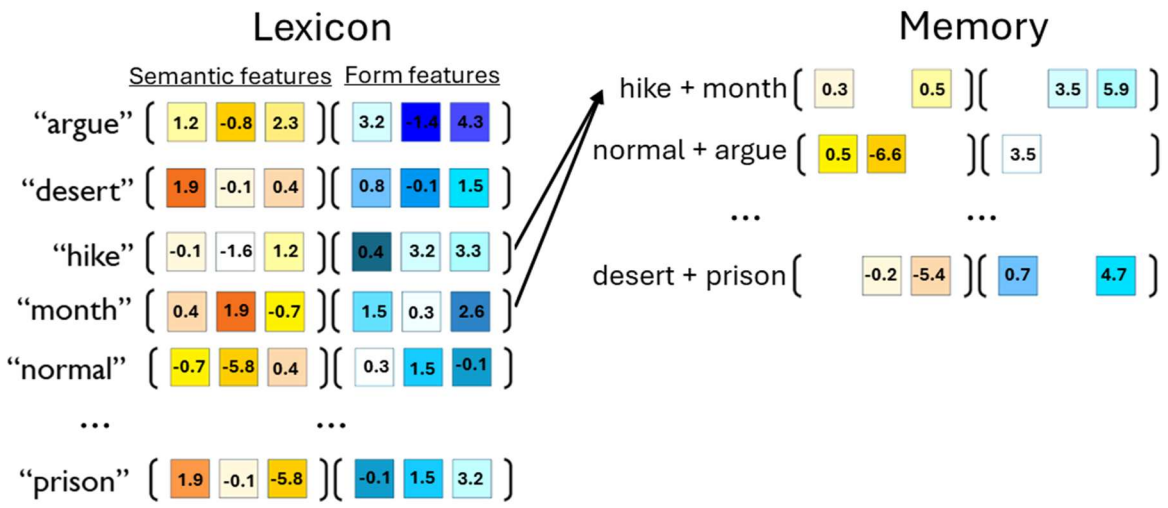
Computational Theory

The word representations used by the model are comprised of vectors representing lexical information. These include vector representations for semantics using LSA (Landauer et al., 1998), phonological representations using Pincelate (Parrish, 2017), and orthographic representations using SERIOL2 (Reid et al., 2023; Whitney & Marton, 2013). LSA encodes the semantic relationships between words based on how they co-occur in a word by document matrix. Pincelate encodes phonological relationships using sequences of phonetic feature pairs from ARPAbet transcriptions of items. Finally, SERIOL2 encodes orthographic relationships using representations for letter pairs that are weighted by their proximity within the word. LSA and SERIOL2 use singular value decomposition (SVD) and Pincelate uses principal component analysis (PCA) to reduce the information about word/sound/letter co-occurrences to 300 latent dimensions. These representations make up the lexicon, or semantic memory, of the model. To simulate a two-stage retrieval process with lemma selection leading to lexeme selection, semantic information is encoded separately from phonological and orthographic information. The representations for each item in the lexicon are composed of a 300-dimensional semantic vector and a 300-dimensional word form vector which is the sum of the phonological and orthographic vectors.

Figure 27 contains a visualization of the model's lexicon and process of memory encoding.

Figure 27

Encoding traces to memory.

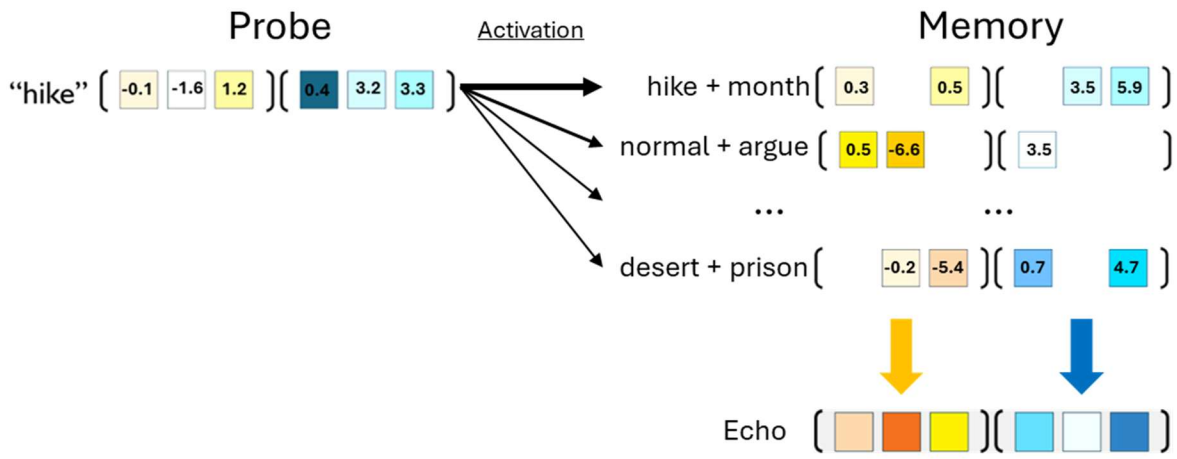


Note. Words are represented by 300 semantic features and 300 combined phonological and orthographic features, but only three dimensions are displayed. Only six of the 52,176 representations in the lexicon are displayed.

To simulate study, the representations for target words (or cue – target pairs) are encoded as traces in an episodic memory matrix, *M,* at a learning rate, *L,* which dictates the probability with which each element is copied from a word's vector representation (all other values are zero). For example, using vectors with 300 dimensions, setting $L = 0.25$ would result in copying on average 75 of the 300 dimensions into memory. Figure 28 contains a visualization of how memory is probed for information in the model.

Figure 28

Probing memory.



Note. Each trace in memory is activated in proportion to its similarity with the probe. The echo contains information elicited by the probe, which is the sum of memory traces weighted by their activation.

In Experiment 1a, participants were presented with semantic clues which varied from one to five words. Probes for the model were created by removing structure words (e.g., to, of) before taking the sum of vectors for each word in the clue. For example, the clue "to set aside" was transformed into a single semantic probe, p = "set" + "aside". In Experiments 1b – 4, the cue – target pairs were encoded by first summing the representations for each item. For example, the pair "limit – beard" was stored as a single memory trace, m_i = "limit" + "beard". The activation of the model depends on the cosine similarity, s, between the probe and each trace in memory which are activated in parallel.

The formula for computing cosine similarity is:

$$s = \frac{\sum_{j=1}^{j=d} p_j \times M_{ij}}{\sqrt{\sum_{j=1}^{j=d} p_j^2} \sqrt{\sum_{j=1}^{j=d} M_{ij}^2}} \qquad [1]$$

where s is the cosine similarity, p_j is the jth feature of the probe, m_{ij} is the jth feature of trace i in memory, and d is the dimensionality of the probe and memory vectors.

The activation produced by a probe is found by cubing the similarity between the probe and a trace. The exponent acts as a weighting function that exaggerates the difference between the most similar and least similar traces in memory. The formula for activation is $a_i = s_i^3$, where a_i is the activation of trace i and s_i is the similarity between the probe and trace i.

The degree of activation for each trace determines the echo output by the model, e. This echo is the activation-weighted sum of memory traces, so that the information returned following a probe reflects each of the memory traces in proportion to their similarity to the probe. The complete formula for computing the echo is:

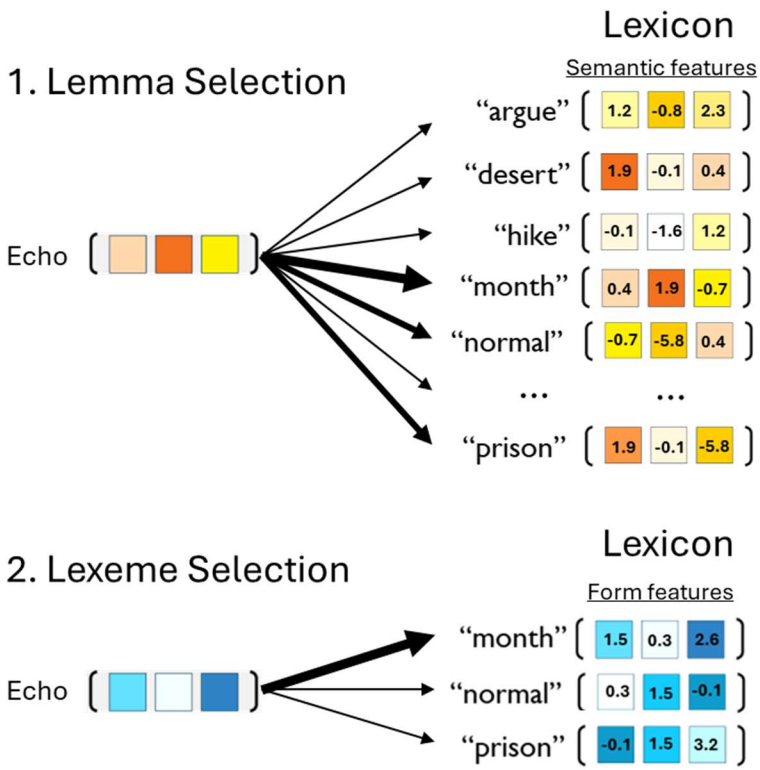
$$e = \sum_{i=1}^{i=m}(M_i \times a_i) = \sum_{i=1}^{i=m} M_i \left(\frac{\sum_{j=1}^{j=d} p_j \times M_{ij}}{\sqrt{\sum_{j=1}^{j=d} p_j^2} \sqrt{\sum_{j=1}^{j=d} M_{ij}^2}} \right)^3 \quad [2]$$

where e is the echo content, m is the number of traces in memory, a_i is the activation of trace M_i in memory, p_j is the jth feature of the probe, M_{ij} is the jth feature of trace M_i, and d is the dimensionality of the vectors.

In simulations, the echo is used to search the lexicon for a target to produce. Figure 29 contains a visualization of the model's two-stage process for reporting a target using information retrieved in the echo.

Figure 29

Lexical retrieval.



Note. Recall requires first identifying a candidate to retrieve through lemma selection, then retrieving the features necessary for production through lexeme selection. Each step involves comparing the elicited information in the echo to items in the lexicon and selecting any items that surpass a threshold.

Lexical retrieval is conditional on the information retrieved in the echo and the word representations. The first stage of lexical retrieval is lemma selection, which requires identifying which item(s) in the lexicon match the semantic information in memory. The echo shares the same structure as the memory traces, so lemma selection occurs by comparing the 300 dimensions of semantic echo content to the semantic vectors in the lexicon. The similarity between an echo, e, and a possible response, r, from the lexicon, s_{er} is compared to a report threshold T. If $s_{er} > T$, then r is a candidate for retrieval. If $s_{er} < T$, that word is not activated sufficiently to be included in the following step. If no candidates are identified in the lexicon, recall fails and the trial is recorded as an omitted response. If one or more candidates are

identified for retrieval, recall proceeds to the second stage of lexeme selection. Lexeme selection requires retrieving the word form information for the candidates identified in lemma selection. To do so, the word form dimensions of the echo are compared to the word form dimensions of candidates. If $s_{er} > T$ during this step, then that candidate is reported. If there are multiple candidates that surpass the threshold, the most activated item is reported. These reports are analyzed to determine if the response was a correct recall, intralist intrusion, or extralist intrusion. Finally, 4AFC responses are simulated by comparing all 600 dimensions in the echo to all 600 dimensions of the representations for each choice on that trial and selecting the choice with the greatest activation. When comparing model simulations to behavioural outcomes, the model's TOTs consisted of omit trials in the top 10% of activation at lexeme selection (i.e., the 10% of omits that came closest the report threshold) and FORGETS consisted of omit trials in the bottom 10% of activation at lexeme selection (i.e., the 10% of omits that were farthest from the report threshold).

Simulations

For each experiment, the model was fit to the empirical data using a grid search method, testing parameterizations of L and T between 0.1 and 1.0. Fits were assessed using mean absolute deviation (MAD) values assessing error for recall performance (proportions of correct recalls, omits, and intrusions) and 4AFC performance (proportions of associate choices during FORGET and TOT trials).

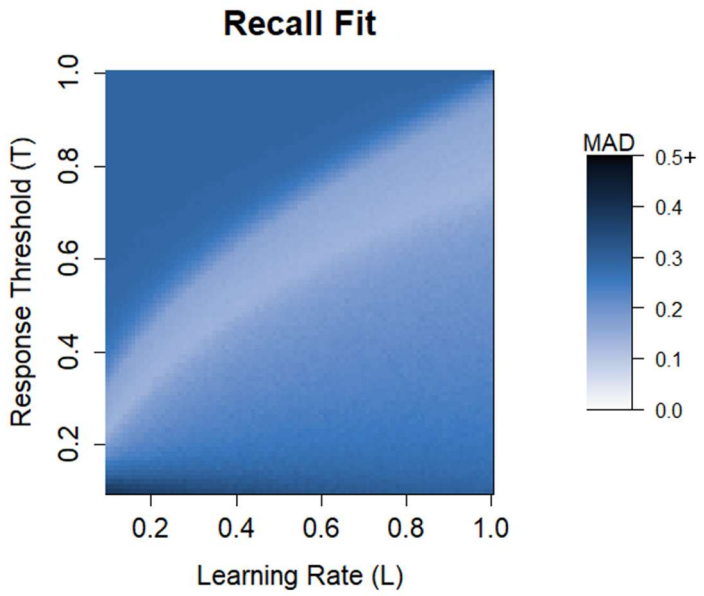
Experiment 1a

The model fits for the recall task showed a linear relationship between L and T such that there was a narrow range of better fits that occurred as L and T increased in proportion with one

another. The fits for the 4AFC task were more homogenous but tended to be worse when parameters (particularly *T*) were more extreme. See Figures 30 and 31 for heatmaps of the recall fits and 4AFC fits, respectively.

Figure 30

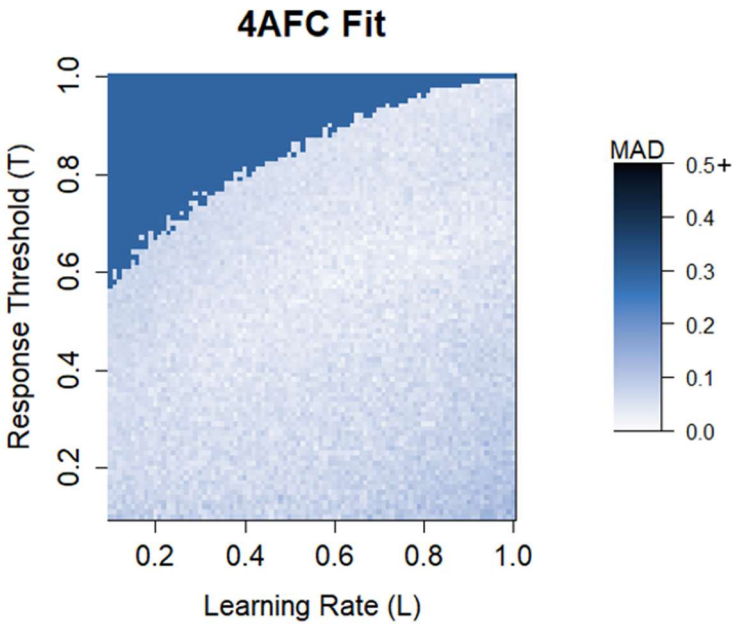
Model fits for recall data.



Note. The model was fit for *L* and *T* parameters from 0.1 to 1.0. MADs were computed for each parameterization based on the proportions of recalls, omits, and intrusions. Lighter cells indicate better fits.

Figure 31

Model fits for 4AFC data.

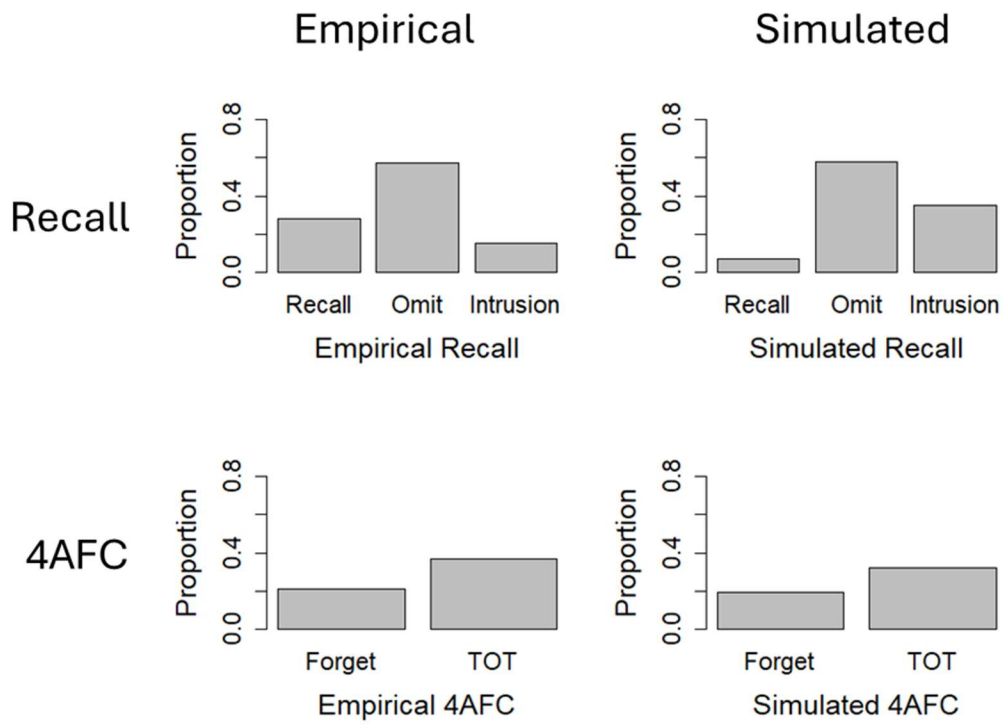


Note. The model was fit for *L* and *T* parameters from 0.1 to 1.0. MADs were computed for each parameterization based on the proportion of associate choices during FORGETs and TOTs. Lighter cells indicate better fits.

The model performed 10000 simulations at the parameters $L = 0.65$, $T = 0.63$ because these parameters had the lowest MAD when summing the recall and 4AFC fits. In these simulations, the model captured 4AFC performance ($MAD = 0.030$) better than recall performance ($MAD = 0.170$). The error in recall performance was caused by underestimating the proportion of correct recalls and overestimating intrusions. The simulated 4AFC results captured the observed improvement in 4AFC performance during TOT trials compared to FORGET trials. Bar graphs of the empirical and simulated results are found in Figure 32.

Figure 32

Simulated and empirical results for Experiment 1a.



