Exploring the Use of Phonological and Semantic Representations in Working Memory

Nelson Cowan

University of Missouri

Dominic Guitard

Université de Moncton

Nathaniel R. Greene

University of Missouri

Sylvain Fiset

Université de Moncton

Authors' Note

This research was supported by NIH Grant R01-21338 to Nelson Cowan. While working on this manuscript, Dominic Guitard was supported by a graduate scholarship from NSERC. We thank Bret A. Glass, Maximilian Stroyeck, and Caleb Burns for assistance in data collection and Nate Rose and an anonymous reviewer for helpful comments.

E-mail correspondence concerning this article should be addressed to Nelson Cowan at cowann@missouri.edu.

Open Practices Statement

The materials are included in this manuscript. The data and the R Markdowns for all experiments are available on the Open Science Framework and will be made public upon publication (https://osf.io/qg9fe/?view_only=047f0f52eac24dc0a2faf0dfbea547f3).

Abstract

In the traditional conception of working memory for word lists, phonological codes are used primarily, and semantic codes are often discarded or ignored. Yet, other evidence dictates an important role for semantic codes. We carried out a pre-planned set of four experiments to determine whether phonological and semantic codes are used similarly or differently. In each trial, random lists of 1, 2, 3, 4, 6, or 8 words were followed by a probe to be judged present in the list or absent from it. Sometimes, a probe was absent from the list but rhymed with a list item (in Experiments 1 and 2) or was a synonym of a list item (in Experiments 3 and 4). A probe that was similar to a list item was to be rejected just like other non-target probes, a *reject-similar* use (in Experiments 1 and 3) or it was to be placed in the same category as list items, an *accept-similar* (in Experiments 2 and 4). The results were comparable in the accept-similar use of both phonological and semantic codes. However, the reject-similar use was interestingly different. Rejecting rhyming items was more difficult than rejecting control words, as expected, whereas rejecting synonyms was easier than rejecting control words, presumably due to a recall-to-reject process. This effect increased with memory load. We discuss theoretically important differences between the use of phonology and semantics in working memory.

Keywords: Working memory; Short-term memory; Phonological representations; Semantic representations; Phonological similarity; Semantic similarity

Exploring the Use of Phonological and Semantic Representations in Working Memory

Working memory (sometimes called short-term memory) refers here to the small amount of information temporarily held in mind and used to carry out various cognitive tasks (Cowan, 2017). It has long been understood that working memory for printed verbal information usually includes phonological information, i.e., information about the speech sound system, and at least sometimes also includes semantic information, i.e., information about the meaning (for reviews see Cowan, 1988; Craik, 2020). What has not been made clear is how these information codes come into play in working memory, which we address by introducing different codes and tasks in four experiments. A deeper understanding of how phonological and semantic codes are used in working memory can provide valuable insight into whether these codes are always beneficial to, or perhaps sometimes are detrimental to, performance in working memory.

Exploring the Reject-similar and Accept-similar Use of Phonological and Semantic Codes

For each code, one question is, can it be ignored if doing so is helpful? To address that question, we use a *reject-similar* situation in which a list is followed by a probe item to be judged present in the list or absent from it, and the absent probes (lures) to be rejected include items that share some phonological and orthographic characteristics with the target (e.g., *boat* presented in the list and *coat* as the probe, for a *reject-rhyme* response) or items that are synonyms with the target (e.g., *boat* presented in the list and *ship* as the probe, for a *reject-synonym* response), plus neutral lures that are not designed to share phonological or semantic codes with list items. If a type of feature can be successfully excluded, then it should be no harder to respond correctly to such a probe (e.g., overlooking similarities in phonology in order to judge *coat* to have been absent from the list containing *boat*) compared to a neutral lure (e.g., judging *sock* to have been absent from the list). If the feature cannot be ignored, however, then the similarity to a target

should be detrimental, resulting in more failures to correctly reject similar lures than neutral lures. If phonological information dominates over semantic information in working memory (e.g., Baddeley, 1966a, 1966b; Craik, 2020), one would expect that it cannot be excluded, whereas semantic information should be easier to exclude. This question is examined in Experiment 1 for phonological/orthographic information and in Experiment 3 for semantic information. (To foreshadow the results, unexpectedly, the semantic information was helpful rather than harmful, in ways that allow a revised theoretical conception of the use of semantic information.)

For each code, a second question is how well it can be used when the task demands it. To examine this question, in an *accept-similar* task, we instructed participants to respond in the same way to targets (e.g., *boat*) and to probes that were not in the list but rhymed with a list item (e.g., *coat*, in Experiment 2, for an *accept-rhyme* response) or to probes that had the same meaning as a list item (e.g., *ship*, in Experiment 4, for an *accept-synonym* response), but