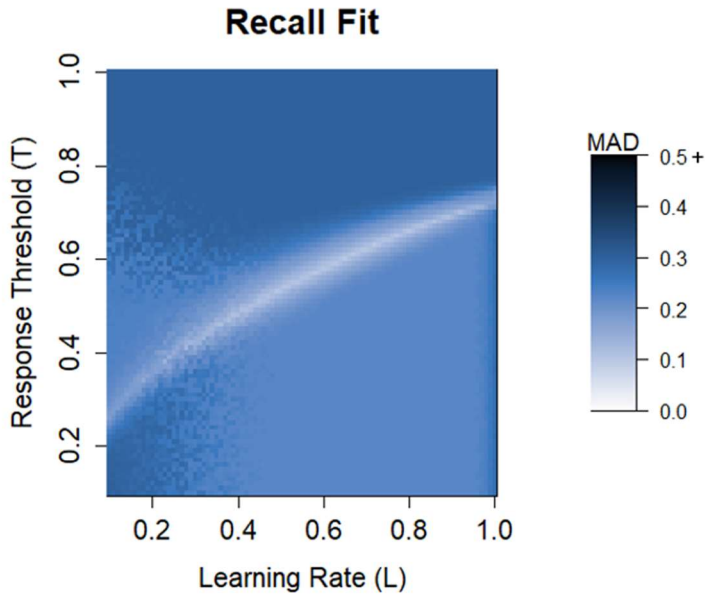
Note. A summary of performance on the recall and 4AFC tasks, with empirical results on the left and simulations on the right. The model predicted fewer recalls and more intrusions than observed in the recall task, but correctly predicted the increase in associate choices during TOTs compared to FORGETs in the 4AFC task.

Experiment 1b

The model fits for the recall task showed a linear relationship between *L* and *T*, resulting in a linear region of closer fits as both parameters increased. The fits for the 4AFC task rapidly dropped off at certain thresholds, as higher *L* values and lower *T* values each led to increased recall responses and thus fewer 4AFC responses. This led to more simulations in which there were either no trials to measure or highly skewed 4AFC responses that massively inflated the error. Within the range of functional parameters, however, the 4AFC fits were consistent. See Figures 33 and 34 for heatmaps of the recall fits and 4AFC fits, respectively.

Figure 33

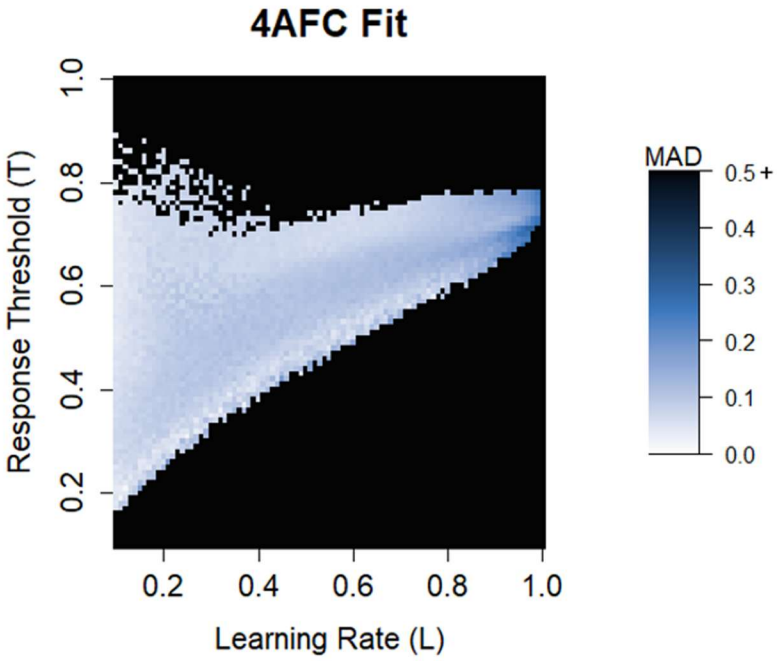
Model fits for recall data.



Note. The model was fit for *L* and *T* parameters from 0.1 to 1.0. MADs were computed for each parameterization based on the proportions of recalls, omits, and intrusions. Lighter cells indicate better fits.

Figure 34

Model fits for 4AFC data.

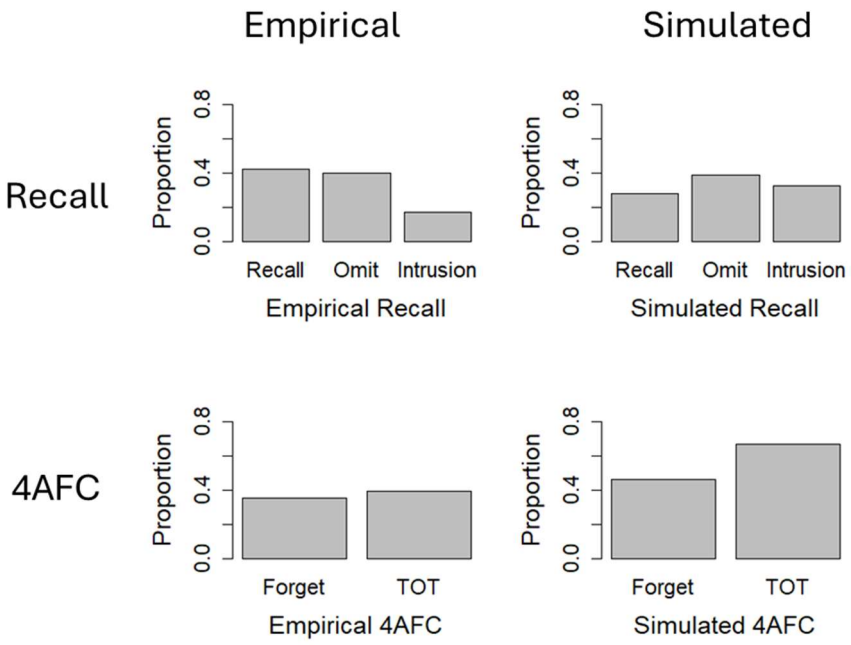


Note. The model was fit for *L* and *T* parameters from 0.1 to 1.0. MADs were computed for each parameterization based on the proportion of associate choices during FORGETs and TOTs. Lighter cells indicate better fits.

The model performed 10000 simulations at the parameters $L = 0.65$, $T = 0.60$ because this parameterization was among the best fits for both recall and 4AFC performance. The model captured recall performance ($MAD = 0.127$) better than 4AFC performance ($MAD = 0.190$). Examining the recall results showed that the model matched the proportion of omits closely but produced fewer correct recalls and more intrusions than participants did. Turning to the 4AFC results, the model predicted a large difference in associate choice proportions between FORGET and TOT trials as opposed to the insignificant difference observed empirically. The predicted associate choice proportion was slightly higher than observed for the FORGET trials, and much higher than observed for the TOT trials. Bar graphs of the empirical and simulated results can be found in Figure 35.

Figure 35

Simulated and empirical results for Experiment 1b.



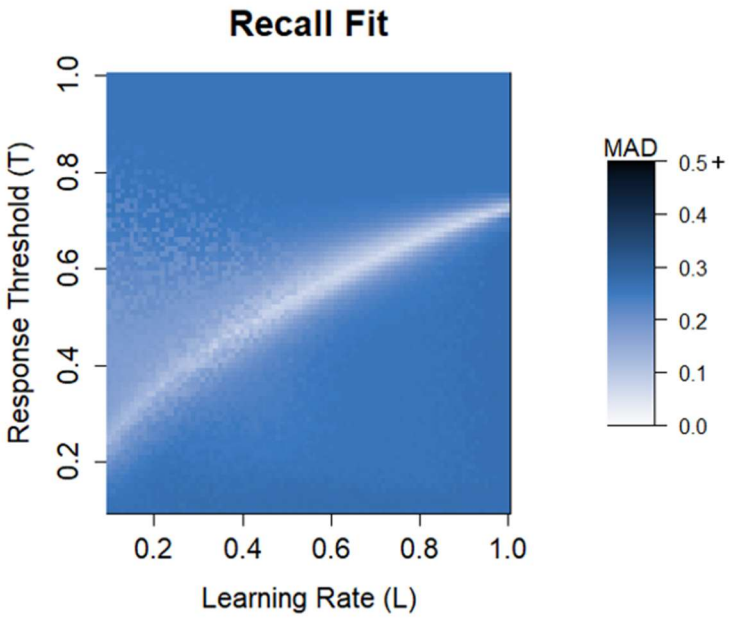
Note. A summary of performance on the recall and 4AFC tasks, with empirical results on the left and simulations on the right. The model predicted fewer recalls and more intrusions than observed in the recall task, as well as a strong increase in associate choices during TOTs that was not observed in the 4AFC task.

Experiment 2

The fits for the recall task showed a linear relationship between *L* and *T*, resulting in a linear region of improved fits. The fits for the 4AFC task rapidly dropped off at certain thresholds, as higher *L* values and lower *T* values each led to increased recall responses and thus fewer 4AFC responses. This led to more simulations in which there were either no trials to measure or highly skewed 4AFC responses that inflated the error. Within the range of functional parameters, however, the 4AFC fits were relatively consistent. The best fits fell within a linear region similar to the best fits for recall responses, but with slightly higher *T* values. See Figures 36 and 37 for heatmaps of the recall fits and 4AFC fits, respectively.

Figure 36

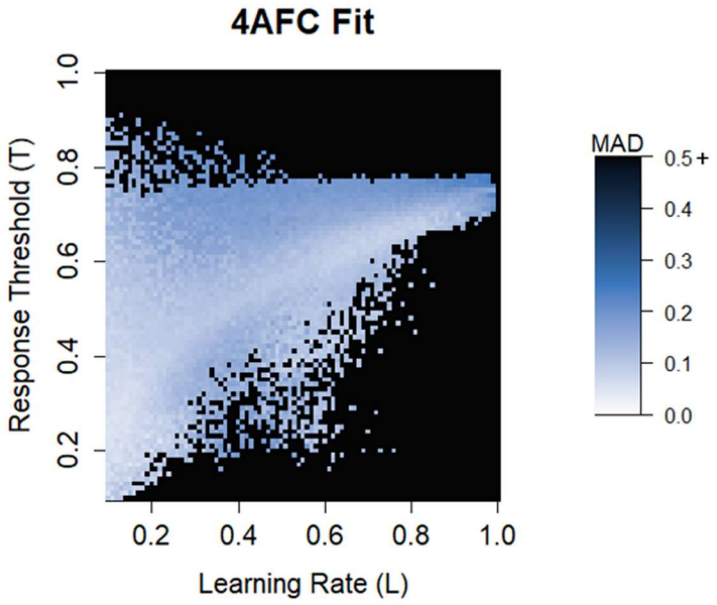
Model fits for recall data.



Note. The model was fit for *L* and *T* parameters from 0.1 to 1.0. MADs were computed for each parameterization based on the proportions of recalls, omits, and intrusions. Lighter cells indicate better fits.

Figure 37

Model fits for 4AFC data.

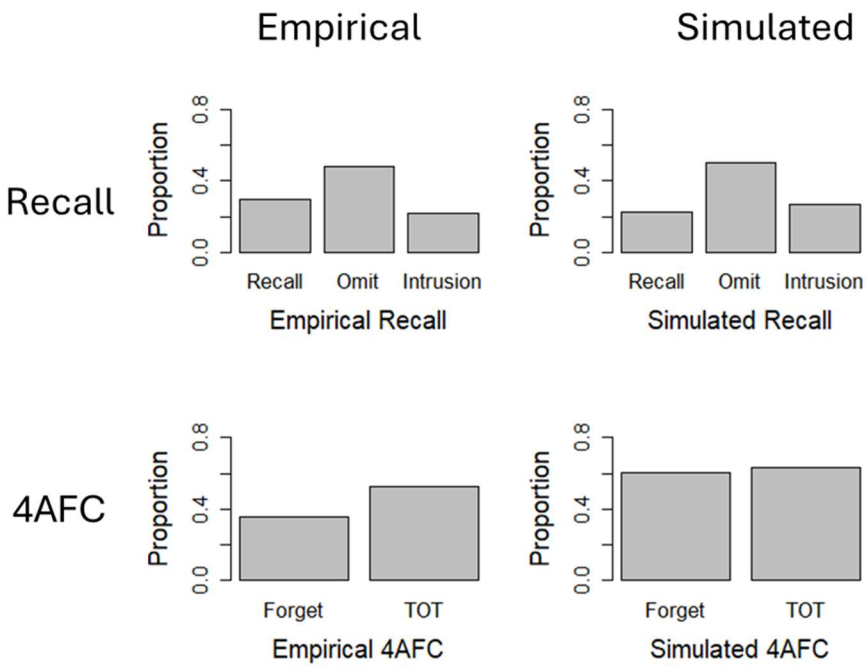


Note. The model was fit for *L* and *T* parameters from 0.1 to 1.0. MADs were computed for each parameterization based on the proportion of associate choices during FORGETs and TOTs. Lighter cells indicate better fits.

The model performed 10000 simulations at the parameters $L = 0.47$, $T = 0.53$ because this parameterization balanced the error for both recall and 4AFC performance. These simulations captured recall performance ($MAD = 0.110$) better than 4AFC performance ($MAD = 0.177$). Examining the recall results showed that the model slightly underestimated the proportion of recalls and overestimated intrusions. The model fared worse with the 4AFC task, failing to predict the observed difference in associate choice proportions between TOT and FORGET trials and overestimating the proportion of associate choices for both trial types. Bar graphs of the empirical and simulated results are found in Figure 38.

Figure 38

Simulated and empirical results for Experiment 2.



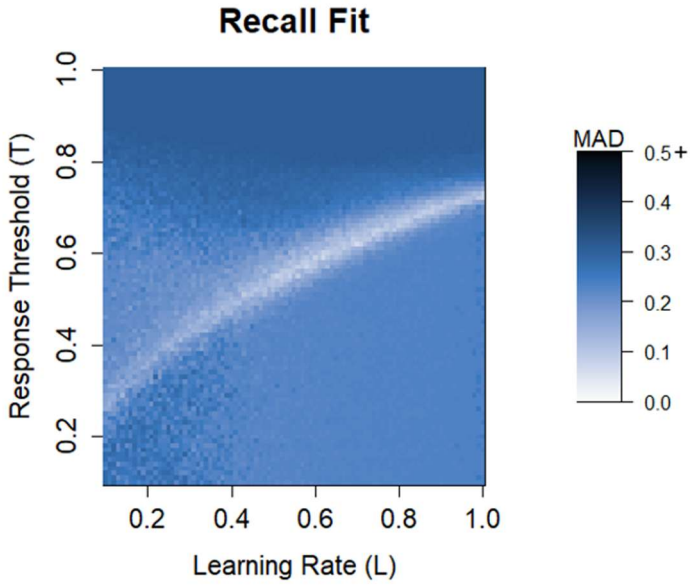
Note. A summary of performance on the recall and 4AFC tasks, with empirical results on the left and simulations on the right. The model predicted a slightly lower proportion of recalls and higher proportion of intrusions than observed, and failed to predict worse 4AFC performance during FORGET trials compared to TOT trials.

Experiment 3

The target "glove" was assigned "grove" as an alternative orthographic associate because "clove" did not appear in the model's lexicon. Fits for the recall task show a linear relationship between *L* and *T*, producing a linear region of better fits. The 4AFC fits continued to show a triangular region of functional parameterizations, outside of which the error rate suddenly increased, but the fits within the range of functional parameters were less uniform than previous fitting routines. The fits for the 4AFC task also demonstrated a linear region of best fits, which was similar to the region for recall but with slightly lower *T* values. Figures 39 and 40 contain heatmaps of the recall and 4AFC fits, respectively.

Figure 39

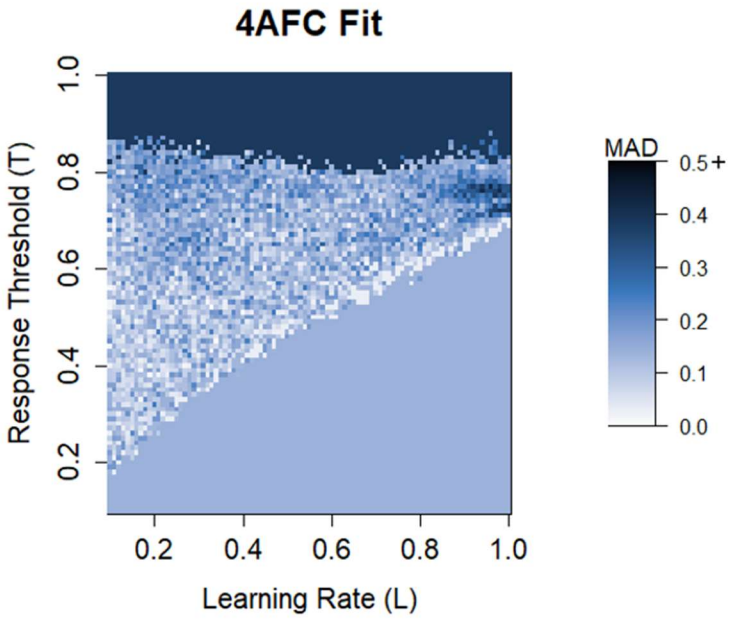
Model fits for recall data.



Note. The model was fit for *L* and *T* parameters from 0.1 to 1.0. MADs were computed for each parameterization based on the proportions of recalls, omits, and intrusions. Lighter cells indicate better fits.

Figure 40

Model fits for 4AFC data.

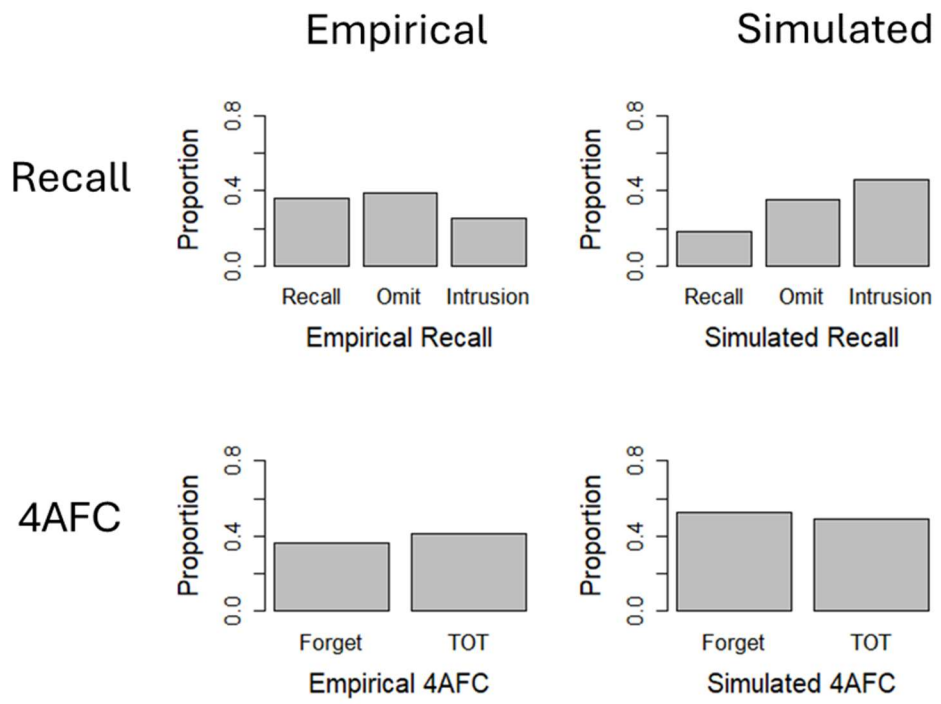


Note. The model was fit for *L* and *T* parameters from 0.1 to 1.0. MADs were computed for each parameterization based on the proportion of associate choices during FORGETs and TOTs. Lighter cells indicate better fits.

The model performed 10000 repeated simulations at the parameters $L = 0.34$, $T = 0.45$ because this parameterization balanced the error for both recall and 4AFC performance. The model captured 4AFC results ($MAD = 0.116$) better than recall results ($MAD = 0.161$). The simulations continued to overestimate intrusions and underestimate recalls. The model was able to predict no difference between FORGET and TOT trials but overestimated the proportions of associate choice in both cases. Bar graphs of the empirical and simulated results are found in Figure 41.

Figure 41

Simulated and empirical results for Experiment 3.



Note. A summary of performance on the recall and 4AFC tasks, with empirical results on the left and simulations on the right. The model predicted fewer recalls and more intrusions than

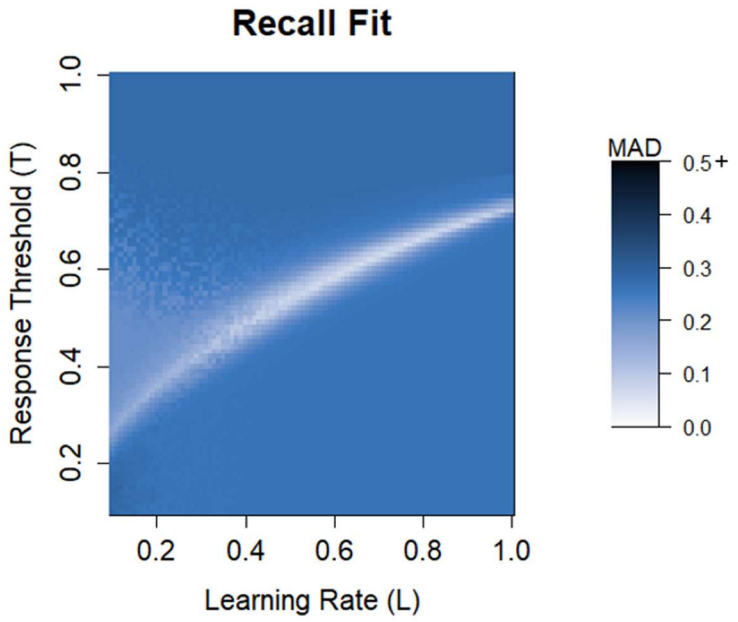
observed in the recall task. It did not predict a difference between FORGET and TOT trials but did overestimate the proportion of associate choice overall.

Experiment 4

The target "ache" was assigned "sachet" as an alternative associate because "achoo" did not appear in the vocabulary. The model fits for the recall task produced a linear strip of best fits that occurred as *L* and *T* increased proportionately with one another. The fits for the 4AFC task continued to show a sharp drop off around $T = 0.8$, and the fits were best with lower values of *L* and *T.* Figures 42 and 43 contain heatmaps of the recall and 4AFC fits, respectively.

Figure 42

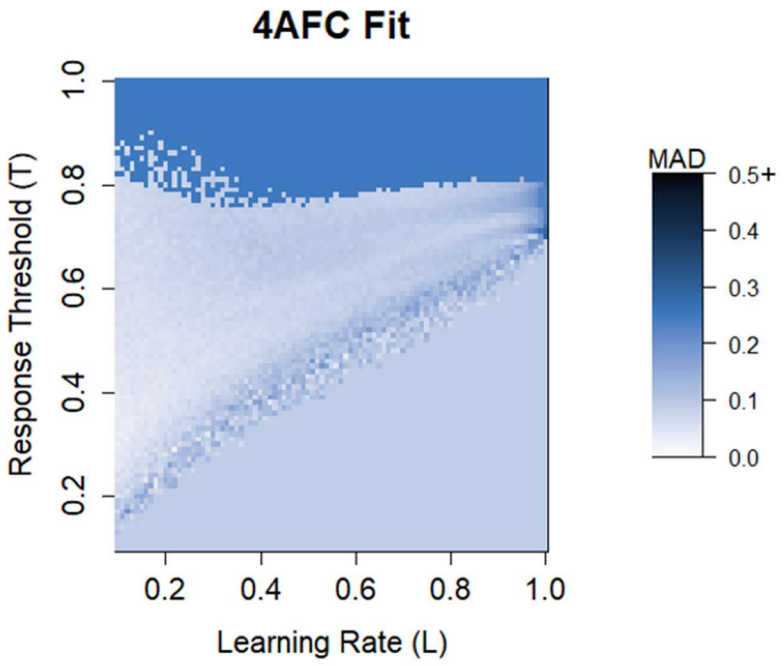
Model fits for recall data.



Note. The model was fit for *L* and *T* parameters from 0.1 to 1.0. MADs were computed for each parameterization based on the proportions of recalls, omits, and intrusions. Lighter cells indicate better fits.

Figure 43

Model fits for 4AFC data.

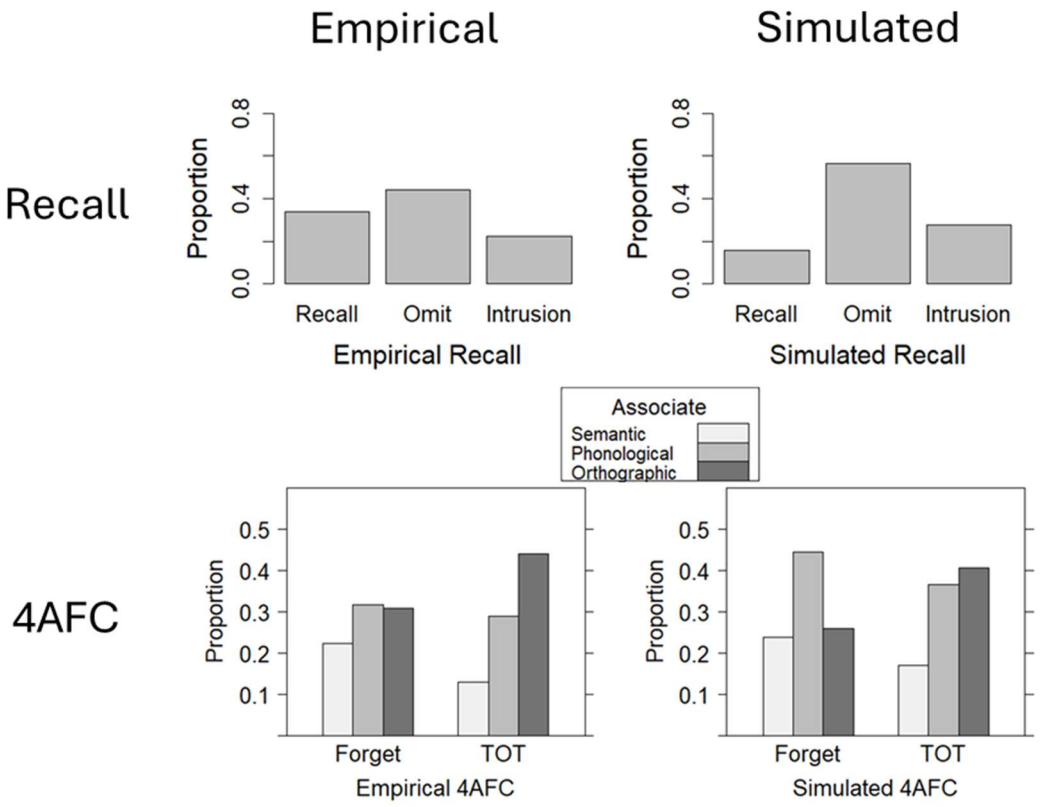


Note. The model was fit for *L* and *T* parameters from 0.1 to 1.0. MADs were computed for each parameterization based on the proportions of all 4AFC responses during FORGETs and TOTs. Lighter cells indicate better fits.

The model performed 10000 simulations at the parameters $L = 0.40$, $T = 0.5$ because these parameters balanced the error for both recall and 4AFC performance. These simulations captured 4AFC performance ($MAD = 0.064$) better than recall performance ($MAD = 0.119$). For the recall fits, the model underestimated the proportion of recalls and overestimated the proportions of omits and intrusions. The 4AFC simulations captured the trends in associate choices, including the relationships between each type of associate for FORGET and TOT trials and how those proportions changed between FORGET and TOT trials. The proportions for phonological associate choices were overestimated and the proportions of orthographic were slightly underestimated, but not sufficiently to change their positions relative to one another. Bar graphs of the empirical and simulated results are found in Figure 44.

Figure 44

Simulated and empirical results for Experiment 4.



Note. A summary of performance on the recall and 4AFC tasks, with empirical results on the left and simulations on the right. The model predicted fewer recalls, more omits, and more intrusions than observed in the recall task. It also predicted the decrease in semantic and phonological associate choice, and the increase in orthographic associate choice between FORGET and TOT trials.

Discussion

Across all experiments, the model tended to underperform on the recall task by producing more intrusions and fewer correct recalls. There were mixed results when predicting 4AFC choices. The model fit Experiment 1a relatively closely, approximately captured the relationships in associate choice data (but not magnitudes of proportions) for Experiments 1b, 3, and 4, and failed to predict the difference in associate choice rates in Experiment 2 entirely by

overestimating the associate choice proportion during FORGET trials. The fitting process showed that the model was capable of better fits for recall responses or 4AFC choices individually, but in most cases no individual set of parameters were optimal for both tasks.

Understanding the successes and failures of the model can provide insight to further refine the theory of lexical retrieval posed by the model. The consistent shortcoming across all simulations was the prevalence of intrusions. This likely arose due to the structure of the representations and how the two-stage retrieval process was implemented. To allow for an independent semantic stage and a combined phonological and orthographic stage during retrieval, the vectors for each item in the lexicon included 300 dimensions for semantics and 300 dimensions containing the sum of the phonological and orthographic vectors. The result is that semantic information is clearer than phonological or orthographic information. This made it more likely that the correct item would be included in the candidate set during lemma selection, but the noisier representations for phonology and orthography increased the likelihood of extralist intrusions.

Future versions of the model could improve upon this in a few ways. The semantic, phonological, and orthographic representations could all be summed when items are encoded to memory. This would result in noisier memory traces but would also impact all three modalities more equally. Memory probes would also be composed of representations for all three modalities. The two-stage retrieval process would require weighing the three modalities when creating probes, with an initial probe weighted more heavily for semantic information during lemma selection, and another weighted for phonological and orthographic information during lexeme selection. An additional benefit to using composite representations for all three modalities would be the use of a single threshold to allow retrieval failure at either stage of

retrieval. Because the semantic representations are less noisy in the current model, fitting the parameters for recall responses seemed to result in semantic candidates always being identified, and omits only occurring at the second stage of retrieval. Having equally fuzzy encoding of all three modalities should mean that one set of parameters would be more likely to produce recall failures during both lemma selection and lexeme selection. This should enable the model to function better with higher values of *T*, thus filtering out more intrusions during lemma selection and not forcing the model to try and fit proportions of recalls, omits, and intrusions during lexeme selection alone. With failure occurring at both stages, the model will do a better job of differentiating retrieval failure (when failure occurs at lemma selection) from TOTs arising from transmission deficit or step two retrieval failures (Burke et al., 1991; Gollan & Brown, 2006; L. E. James & Burke, 2000).

Finally, an ongoing retrieval process would allow the model to account for retrieval time and could enable better performance. This would involve taking the echo content retrieved from a probe and creating a loop that iteratively searches memory using the elicited information, which is a solution for partial recall proposed by Hintzman (1986). This provides value to the theory because it captures two key features of the TOT experience – their ongoing nature and the tendency to identify a detail that feels right and latch onto it in the hopes it will help identify the target. This process may also be necessary to work with noisier representations using the three combined modalities. By introducing an evolving probe that gets shaped by the contents of memory, the model can gradually narrow in on a target or become stuck in a dead end of conflicting information. The dead end outcome is also an important element of the TOT phenomenon, which commonly reoccurs for the same items when repeatedly tested (Oliver & Humphreys, 2019; Warriner & Humphreys, 2008). The process would still be a two-stage

retrieval, but each stage could occur through this looping search instead of a single instantaneous decision. This is in keeping with the subjective experience of TOTs as well findings that TOTs require ongoing attention and working memory (Ryan et al., 1982; Schwartz, 2008).

General Discussion

In this thesis, my aim was to investigate how TOTs arise during the process of memory retrieval. I first conducted a series of experiments to obtain empirical data to evaluate established findings in the literature. Two-stage lexical access theories and direct-access TOT theories generally agree that semantic information must be processed as a first step which allows the ensuing retrieval of phonological information. The availability of these and other modalities of lexical information during TOTs were explored independently and in relation to one another. Experiments 1a adds to the body of support for early semantic processing by allowing participants to demonstrate semantic knowledge for unrecalled words. Whereas past studies have prospected for TOTs using definitions and picture-naming tasks, in Experiment 1b I found that semantic information was not accessible during TOTs when retrieval is prompted by a paired cue word. Phonological interlopers are the most salient and well-documented component of TOTs but, aside from guessing the first letter, freely reported features are frequently inaccurate (Huebert et al., 2023). Experiment 2 invited participants to identify homophones of unrecalled targets as well as rhyming near-homophones with different first letters to provide insight into the retrieval of phonological information including, but not limited to, first letters. Unlike the semantic case, participants could identify phonological associates more often during TOTs than FORGETs. It was also found that participants performed better when identifying homophone associates than near-homophones that did not share a first letter with the target, and that TOTs were associated with a significant increase in choosing homophones but not near-homophones.

This shows that first letter interlopers during TOTs occurred with sufficient frequency and accuracy to impact 4AFC performance. Experiment 3 tested for orthographic retrieval during TOTs using words where orthography and phonology diverge. While participants did not choose the associate significantly more often during TOTs, the analysis of FOKs did show a significant effect. This could have occurred due to the lower rate of TOTs in the experiment, or participants may have had access to orthographic information but shifted their attention towards phonology during TOTs, thus ignoring the orthographically similar but phonologically distinct associates. In Experiment 4, I placed all three associate types in competition and found that only orthographic associate choices significantly increased between FORGET and TOT trials. The TOT and FOK analyses both indicate that as participants became increasingly confident in being able to access a target, they increasingly identified the orthographic associate even though the phonological associate was the only option selected significantly above chance levels during FORGET and DK trials.

Taking a step back to consider all of the results together allows for more insights. Across all experiments, participants showed a tendency to recall extralist intrusions that were more orthographically than phonologically similar to the target. This trend occurred during both initial and final recall tasks, and during FORGETs and TOTs. Behaviour was strongly shaped by the task at hand and the modality highlighted in the materials. When attempting to report a studied word by typing, participants were biased towards orthography more than when choosing associates. Furthermore, the final recall task was impacted by the types of associates they had just viewed on the preceding 4AFC trial. Associate choices were more often associated with recalling the correct target than TOTs were, and extralist intrusions in final tended to be more similar in the modality of the associate than extralist intrusions during initial recall. This

opportunistic use of information from the associates cumulated in Experiment 4, where their recognition of target information in the orthographic associate followed the same direction as their bias when recalling target information. Why then, would additional phonological and semantic cueing lead to a preference for orthography? Firstly, the orthographic associate was closest in nature to the target stimuli exactly as it was presented in a typed form. Secondly, the insignificant increase in orthographic associate choices during TOTs in Experiment 3 and the insignificant increase in near-homophone associate choices during TOTs in Experiment 2 share a similar level of orthographic/phonological conflict. This type of conflict has been shown to impair rhyme judgements in the case of an orthographic mismatch, and spelling similarity judgements in the case of phonological mismatch (Polich et al., 1983; Welcome & Alton, 2015). Viewing the phonological associate during Experiment 4 may have offered the additional push participants needed to overcome this conflict during a TOT and recognize that the orthographic associate was most like the item they studied. Differences in orthographic/phonological conflict can explain how near-homophone associates benefitted less than homophones during TOTs, and why orthographic associates benefitted less from TOTs during Experiment 3 than Experiment 4.

The modeling work showed promise; however some additional work will be needed to enable this implementation of two-stage lexical retrieval to capture recall and TOT. The shortcoming appeared to arise from how the lexicon of word representations was constructed, with one semantic vector and a summed phonological/orthographic vector. The model tended to retrieve clear semantics and foggy phonology and orthography, resulting in either too many intrusions (with a lower response threshold) or too many omits and highly biased 4AFC decisions (with a higher response threshold). This is also likely why the model predicted a strong effect of TOTs on semantic associate choice, but not phonological or orthographic associate

choice. The separation of semantic and phonological/orthographic information need not be abandoned to overcome this problem, but a more nuanced approach is needed. The system could use representations consisting of weighted sums of the vectors for each modality. Each modality would be stored with equal weights, and therefore similar amounts of noise from the other modalities. Probes would be constructed with biased weights to force the model to attend to semantics during lemma selection, then phonology and orthography during lexeme selection. This should lead to improved model performance, simplify the assumptions around semantic memory using a lexicon of combined representations, and retain the two-stage retrieval theory at the core of the model. Although flawed, the model captured the relationships (if not the magnitudes) of many of the empirical outcomes, including the 4AFC results of Experiments 1a and 4 which had more complex relationships between the stimuli. Furthermore, the simulations presented were in most cases a compromise between fitting the recall and the 4AFC data. For these reasons, I am optimistic that with some modifications the model will be flexible enough to accommodate both tasks at once and to generate new testable hypotheses. Finally, future modelling could further explore the impacts of orthographic/phonological conflict using representations for orthography and phonology that facilitate within-item comparisons.

The experimental work has contributed to the literature on TOTs by supporting the use of paired associate recall tasks to elicit TOTs, with experiments designed to address criticisms of similar work in the past (A. S. Brown, 2011; Ryan et al., 1982; Valentine et al., 1996). Through this task, the experiments targeted established TOT phenomena from a different angle. The validity of the experimental design was affirmed by replication of familiar results including the strong focus on phonology during TOTs. Contrasting outcomes between paired associate recall and a cued recall task (resembling the more traditional prospecting task) showed that participants

may require more semantic input to shift their focus to semantic information during TOTs. This work highlights the impact of task demands in lexical retrieval as participants displayed a strong orthographic bias when attempting to type targets, and this bias extended into 4AFC decisions during TOTs only when more lexical information was available. Orthographic/phonological conflict has been shown to impact judgements of rhyme, visual similarity, and lexical decision (Lim et al., 2023; Polich et al., 1983; Welcome & Alton, 2015), and the results of Experiments 2 – 4 showed an impact on the perceived similarity of targets and associates during TOTs. This occurred across different stimuli including weak and strong phonology-to-orthography mismatches (i.e., homophones and near-homophones) and strong orthography-to-phonology mismatches (near-homographs). Future research could employ auditory stimuli and even spoken responses to determine whether the orthographic response bias gives way to a phonological bias, and whether the impact of orthographic/phonological conflict is strengthened in the auditory modality (see Cortese, 1998; Cortese et al., 2004). These findings made use of the TOT phenomenon to explore the inner workings of lexical retrieval, showing that retrieving a word requires coordinating multiple modalities of interrelated lexical information in a task-oriented fashion.

References

Baddeley, A., & Lewis, V. (2017). Inner Active Processes in Reading: The Inner Voice, the Inner Ear, and. In C. A. Perfetti & A. M. Lesgold (Eds.), *Interactive processes in reading* (Vol. 6, pp. 107–129). Taylor & Francis Group. https://doi.org/10.4324/9781315108506-5

Bahrick, H. P., Baker, M. K., Hall, L. K., & Abrams, L. (2011). How should we define and differentiate metacognitions? In *Successful Remembering and Successful Forgetting: A Festschrift in Honor of Robert A. Bjork* (1st Edition, pp. 329–346). Psychology Press.

Brown, A. S. (2011). *The Tip of the Tongue State*. Psychology Press. https://doi.org/10.4324/9780203582961

Brown, R., & McNeill, D. (1966). The "Tip of the Tongue" Phenomenon. *Journal of Verbal Learning and Verbal Behavior*, *5*, 325–337.

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

Burke, D. M., MacKay, D. G., Worthley, J. S., & Wade, E. (1991). On the tip of the tongue: What causes word finding failures in young and older adults? *Journal of Memory and Language*, *30*, 542–579.

Caramazza, A., & Miozzo, M. (1997). The relation between syntactic and phonological knowledge in lexical access: Evidence from the `tip-of-the-tongue' phenomenon. *Cognition*, *64*(3), 309–343. https://doi.org/10.1016/S0010-0277(97)00031-0

Caramazza, A., & Miozzo, M. (1998). More is not always better: A response to Roelofs, Meyer, and Levelt. *Cognition*, *69*, 231–241.

Cortese, M. J. (1998). Revisiting Serial Position Effects in Reading. *Journal of Memory and Language*, *39*(4), 652–665. https://doi.org/10.1006/jmla.1998.2603

Cortese, M. J., Watson, J. M., Wang, J., & Fugett, A. (2004). Relating distinctive orthographic and phonological processes to episodic memory performance. *Memory & Cognition*, *32*(4), 632–639. https://doi.org/10.3758/BF03195854

D'Angelo, M. C., & Humphreys, K. R. (2015). Tip-of-the-tongue states reoccur because of implicit learning, but resolving them helps. *Cognition*, *142*, 166–190. https://doi.org/10.1016/j.cognition.2015.05.019

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Gollan, T. H., & Brown, A. S. (2006). From tip-of-the-tongue (TOT) data to theoretical implications in two steps: When more TOTs means better retrieval. *Journal of Experimental Psychology: General*, *135*(3), 462–483. https://doi.org/10.1037/0096-3445.135.3.462

Günther, F., Dudschig, C., & Kaup, B. (2015). LSAfun—An R package for computations based on Latent Semantic Analysis. *Behavior Research Methods*, *47*(4), 930–944. https://doi.org/10.3758/s13428-014-0529-0

Hanly, S., & Vandenberg, B. (2010). Tip-of-the-Tongue and Word Retrieval Deficits in Dyslexia. *Journal of Learning Disabilities*, *43*(1), 15–23. https://doi.org/10.1177/0022219409338744

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

Hintzman, D. L. (1986). “Schema Abstraction” in a Multiple-Trace Memory Model. *Psychological Review*, *93*(4), 411–428.

Huebert, A. M., McNeely-White, K. L., & Cleary, A. M. (2023). On the relationship between tip-of-the-tongue states and partial recollective experience: Illusory partial recollective access during tip-of-the-tongue states. *Journal of Experimental Psychology: General*, *152*(2), 542–570. https://doi.org/10.1037/xge0001292

James, L. E., & Burke, D. M. (2000). Phonological priming effects on word retrieval and tip-of-the-tongue experiences in young and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(6), 1378–1391. https://doi.org/10.1037/0278-7393.26.6.1378

James, W. (1890). *The Principles of Psychology: Vol. I*. Henry Holt and Co.

Kahana, M. J. (2002). Associative symmetry and memory theory. *Memory & Cognition*, *30*(6), 823–840. https://doi.org/10.3758/BF03195769

Kumar, A. A., Balota, D. A., Habbert, J., Scaltritti, M., & Maddox, G. B. (2019). Converging semantic and phonological information in lexical retrieval and selection in young and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(12), 2267–2289. https://doi.org/10.1037/xlm0000699

Landauer, T. K., & Dumais, S. T. (1997). A Solution to Plato’s Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, *104*(2), 211–240.

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

Levelt, W. J. M. (1992). Accessing words in speech production: Stages, processes and representations. *Cognition*, *42*(1), 1–22. https://doi.org/10.1016/0010-0277(92)90038-J

Levelt, W. J. M. (1993). *Speaking: From Intention to Articulation*. MIT Press.

Lim, A., O'Brien, B., & Onnis, L. (2023). Orthography-phonology consistency in English: Theory- and data-driven measures and their impact on auditory vs. visual word recognition. *Behavior Research Methods*, *56*(3), 1283–1313. https://doi.org/10.3758/s13428-023-02094-5

Loo, M. P. J. van der. (2014). The stringdist Package for Approximate String Matching. *The R Journal*, *6*(1), 111–122.

MacLeod, C. M., & Kampe, K. E. (1996). *Word Frequency Effects on Recall, Recognition, and Word Fragment Completion Tests*.

Maril, A., Simons, J. S., Weaver, J. J., & Schacter, D. L. (2005). Graded recall success: An event-related fMRI comparison of tip of the tongue and feeling of knowing. *NeuroImage*, *24*(4), 1130–1138. https://doi.org/10.1016/j.neuroimage.2004.10.024

Metcalfe, J., Schwartz, B. L., & Joaquim, S. G. (1993). The cue-familiarity heuristic in metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(4), 851–861. https://doi.org/10.1037/0278-7393.19.4.851

Meyer, A. S., & Bock, K. (1992). The tip-of-the-tongue phenomenon: Blocking or partial activation? *Memory & Cognition*, *20*(6), 715–726. https://doi.org/10.3758/BF03202721

Miralpeix, I. (2019). L1 and L2 Vocabulary Size and Growth. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (1st ed., pp. 189–206). Routledge. https://doi.org/10.4324/9780429291586-13

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407. https://doi.org/10.3758/BF03195588

Nosofsky, R. M. (1986). Attention, Similarity, and the Identification-Categorization Relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Oliver, L. K., & Humphreys, K. R. (2019). Phonological Interlopers Tend to Repeat When Tip-of-the-Tongue States Repeat. *Frontiers in Psychology*, *10*. https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00341

Osth, A. F., & Zhang, L. (2024). Integrating word-form representations with global similarity computation in recognition memory. *Psychonomic Bulletin & Review*, *31*(3), 1000–1031. https://doi.org/10.3758/s13423-023-02402-2

Parrish, A. (2017). Poetic sound similarity vectors using phonetic features. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, *13*(2), 99–106. https://ojs.aaai.org/index.php/AIIDE/article/view/12971

Pexman, P. M., Lupker, S. J., & Jared, D. (2001). Homophone effects in lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(1), 139–156. https://doi.org/10.1037/0278-7393.27.1.139

Polich, J., McCarthy, G., Wang, W. S., & Donchin, E. (1983). When words collide: Orthographic and phonological interference during word processing. *Biological Psychology*, *16*(3–4), 155–180. https://doi.org/10.1016/0301-0511(83)90022-4

Reid, J. N., Yang, H., & Jamieson, R. K. (2023). A computational account of item-based directed forgetting for nonwords: Incorporating orthographic representations in MINERVA 2. *Memory & Cognition*, *51*(8), 1785–1806. https://doi.org/10.3758/s13421-023-01433-3

Roelofs, A., Meyer, A. S., & Levelt, W. J. M. (1998). A case for the lemma/lexeme distinction in models of speaking: Comment on Caramazza and Miozzo (1997). *Cognition*, *69*.

Ryan, M. P., Petty, C. R., & Wenzlaff, R. M. (1982). Motivated remembering efforts during tip-of-the-tongue states. *Acta Psychologica*, *51*(2), 137–147. https://doi.org/10.1016/0001-6918(82)90058-0

Schwartz, B. L. (1998). Illusory Tip-of-the-tongue States. *Memory*, *6*(6), 623–642. https://doi.org/10.1080/741943371

Schwartz, B. L. (2008). Working memory load differentially affects tip-of-the-tongue states and feeling-of-knowing judgments. *Memory & Cognition*, *36*(1), 9–19. https://doi.org/10.3758/MC.36.1.9

Schwartz, B. L., & Metcalfe, J. (2011). Tip-of-the-tongue (TOT) states: Retrieval, behavior, and experience. *Memory & Cognition*, *39*(5), 737–749. https://doi.org/10.3758/s13421-010-0066-8

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

Umanath, S., Coane, J. H., Renaker, J. T., Whitman, K., Lee, A. A., & Kim, S. (2025). Using the phenomenology of knowledge-based retrieval failures in younger and older adults to characterize proximity to retrieval success and identify a Zone of Proximal Retrieval.

Journal of Memory and Language, *140*, 104582. https://doi.org/10.1016/j.jml.2024.104582

Valentine, T., Bredart, S., & Brennen, T. (1996). *The Cognitive Psychology of Proper Names*. Routledge.

Warriner, A. B., & Humphreys, K. R. (2008). Short Article: Learning to Fail: Reoccurring Tip-of-the-Tongue States. *Quarterly Journal of Experimental Psychology*, *61*(4), 535–542. https://doi.org/10.1080/17470210701728867

Welcome, S. E., & Alton, A. C. (2015). Individual Differences in the Effect of Orthographic/Phonological Conflict on Rhyme and Spelling Decisions. *PLOS ONE*, *10*(3), e0119734. https://doi.org/10.1371/journal.pone.0119734

Whitney, C., & Marton, Y. (2013). *The SERIOL2 Model of Orthographic Processing*. Retrieved from https://www.researchgate.net/publication/237065841_The_SERIOL2_Model_of_Orthographic_Processing

Appendix A

Appendix A. Stimuli used in Experiment 1a, taken in part from Kumar et al., 2019.

List A	Target	Clue	Associate	Foil 1	Foil 2	Foil 3
	Allocate	To set aside	Budget	Abandon	Forget	Store
	Decanter	Holds drinks	Bottle	Mug	Cup	Coaster
	Deplete	To exhaust something	Expend	Exercise	Tire	Discharge
	Linen	Plant-based material	Sheets	Wicker	Table	Paper
	Sextant	Related to navigating	Sailing	Landmark	GPS	Road
	Tranquilizer	Causes a sedative effect	Depressant	Fatigue	Tea	Nap

List B	Target	Clue	Associate	Foil 1	Foil 2	Foil 3
	Abdicate	Abandon a position	Renounce	Retreat	Travel	Yield
	Asbestos	Insulate	Cancer	Walls	Isolate	Foam
	Incubate	Hold at a certain temperature	Nest	Fridge	Oven	Cook
	Instigate	To set in motion	Provoke	Roll	Throw	Commence
	Javelin	Something you throw	Spear	Ball	Rock	Punches
	Tariff	A type of fee	Tax	Parking	Reservation	Interest

Appendix B

Appendix B. Stimuli for use in Experiment 1b, taken from Nelson et al., 2004.

Cue	Target	Associate	Foil 1	Foil 2	Foil 3
shun	admire	respect	mass	taste	body
normal	argue	debate	yellow	steel	bike
ready	customer	service	brown	pile	running
cork	drive	car	dancer	squeal	diary
forest	dry	wet	bitter	gull	bureau
pretty	eggs	bacon	can	twist	stairs
chip	federal	government	equation	napkin	method
sorry	fish	water	analyze	trade	noise
jeans	flower	petals	honest	news	speed
channel	grant	money	wood	force	mat
safety	inhale	exhale	toasty	blood	picture
disease	kid	child	casual	extra	dove
corner	law	rule	pottery	nice	top
write	mean	rude	pad	lose	ray
hike	month	year	man	traffic	peace
cold	nervous	tense	walker	shorts	jewelry
desert	prison	criminal	manager	perfume	jump
tumor	real	life	mistreat	compete	freedom
promise	room	house	ant	teeth	fog
margin	sky	cloud	dresser	sin	child
crisis	tan	bronze	flip	leave	business
single	tumble	fall	member	twin	dull
sliver	voice	sing	penguin	husband	mess
food	warn	danger	party	chain	dating

Appendix C

Appendix C. Stimuli for use in Experiment 2, with targets and associates consisting of homophone pairs taken from Pexman et al., 2001 and near-homophone rhyming pairs from Polich et al., 1983.

Homophones	**Cue**	**Target**	**Associate**	**Foil 1**	**Foil 2**	**Foil 3**
	quantity	aloud	allowed	thief	liquor	pearl
	whisper	blue	blew	maid	understand	odd
	medal	site	sight	ask	please	minus
	spoon	flower	flour	people	beam	neck
	woods	higher	hire	parent	covered	plant
	descent	nun	none	car	dream	opinion
	good	pear	pair	term	touch	rooster
	bounty	rows	rose	talent	care	helpful
	diet	sale	sail	team	brass	hedge
	dive	size	sighs	because	pastor	cobra
	stupid	whether	weather	prove	harsh	squeeze
	natural	would	wood	talk	special	lobster
Near-Homophones	**Cue**	**Target**	**Associate**	**Foil 1**	**Foil 2**	**Foil 3**
	date	ache	make	radio	sauce	specific
	grade	chew	shoe	fun	experiment	help
	fake	comb	roam	rest	drunk	blast
	whiskers	dual	fool	rub	mercy	factory
	merit	fate	bait	rubber	replace	choose
	seven	float	quote	captain	end	reason
	schedule	gait	hate	purpose	double	art
	planet	hail	male	machine	poetry	none
	look	jewel	fool	dark	peer	fast
	shake	known	phone	scare	famous	fly
	play	loan	bone	scheme	band	golden
	lost	moose	juice	sometimes	left	tax

Appendix D

Appendix D. Stimuli for use in Experiment 3, with targets and orthographic associate words taken from Polich et al. (1983).

Cue	Target	Associate	Foil 1	Foil 2	Foil 3
kiss	bone	gone	cake	string	two
close	boot	foot	style	tell	hut
wake	comb	tomb	tale	ball	eye
guess	crow	brow	dope	way	germ
cheat	drove	prove	cab	bean	lake
mark	fear	bear	chance	name	toast
stud	freak	break	cold	cane	brain
church	give	hive	fish	sleep	shout
head	glove	clove	crack	noise	tool
sweep	have	gave	tin	ride	dance
cup	hood	food	arts	nice	cure
roam	hush	push	search	cheap	bug
sale	lose	nose	top	bird	crops
horse	moth	both	slope	meat	buy
see	move	love	cooked	drink	yard
muck	pull	dull	crest	drift	boat
choir	rough	dough	wine	seize	cops
calm	sour	four	fair	tart	quick
ace	toll	doll	mild	hard	start
sport	touch	couch	right	haul	wax
book	wand	hand	truck	snap	light
home	wasp	rasp	board	glance	plane
grass	worse	horse	surf	blood	toy
sick	your	hour	wet	death	slow

Appendix E

Appendix E. Stimuli for use in Experiment 4, taken in part from Pexman et al. (2001) and Polich et al. (1983).

Cue	Target	Semantic Associate	Orthographic Associate	Phonological Associate	Foil
desk	ache	pain	achoo	make	handle
limit	beard	mustache	heard	weird	skinny
hush	bone	dog	gone	own	slack
hip	bough	tree	cough	cow	fuse
garbage	comb	brush	tomb	roam	shun
pants	dough	bread	rough	doe	worm
unique	fear	scared	bear	mere	eat
hurry	flower	petal	lower	flour	tooth
award	foul	smell	soul	owl	rival
warm	great	good	treat	freight	line
dirty	here	absent	were	ear	treasure
foam	lose	win	nose	news	sauce
grab	none	zero	cone	nun	seat
strong	pear	fruit	rear	pair	rake
square	pour	spill	tour	oar	girl
tray	real	fake	realm	feel	call
brittle	rows	columns	brows	rose	just
crane	steak	beef	steal	stake	drink
crash	taste	food	caste	laced	coal
motor	tier	layer	tiger	sphere	prank
branch	tough	hard	ought	fluff	plan
green	wear	clothes	hear	fare	slob
jerk	weight	heavy	height	wait	frisk
text	wood	log	mood	would	tie

Appendix F

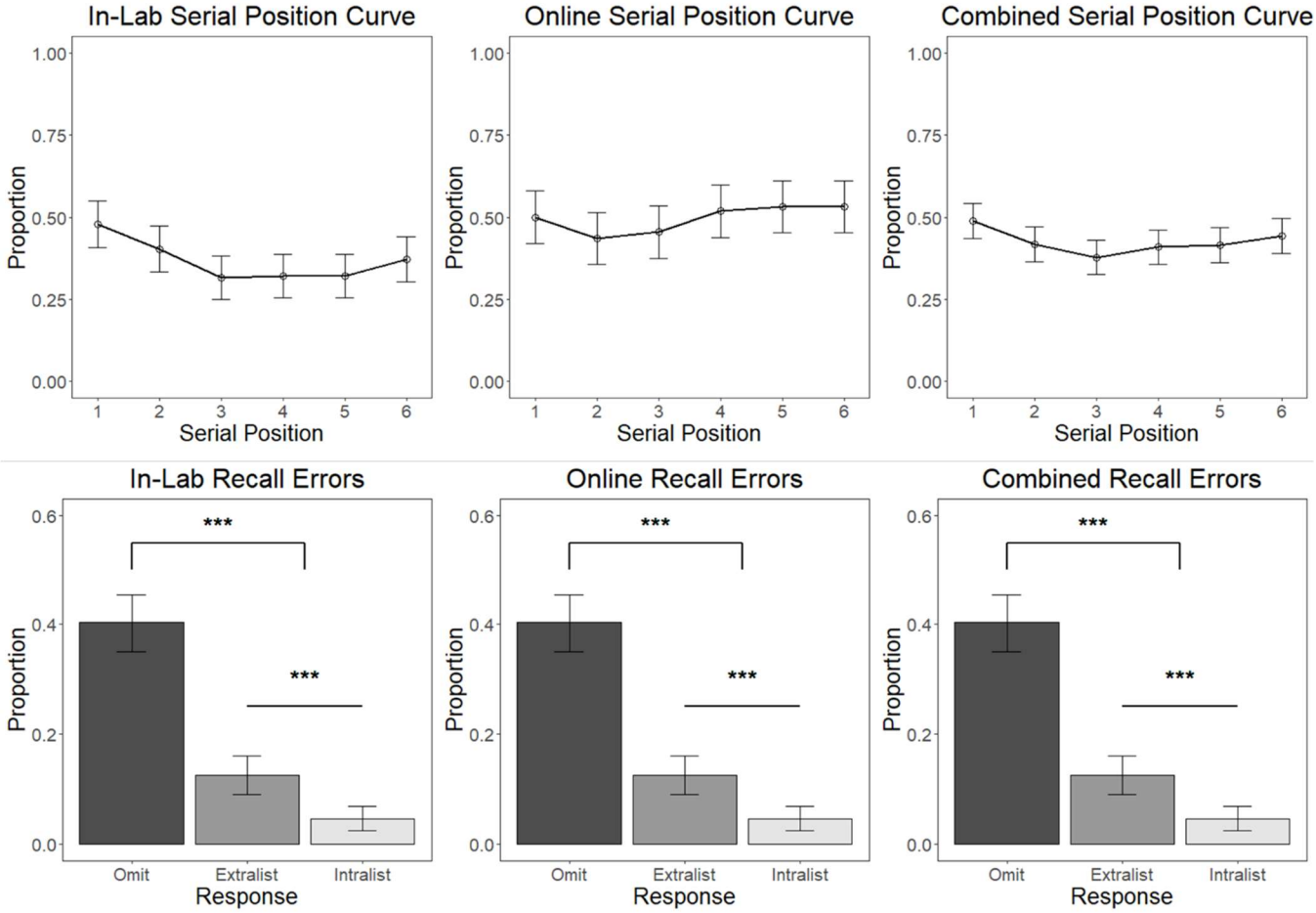
Appendix F. Comparison of data collected online versus in-lab.

Experiment 1b

Figure F1 displays the serial position curves and error proportions for both versions of the experiment independently and combined.

Figure F1

Comparison of recall data collected in-lab and online for Experiment 1b.



Note. Both versions of the experiment showed similar proportions of correct recall, omits, extralist intrusions, and intralist intrusions.

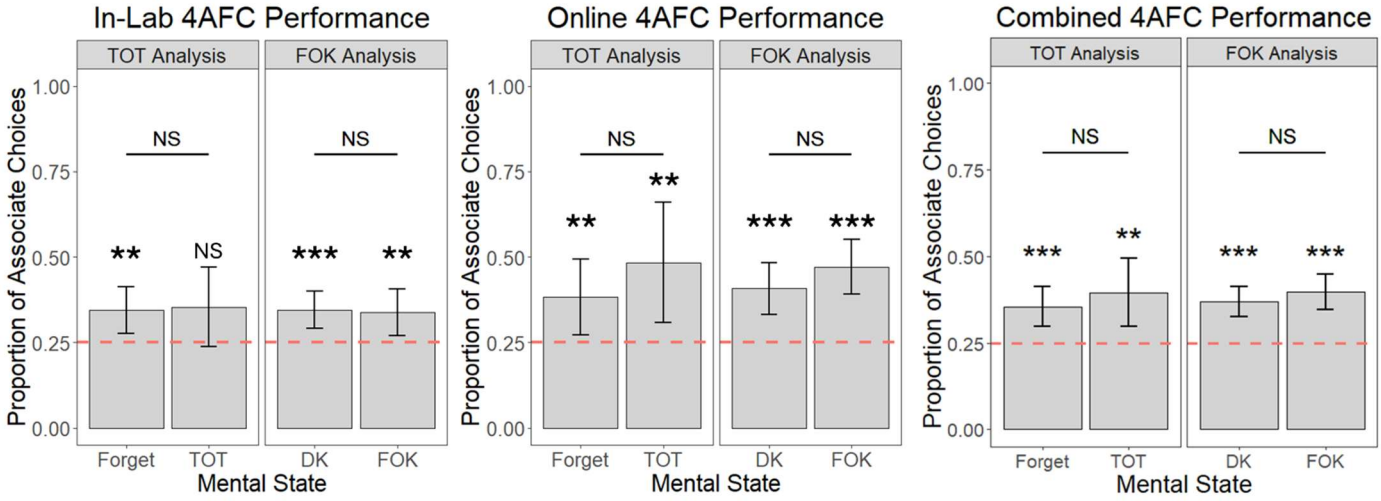
One-way within-subjects ANOVAs showed a significant main effect of serial position on recall performance for the in-lab version of the experiment, $F(5, 240) = 4.58$, $p < .001$, but not

the online version, $F(5, 190) = 1.35$, $p = .246$. One-way within-subjects ANOVAs tested for a main effect of error type, with planned contrasts for omits vs. intrusions, and extralist vs. intralist intrusions. There was a main effect of error type for the in-lab version of the experiment, $F(2, 96) = 57.65$, $p < .001$, as well as the online version, $F(2, 76) = 67.45$, $p < .001$. There were significantly more omits than intrusions in-lab, $F(1, 49) = 451.23$, $p < .001$, and online, $F(1, 38) = 410.04$, $p < .001$. There were also more extralist than intralist intrusions in-lab, $F(1, 49) = 27.05$, $p < .001$, and online, $F(1, 38) = 12.65$, $p = .001$.

Figure F2 displays bar graphs of associate choice proportions and test results.

Figure F2

Comparison of 4AFC data collected in-lab and online for Experiment 1b.



Note. Both versions of the experiment showed no significant difference in performance during TOTs or FOKs, although the difference was larger in magnitude in the online version of the experiment.

For the analysis of 4AFC responses, two-sided tests of proportions were used to compare the observed proportions to random chance levels (0.25). For the in-lab version, the proportion of associate choices were significantly above chance during FORGETs, $z = 2.96$, $p = .003$, DKs, $z = 3.84$, $p < .001$, and FOKs, $z = 2.85$, $p = .004$, but not TOTs, $z = 1.93$, $p = .053$. For the online

version, the proportion of associate choices were higher for FORGETs $z = 2.64$, $p = .008$, TOTs, $z = 3.01$, $p = .003$, DKs, $z = 5.05$, $p < .001$, and FOKs, $z = 6.35$, $p < .001$.

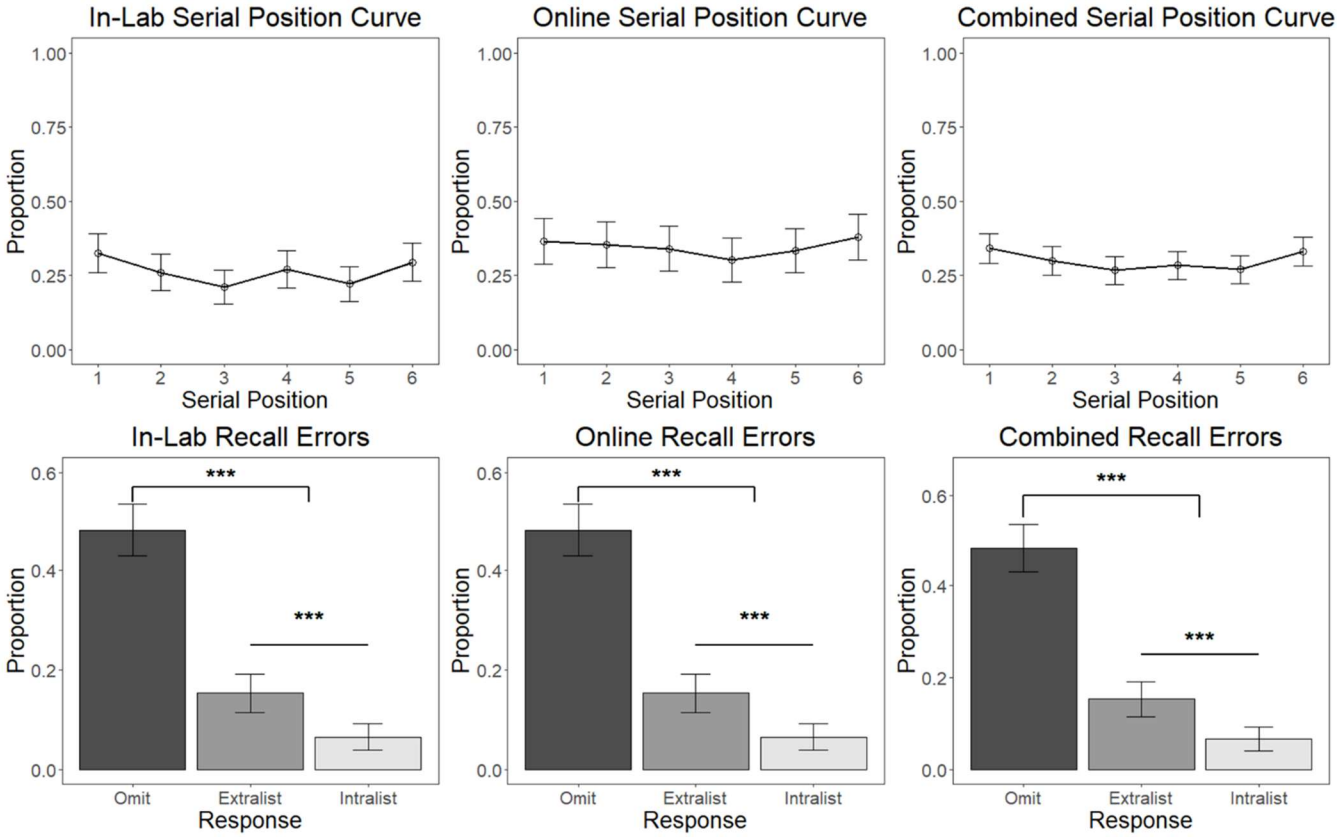
Two-sided tests of proportions also compared associate choice rates for FORGETs vs. TOTs, and DKs vs. FOKs. There were no significant changes between FORGET and TOT trials, or DK and FOK trials in either version of the experiment.

Experiment 2

Figure F3 displays the serial position curves and error proportions for both versions of the experiment.

Figure F3

Comparison of recall data collected in-lab and online for Experiment 2.



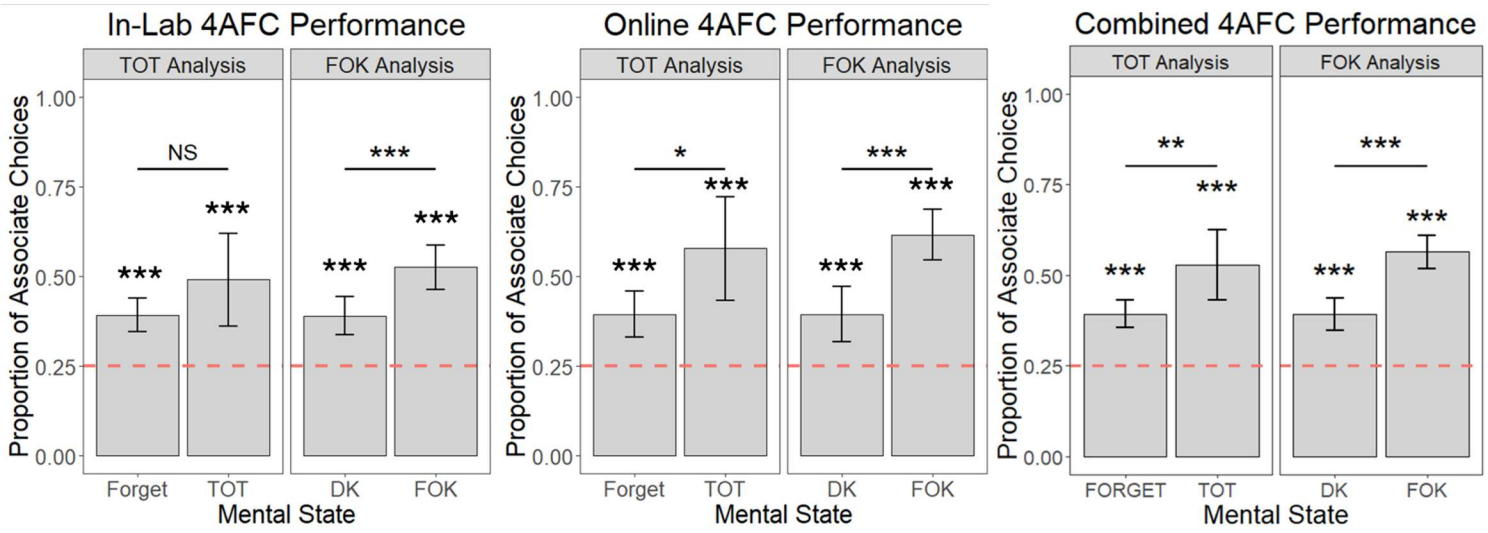
Note. Both versions of the experiment showed similar proportions of correct recall, omits, extralist intrusions, and intralist intrusions.

One-way within-subjects ANOVAs showed no significant main effect of serial position on recall performance for the in-lab version of the experiment, $F(5, 250) = 2.23$, $p = .052$, or the online version, $F(5, 190) = 0.65$, $p = .666$. One-way within-subjects ANOVAs tested for a main effect of error type, with planned contrasts for omits vs. intrusions, and extralist vs. intralist intrusions. There was a main effect for error type for the in-lab version of the experiment, $F(2, 100) = 82.36$, $p < .001$, as well as the online version, $F(2, 76) = 35.29$, $p < .001$. There were significantly more omits than intrusions in-lab, $F(1, 50) = 652.76$, $p < .001$, and online, $F(1, 38) = 298.54$, $p < .001$. There were also more extralist than intralist intrusions in-lab, $F(1, 50) = 23.99$, $p < .001$, and online, $F(1, 38) = 14.62$, $p < .001$.

Figure F4 displays bar graphs of associate choice proportions and test results.

Figure F4

Comparison of 4AFC data collected in-lab and online for Experiment 2.



Note. Both versions of the experiment showed similar increases in associate choice during TOTs and FOKs. The increase in performance during TOTs was not significant for the in-lab version of the experiment but was significant for the online version.

Two-sided tests of proportions compared the proportions of associate choices to random chance levels (0.25), and to compare associate choice rates for FORGETs vs. TOTs, and DKs vs.

FOKs. For the in-lab version, the proportion of associate choices was significantly above chance during FORGETs, $z = 6.67$, $p < .001$, TOTs, $z = 4.21$, $p < .001$, DKs, $z = 6.55$, $p < .001$, and FOKs, $z = 9.74$, $p < .001$. For the online version, the proportion of associate choices was significantly above chance levels during FORGETs, $z = 5.05$, $p < .001$, TOTs, $z = 5.08$, $p < .001$, DKs, $z = 5.05$, $p < .001$, and FOKs, $z = 11.09$, $p < .001$.

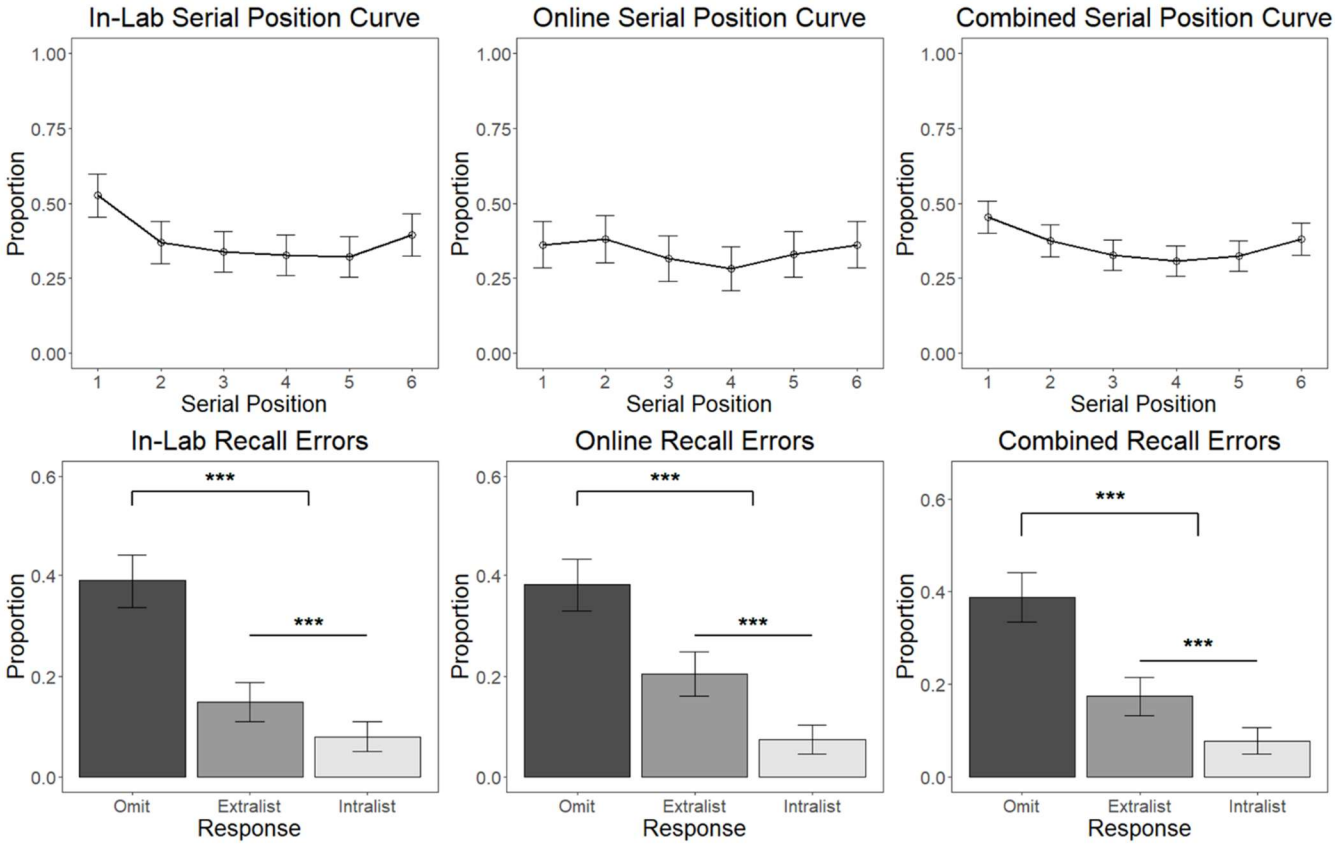
Two-sided tests of proportions also compared associate choice rates for FORGETs vs. TOTs, and DKs vs. FOKs. For the in-lab experiment, there was no significant difference between FORGETs and TOTs, $z = 1.42$, $p = .156$, but participants chose the phonological associate more often during FOKs than DKs, $z = 3.31$, $p < .001$. Online, there was a significant increase in associate choice for TOTs compared to FORGETs, $z = 2.27$, $p = .023$, and for FOKs compared to DKs, $z = 4.39$, $p < .001$.

Experiment 3

Figure F5 displays the serial position curves and error proportions for both versions of the experiment.

Figure F5

Comparison of recall data collected in-lab and online for Experiment 3.



Note. Both versions of the experiment showed similar proportions of correct recall, omits, extralist intrusions, and intralist intrusions.

One-way within-subjects ANOVAs showed a significant main effect of serial position on recall performance for the in-lab version of the experiment, $F(5, 235) = 5.68$, $p < .001$, but not for the online version, $F(5, 185) = 0.99$, $p = .426$.

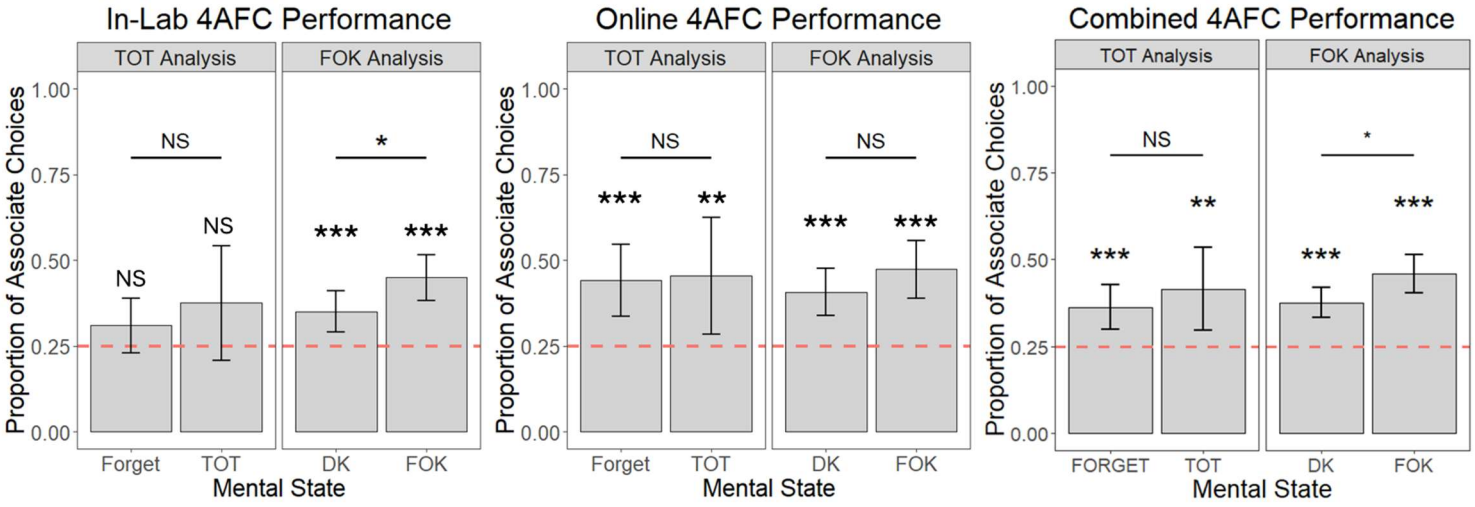
One-way within-subjects ANOVAs tested for a main effect of error type, with planned contrasts for omits vs. intrusions and extralist vs. intralist intrusions. There was a main effect of error type for the in-lab version of the experiment, $F(2, 94) = 39.62$, $p < .001$, as well as the online version, $F(2, 74) = 16.90$, $p < .001$. There were significantly more omits than intrusions in-lab, $F(1, 47) = 281.60$, $p < .001$, and online, $F(1, 37) = 136.27$, $p < .001$. There were also

more extralist than intralist intrusions in-lab, $F(1, 47) = 13.16$, $p < .001$, and online, $F(1, 37) = 29.77$, $p < .001$.

Figure F6 displays bar graphs of associate choice proportions and test results.

Figure F6

Comparison of 4AFC data collected in-lab and online for Experiment 3.



Note. Associate choices during FORGET and TOT trials were slightly lower on the in-lab version of the experiment. The increase in associate choices during FOK trials was significant for the in-lab version but not the online version.

Two-sided tests of proportions compared the proportions of associate choices to random chance levels (0.25), For the in-lab version, the associate choice proportions did not differ from chance during FORGETs, $z = 1.54$, $p = .123$, or TOTs, $z = 1.63$, $p = .103$, but were above chance during both DKs, $z = 3.72$, $p < .001$, and FOKs, $z = 6.20$, $p < .001$. For the online version, the associate choices were above chance levels during FORGETs, $z = 4.11$, $p < .001$, TOTs, $z = 2.71$, $p = .007$, DKs, $z = 5.34$, $p < .001$, and FOKs, $z = 5.90$, $p < .001$.

Two-sided tests of proportions also compared associate choice rates between FORGETs vs. TOTs, and DKs vs. FOKs. For the in-lab version, there was no difference between TOTs and

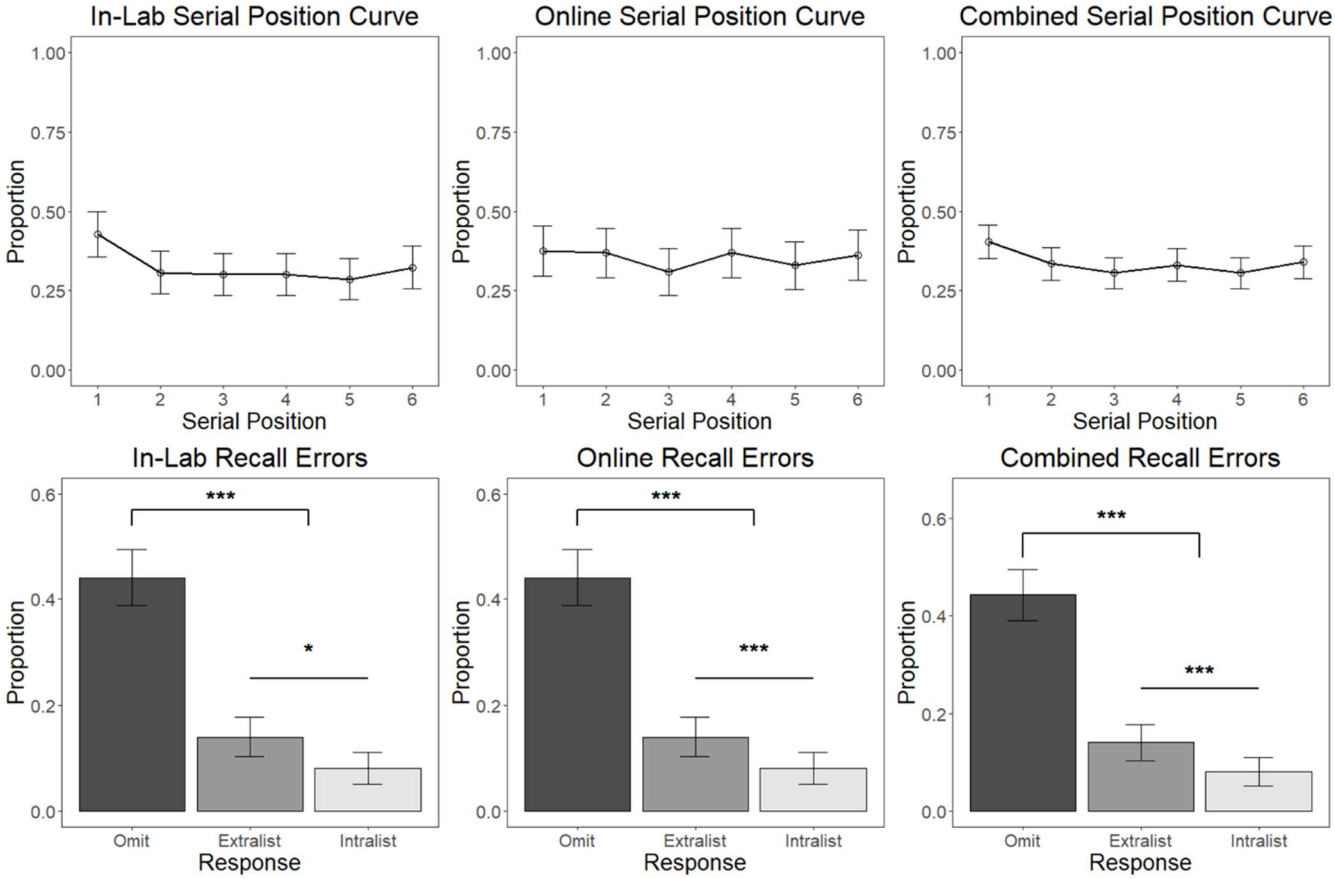
FORGETs, $z = 0.71$, $p = .310$, but the associate was chosen more often during FOKs than DKs, $z = 2.11$, $p = .035$. There were no differences in either comparison online.

Experiment 4

Figure F7 displays the serial position curves and the error proportions for both versions of the experiment.

Figure F7

Comparison of recall data collected in-lab and online for Experiment 4.



Note. Both versions of the experiment showed similar proportions of correct recall, omits, extralist intrusions, and intralist intrusions.

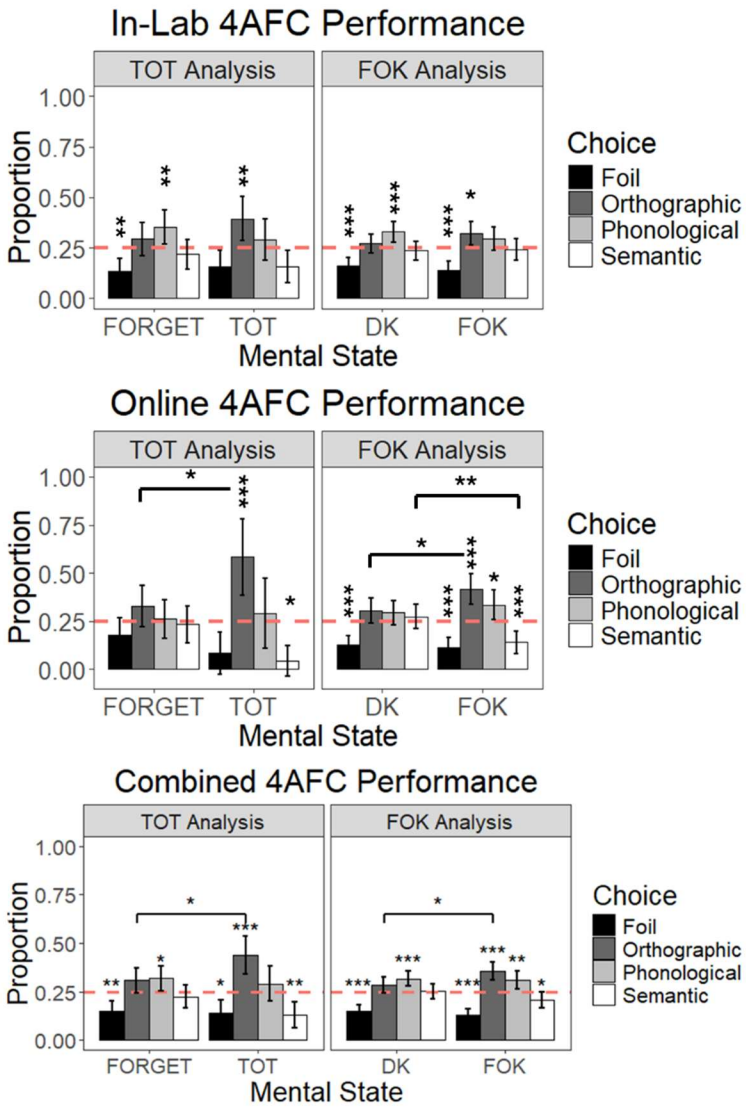
One-way within-subjects ANOVAs showed a significant main effect of serial position on recall performance for the in-lab version of the experiment, $F(5, 235) = 2.54$, $p = .029$, but not for the online version, $F(5, 185) = 0.69$, $p = .633$.

One-way within-subjects ANOVAs tested for a main effect of error type, with planned contrasts for omits vs. intrusions and extralist vs. intralist intrusions. There was a main effect of error type for the in-lab version of the experiment, $F(2, 94) = 102.90$, $p < .001$, as well as the online version, $F(2, 74) = 16.73$, $p < .001$. There were significantly more omits than intrusions in-lab, $F(1, 47) = 686.42$, $p < .001$, and online, $F(1, 37) = 160.51$, $p < .001$. There were also more extralist than intralist intrusions in-lab $F(1, 47) = 6.33$, $p = .015$, and online $F(1, 37) = 12.64$, $p = .001$.

Figure F8 displays bar graphs of 4AFC response proportions.

Figure F8

Comparison of 4AFC data collected in-lab and online for Experiment 4.



Note. The proportions of associate choices were similar in direction between both versions of the experiment. The online version of the experiment showed more extreme differences in magnitudes for TOT trials but was otherwise similar. The increase in orthographic associate choice during TOTs and FOKs was significant in the online, but not the in-lab version of the experiment, as was the decrease in semantic associate choice during FOKs.

For the analysis of 4AFC responses, two-sided tests of proportions were used to compare the observed proportions to random chance levels (0.25). Two-sided tests of proportions were also used to detect changes in proportions between TOT and FOK trials. Tables F1 and F2 contain the proportions and test results.

Table F1

4AFC choice proportions for data collected in-lab and online, with two-way tests of proportions for FORGET vs TOT trials and DK vs FOK trials.

	FORGET	**TOT**		**DK**	**FOK**	
In-Lab	Proportion	Proportion	*z*	Proportion	Proportion	*z*
Foil	0.13	0.16	0.455	0.16	0.14	0.726
Semantic	0.22	0.16	1.042	0.24	0.24	0.201
Phonological	0.35	0.29	0.920	0.33	0.30	0.873
Orthographic	0.29	0.39	1.454	0.27	0.32	1.276
	FORGET	**TOT**		**DK**	**FOK**	
Online	Proportion	Proportion	*z*	Proportion	Proportion	*z*
Foil	0.18	0.08	-	0.13	0.11	0.443
Semantic	0.23	0.04	-	0.27	0.14	2.99**
Phonological	0.26	0.29	0.301	0.29	0.33	0.767
Orthographic	0.33	0.58	2.216*	0.42	0.30	2.141*
	FORGET	TOT		DK	FOK	
Combined	Proportion	Proportion	*z*	Proportion	Proportion	*z*
Foil	0.15	0.14	0.253	0.15	0.13	0.827
Semantic	0.22	0.13	0.935	0.25	0.21	1.610
Phonological	0.32	0.29	0.487	0.32	0.31	0.22
Orthographic	0.31	0.44	2.252*	0.28	0.36	2.30*

Note. Empty cells were due to insufficient numbers of observations for statistical comparison.
* $p < .05$. ** $p < .01$.

Table F2

4AFC choice proportions for data collected in-lab and online, with two-way tests comparing observed proportions to random chance (0.25).

	Foil		Semantic		Phonological		Orthographic	
In-Lab	Prop	*z*	Prop	*z*	Prop	*z*	Prop	*z*
FORGET	0.13	2.911**	0.22	0.794	0.35	2.593*	0.29	1.111
TOT	0.16	1.854	0.16	1.854	0.29	0.795	0.39	2.914
DK	0.16	3.641***	0.24	0.548	0.33	3.319***	0.27	0.870
FOK	0.14	4.017***	0.24	0.219	0.30	1.680	0.32	2.556
	Foil		Semantic		Phonological		Orthographic	
Online	Prop	*z*	Prop	*z*	Prop	*z*	Prop	*z*
FORGET	0.18	1.419	0.23	0.338	0.26	0.203	0.33	1.554
TOT	0.08	1.886	0.04	2.357*	0.29	0.471	0.58	3.771***
DK	0.13	3.990***	0.27	0.782	0.29	1.440	0.42	1.769
FOK	0.11	3.849***	0.14	3.079**	0.33	2.309*	0.30	4.619***
	Foil		Semantic		Phonological		Orthographic	
Combined	Prop	*z*	Prop	*z*	Prop	*z*	Prop	*z*
FORGET	0.15	3.167**	0.22	0.833	0.32	2.167*	0.31	1.833
TOT	0.14	2.540*	0.13	2.771**	0.29	0.924	0.44	4.388***
DK	0.15	5.327***	0.25	0.051	0.32	3.501***	0.28	1.776
FOK	0.13	5.526***	0.21	2.036*	0.31	2.734**	0.36	4.828***

Note. * $p < .05$. ** $p < .01$.

Appendix G

Appendix G. Analyses from Experiments 1b, 2, 3, and 4 with exclusion criteria requiring participants to meet performance thresholds on the recall tasks. Participants were excluded from these analyses if they recalled less than four targets correctly, omitted responses 20 or more times, or reported 12 or more extralist intrusions.

Experiment 1b

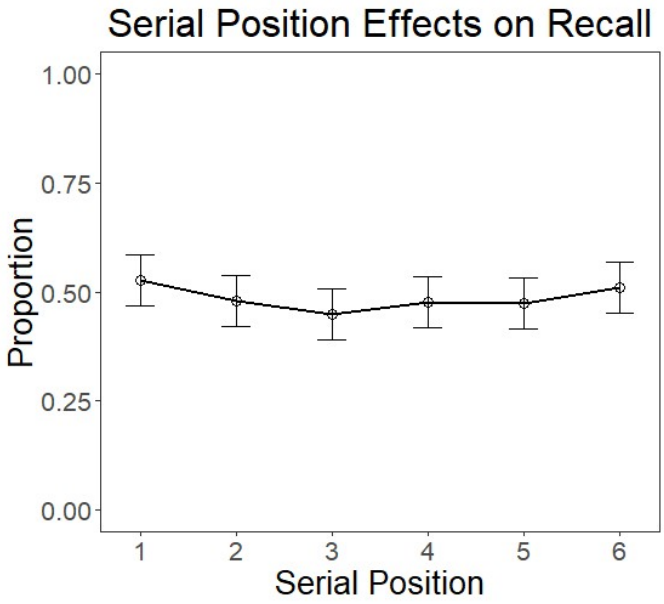
An additional 15 participants were excluded from the analysis for not meeting the required recall performance. Thirteen of these participants completed the experiment in the lab and the other two participated online. This left a sample of 73 for the following analyses (36 in-person and 37 online participants).

Recall:

The serial position curve for recall is displayed in Figure G1.

Figure G1

Recall performance.



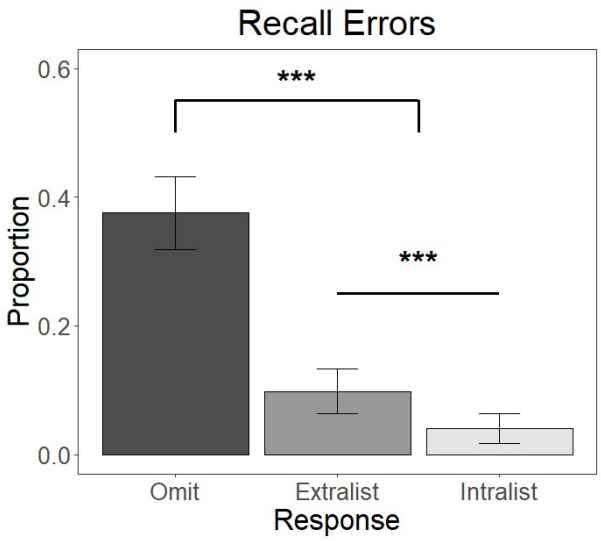
Note. The proportions of correct recalls in Experiment 1b. Error bars represent one standard error above and below the observed proportions.

The median and interquartile range for recall performance are presented due to skewed distributions of intrusion errors. The proportion of recalls was greatest (Mdn = 0.46, IQR = 0.25, followed closely by omits (Mdn = 0.42, IQR = 0.25), extralist intrusions (Mdn = 0.08, IQR = 0.08), and intralist intrusions (Mdn = 0.04, IQR = 0.08). A one-way within-subjects ANOVA showed that there was no significant main effect of serial position on recall performance, $F(5, 360) = 1.14$, $p = .338$.

Figure G2 displays the mean error proportions.

Figure G2

Recall errors.



Note. The proportion of omits, extralist intrusions, and intralist intrusions in Experiment 1b. Error bars represent one standard error above and below the observed proportions.

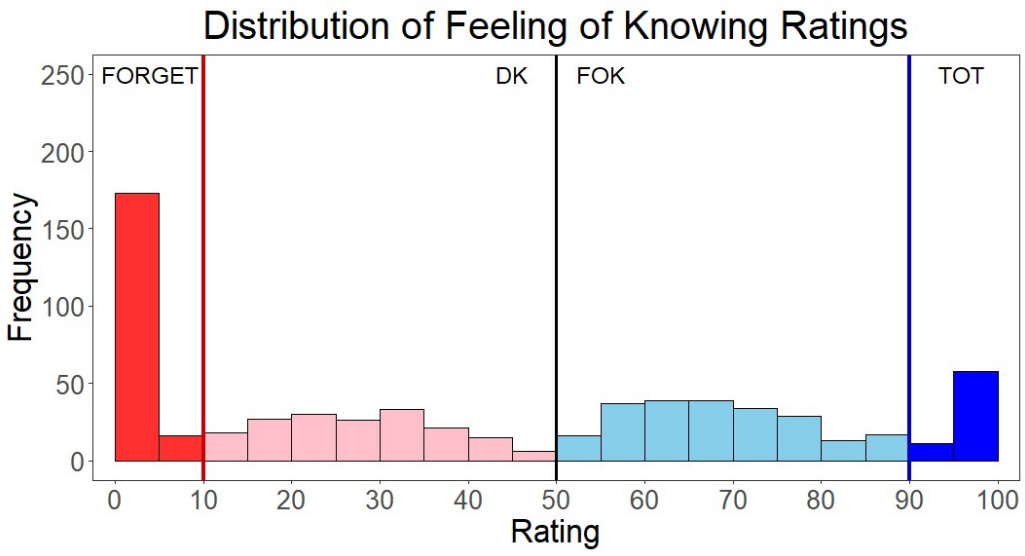
The overall error proportions were analyzed with a one-way within-subjects ANOVA, and there was a significant difference between the three error types, $F(2, 144) = 162.70$, $p < .001$. Planned contrast analysis showed that there were significantly more omits than intrusions, $F(1, 72) = 847.43$, $p < .001$, and significantly more extralist than intralist intrusions, $F(1, 72) = 22.52$, $p < .001$.

TOT ratings:

Figure G3 contains a histogram of sliding-scale ratings and the criteria for each trial type.

Figure G3

Sliding-scale responses.



Note. Ratings of 10 or below (dark red region) were categorized as FORGETs and ratings of 90 or above (dark blue region) were categorized as TOTs. For the analysis of FOKs, all ratings below 50 were categorized as DKs (full red region) and all ratings above 50 were categorized as FOKs (full blue region).

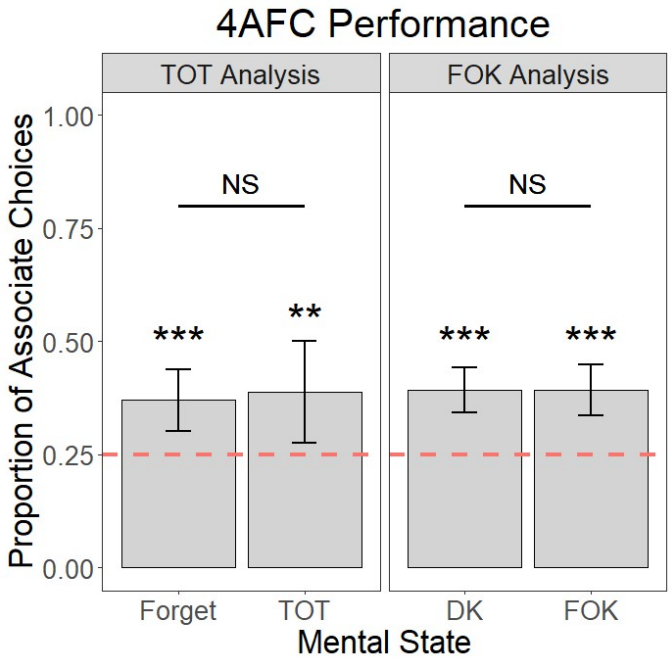
Seventy-two participants contributed a total of 658 omits when a target could not be recalled, with sliding-scale TOT ratings. Responses of 10 or below were classified as FORGET trials, and responses of 90 or above were classified as TOT trials. For the FOK analysis, all responses below 50 were classified as DKs, and responses above 50 were classified as FOKs. Using these cutoffs, there were 189 FORGETs, 72 TOTs, 361 DKs, and 293 FOKs.

4AFC:

Figure G4 displays bar graphs with the associate choice proportions for each trial type.

Figure G4

Performance on the 4AFC task.



Note. The proportion of associate choices was significantly above random chance (red line) for all trial types. The proportion did not significantly differ between FORGET and TOT trials or DK and FOK trials. Error bars indicate the 95% CI for the observed proportions.

A two-sided test of proportions showed that participants did not choose the semantic associate significantly more often during TOT trials than FORGET trials, $z = 0.28$, $p = .782$. The proportion of associate choices during TOT trials (0.39) was significantly above the random chance level of 0.25, $z = 2.72$, $p = .006$, as was the proportion during FORGET trials (0.37), $z = 3.8$, $p < .001$. For the FOK analysis, there was no significant difference in associate choice proportions between DK and FOK trials, $z = 0.81$, $p = .418$. The proportion of associate choices during FOK trials (0.39) was significantly above chance, $z = 6.36$, $p < .001$, as was the proportion during DK trials (0.39), $z = 6.15$, $p < .001$.

Experiment 2

An additional 25 participants were excluded from the analysis for not meeting the required recall performance. Sixteen of these participants completed the experiment in the lab

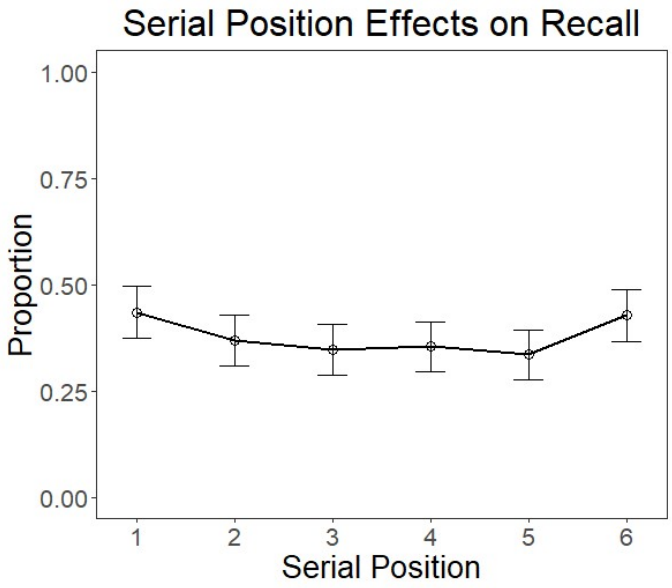
and the other 9 participated online. This left a sample of 65 for the following analyses (35 in-lab and 30 online participants).

Recall:

The serial position curve for recall is displayed in Figure G5.

Figure G5

Recall performance.



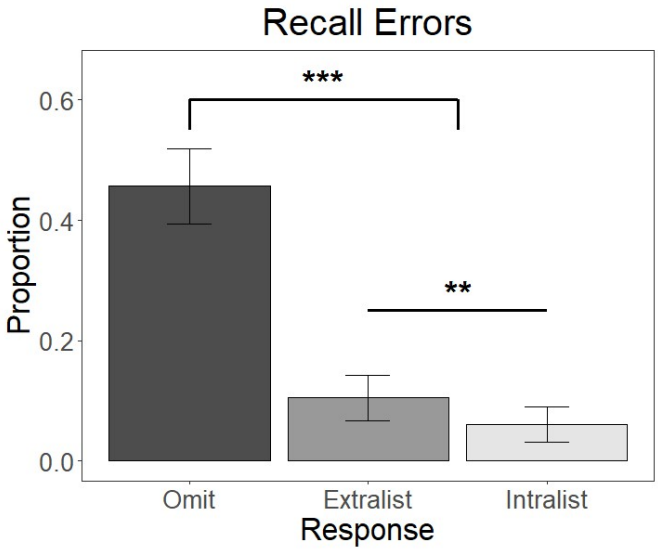
Note. The proportions of correct recalls in Experiment 2. Error bars represent one standard error above and below the observed proportions.

The median and interquartile range for recall performance are presented due to skewed distributions of intrusion errors. The proportion of omits was greatest (Mdn = 0.50, IQR = 0.25), followed by recalls (Mdn = 0.29, IQR = 0.21), extralist intrusions (Mdn = 0.08, IQR = 0.08), and intralist intrusions (Mdn = 0.04, IQR = 0.08). A one-way within-subjects ANOVA showed that there was a significant main effect of serial position on recall performance, $F(5, 320) = 2.33$, $p = .043$.

Figure G6 displays the mean error proportions.

Figure G6

Recall errors.



Note. The proportion of omits, extralist intrusions, and intralist intrusions in Experiment 2. Error bars represent one standard error above and below the observed proportions.

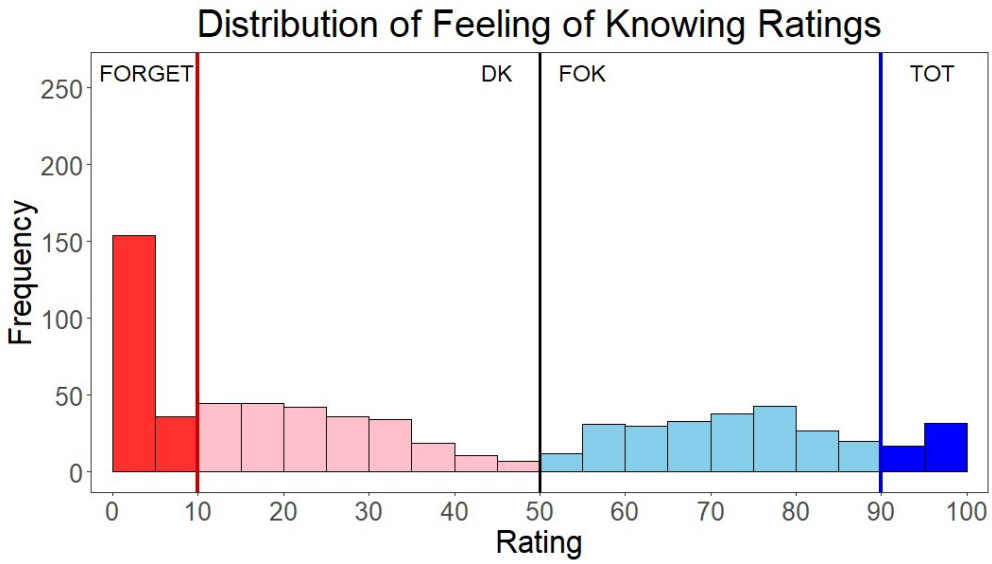
The overall error proportions were analyzed with a one-way within-subjects ANOVA, and there was a significant difference between error types, $F(2, 128) = 170.90$, $p < .001$. Planned contrast analysis showed that there were significantly more omits than intrusions, $F(1, 64) = 922.35$, $p < .001$, and significantly more extralist than intralist intrusions, $F(1, 64) = 9.71$, $p = .003$.

TOT ratings:

Figure G7 contains a histogram of the sliding-scale ratings with the criteria for each trial type.

Figure G7

Sliding-scale responses.



Note. Ratings of 10 or below (dark red region) were categorized as FORGETs and ratings of 90 or above (dark blue region) were categorized as TOTs. For the analysis of FOKs, all ratings below 50 were categorized as DKs (full red region) and all ratings above 50 were categorized as FOKs (full blue region).

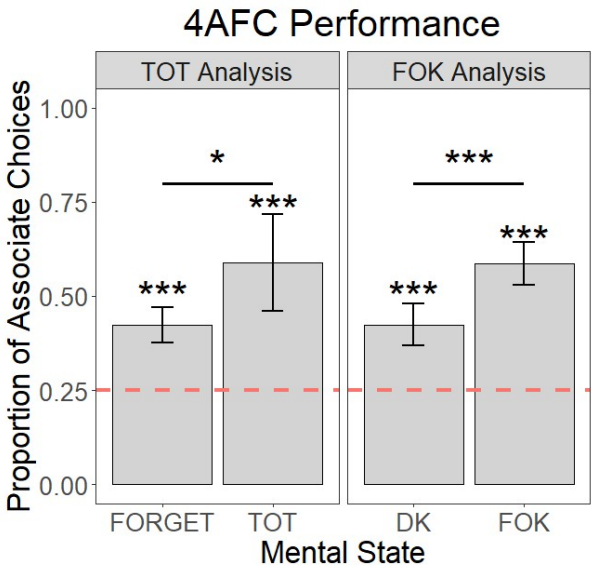
Sixty-two participants contributed 712 omits with sliding-scale TOT ratings. Responses of 10 or below were classified as FORGET trials and responses of 90 or above were classified as TOT trials. For the FOK analysis, all responses below 50 were classified as DKs and responses above 50 were classified as FOKs. Using these cutoffs, there were 429 FORGETs, 56 TOTs, 427 DKs, and 283 FOKs.

4AFC:

Figure G8 displays bar graphs with the associate choice proportions for each trial type.

Figure G8

Performance on the 4AFC task.



Note. The proportion of associate choices was significantly above random chance (red line) for all trial types. The proportion significantly increased between FORGET and TOT trials as well as between DK and FOK trials. Error bars indicate the 95% CI for the observed proportions.

A two-sided test of proportions showed that participants chose the phonological associate significantly more often during TOT trials than FORGET trials, $z = 2.34$, $p = .019$. Two-sided tests of proportions also showed that the proportion of associate choices during TOT trials (0.59) was significantly above random chance (0.25), $z = 5.86$, $p < .001$, as was the proportion during FORGET trials (0.42), $z = 8.33$, $p < .001$. For the FOK analysis, there was a significantly higher proportion of associate choice during FOK trials compared to DK trials, $z = 4.25$, $p < .001$. The proportion during FOK trials (0.59) was significantly above chance, $z = 13.01$, $p < .001$, as was the proportion during DK trials (0.42), $z = 8.30$, $p < .001$.

Experiment 3

An additional 28 participants were excluded from the analysis for not meeting the required recall performance. Fifteen of those participants completed the experiment in the lab

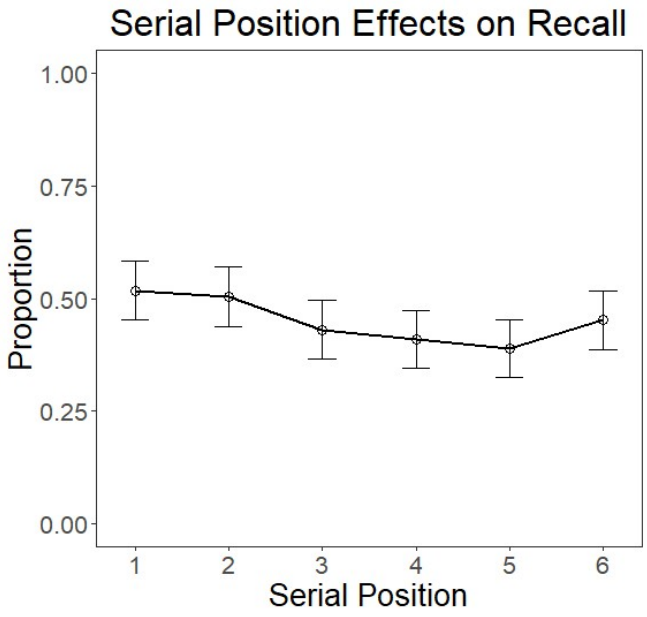
and the other 13 participated online. This left a sample of 58 for the following analyses (33 in-person and 25 online participants).

Recall:

The serial position curve for recall is displayed in Figure G9.

Figure G9

Recall performance.



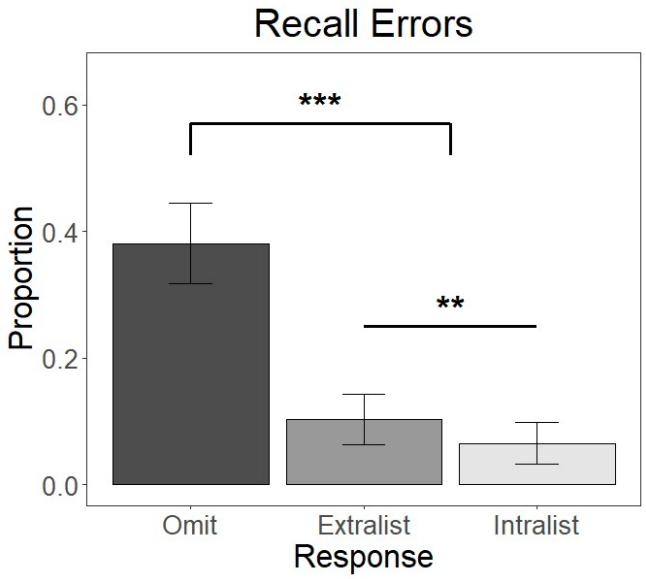
Note. The proportions of correct recalls in Experiment 3. Error bars represent one standard error above and below the observed proportions.

The median and interquartile range for recall performance are presented due to skewed distributions of intrusion errors. The proportion of recalls was greatest (Mdn = 0.42, IQR = 0.29), followed by omits (Mdn = 0.38, IQR = 0.24), extralist intrusions (Mdn = 0.08, IQR = 0.17), and intralist intrusions (Mdn = 0.04, IQR = 0.08). A one-way within-subjects ANOVA showed that there was a significant effect of serial position on recall performance, $F(5, 285) = 2.90$, $p = .014$.

Figure G10 displays the mean error proportions.

Figure G10

Recall errors.



Note. The proportion of omits, extralist intrusions, and intralist intrusions in Experiment 3. Error bars represent one standard error above and below the observed proportions.

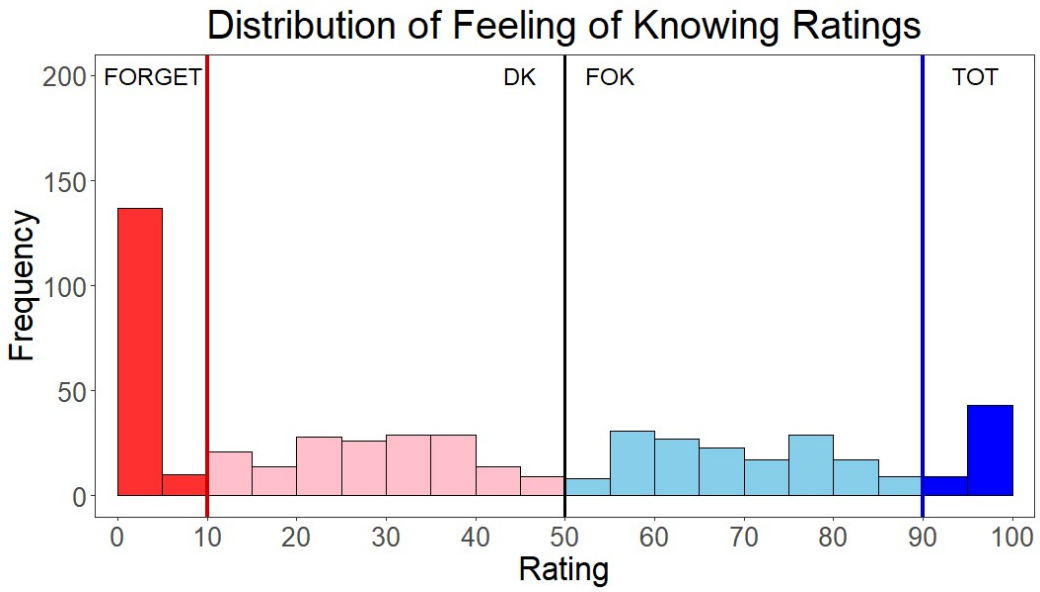
The overall error proportions were analyzed with a one-way within-subjects ANOVA, and there was a significant difference between the three error types, $F(2, 114) = 94.00$, $p < .001$. Planned contrast analysis showed that there were significantly more omits than intrusions, $F(1, 57) = 541.26$, $p < .001$, and significantly more extralist than intralist intrusions, $F(1, 57) = 6.70$, $p = .012$.

TOT ratings:

Figure G11 contains a histogram of the sliding-scale ratings with the criteria for each trial type.

Figure G11

Sliding-scale responses.



Note. Ratings of 10 or below (dark red region) were categorized as FORGETs and ratings of 90 or above (dark blue region) were categorized as TOTs. For the analysis of FOKs, all ratings below 50 were categorized as DKs (full red region) and all ratings above 50 were categorized as FOKs (full blue region).

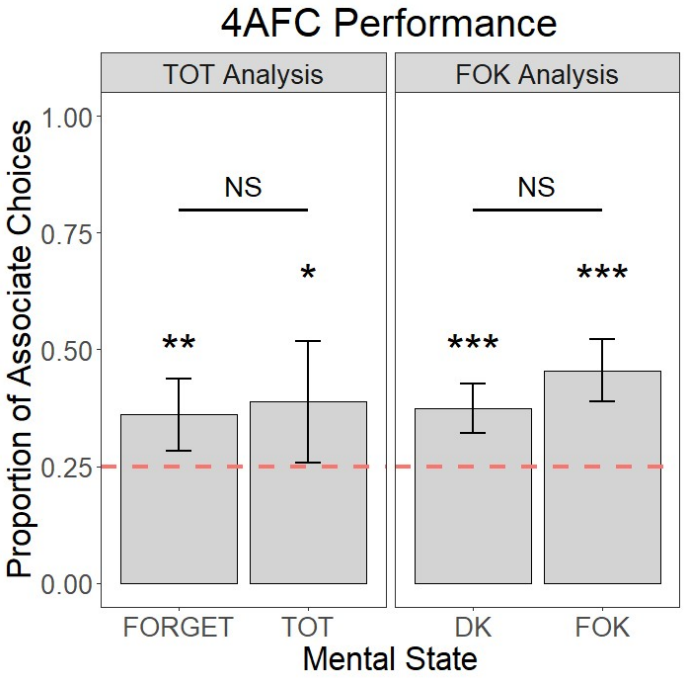
Fifty-six participants contributed a total of 530 omits when a target could not be recalled, with sliding-scale TOT ratings. Responses of 10 or below were classified as FORGET trials and responses of 90 or above were classified as TOT trials. For the FOK analysis, all responses below 50 were classified as DKs and all responses above 50 were classified as FOKs. Using these cutoffs, there were 147 FORGETs, 54 TOTs, 315 DKs, and 213 FOKs.

4AFC:

Figure G12 displays bar graphs with the associate choice proportions for each trial type.

Figure G12

Performance on the 4AFC task.



Note. The proportion of associate choices was significantly above random chance (red line) for all trial types. The proportion did not significantly increase between FORGET and TOT trials or DK and FOK trials. Error bars indicate the 95% CI for the observed proportions.

A two-sided test of proportions showed that participants did not choose the associate significantly more often during TOT trials than FORGET trials, $z = 0.37$, $p = .712$. Two-sided tests of proportions also showed that the proportion of associate choices during TOT trials (0.39) was significantly above random chance (0.25), $z = 2.36$, $p = .018$, as was the proportion during FORGET trials (0.36), $z = 3.10$, $p = .002$. For the FOK analysis, there was no difference in orthographic associate choice during FOK trials compared to DK trials, $z = 1.85$, $p = .064$. The proportion during FOK trials (0.46) was significantly above chance, $z = 6.92$, $p < .001$, as was the proportion during DK trials (0.37), $z = 5.11$, $p < .001$.

Experiment 4

An additional 16 participants were excluded from the analysis for not meeting the required recall performance. Eight of those participants completed the experiment in the lab and

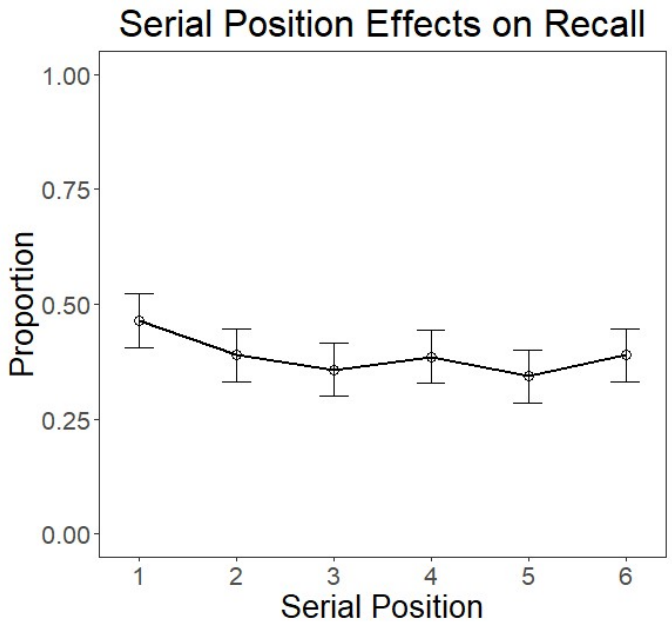
the other 8 participated online. This left a sample of 70 for the following analyses (40 in-person and 30 online participants).

Recall:

The serial position curve for recall is displayed in Figure G13.

Figure G13

Recall performance.



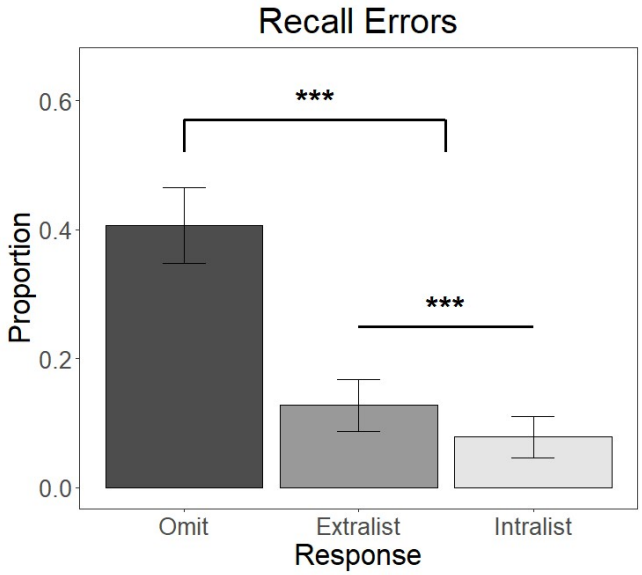
Note. The proportions of correct recalls in Experiment 4. Error bars represent one standard error above and below the observed proportions.

The median and interquartile range for recall performance are presented due to skewed distributions of intrusion errors. The proportion of omits was greatest (Mdn = 0.42, IQR = 0.24), followed by recalls (Mdn = 0.33, IQR = 0.25), extralist intrusions (Mdn = 0.10, IQR = 0.125), and intralist intrusions (0.04, IQR = 0.11). A one-way within-subjects ANOVA showed that there was a significant main effect of serial position on recall performance, $F(5, 345) = 2.47$, $p = .032$.

Figure G14 displays the mean error proportions.

Figure G14

Recall errors.



Note. The proportion of omits, extralist intrusions, and intralist intrusions in Experiment 4. Error bars represent one standard error above and below the observed proportions.

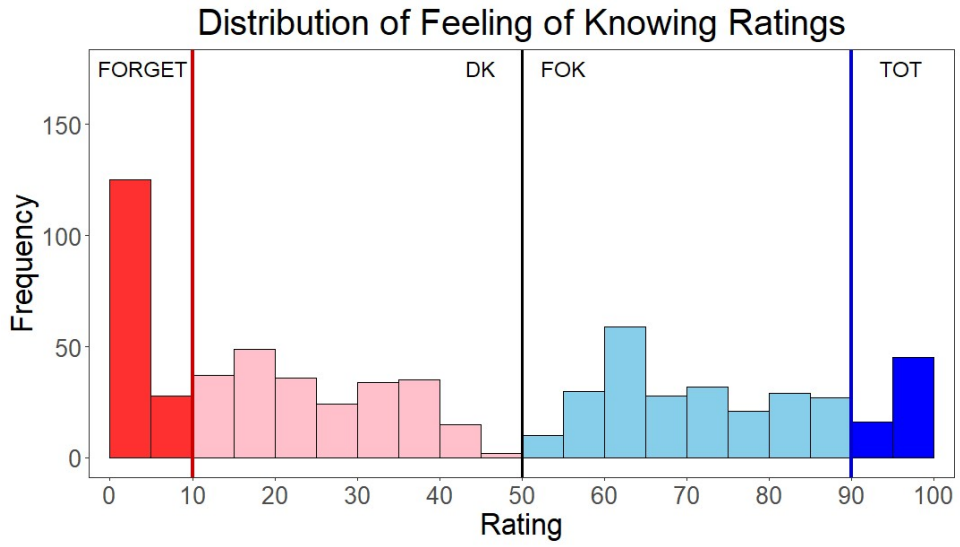
The overall error proportions were analyzed with a one-way within-subjects ANOVA, and there was a significant difference between error types, $F(2, 138) = 89.14$, $p < .001$. Planned contrast analysis showed that there were significantly more omits than intrusions, $F(1, 69) = 592.52$, $p < .001$, and significantly more extralist than intralist intrusions, $F(1, 69) = 11.53$, $p = .001$.

TOT ratings:

Figure G15 contains a histogram of the sliding-scale ratings with the criteria for each trial type.

Figure G15

Sliding-scale responses.



Note. Ratings of 10 or below (dark red region) were categorized as FORGETs and ratings of 90 or above (dark blue region) were categorized as TOTs. For the analysis of FOKs, all ratings below 50 were categorized as DKs (full red region) and all ratings above 50 were categorized as FOKs (full blue region).

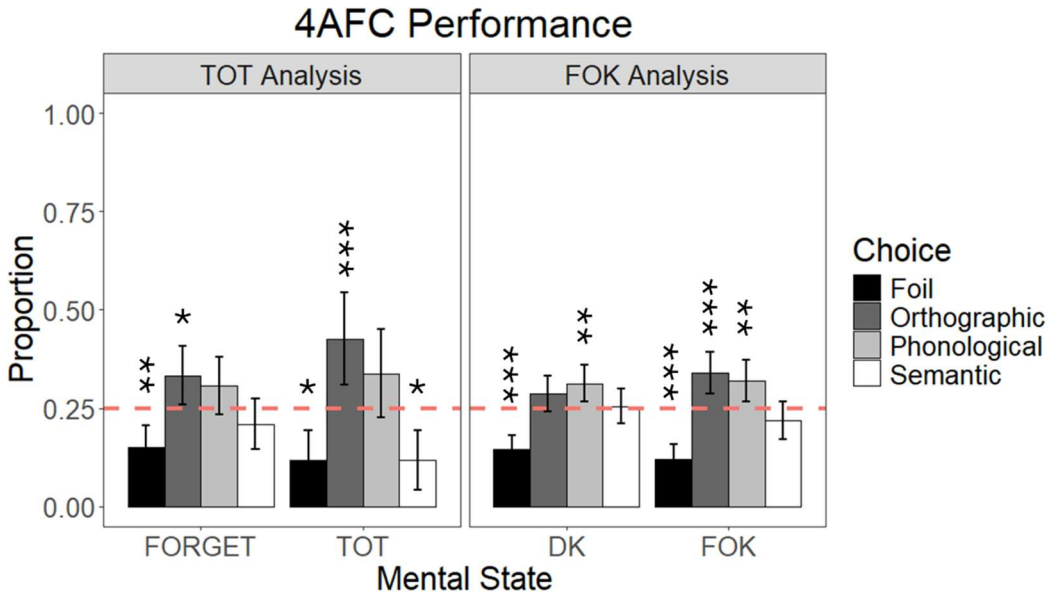
Sixty-four participants contributed a total of 682 omits with sliding-scale TOT ratings. Responses of 10 or below were classified as FORGET trials and responses of 90 or above were classified as TOT trials. For the FOK analysis, all responses below 50 were classified as DKs and responses above 50 were classified as FOKs. Using these cutoffs, there were 153 FORGETs, 68 TOTs, 384 DKs, and 297 FOKs.

4AFC:

Figure G16 displays bar graphs of the 4AFC proportions for each trial type.

Figure G16

Performance on the 4AFC task.



Note. Error bars indicate the 95% CI for the observed proportions.

Two-sided tests of proportions showed that there were no significant differences in foil choice, $z = 0.65$, $p = .519$, orthographic associate choice, $z = 1.33$, $p = .184$, phonological associate choice, $z = 0.46$, $p = .647$, or semantic associate choice, $z = 1.63$, $p = .103$, between FORGETs and TOTs. The proportion of foil choice was significantly below the level of random chance (0.25) during FORGETs (0.15), $z = 2.85$, $p = .004$, and TOTs (0.12), $z = 2.52$, $p = .012$. Orthographic associate choice was significantly above chance levels during FORGETs (0.33), $z = 2.38$, $p = .017$, and TOTs (0.42), $z = 3.36$, $p < .001$. Finally, the semantic associate was chosen significantly below chance levels during TOTs (0.12), $z = 2.52$, $p = .012$. All other tests showed no significant difference from chance.

For the FOK analysis, there was no significant difference in foil choice, $z = 0.93$, $p = .351$, orthographic associate choice, $z = 1.50$, $p = .134$, phonological associate choice, $z = 0.21$, $p = .838$, or semantic associate choice, $z = 1.10$, $p = .270$, between DK and FOK trials. Foil choice was significantly below chance levels during DKs (0.15), $z = 4.71$, $p < .001$, and FOKs (0.12), z

= 5.13, $p < .001$. Orthographic associate choice was significantly above chance levels during FOKs (0.34), $z = 3.58$, $p < .001$. Finally, phonological associate choice was significantly above chance levels during DKs (0.31), $z = 2.83$, $p = .005$, and FOKs (0.32), $z = 2.78$, $p = .005$. All other tests showed no significant difference from chance.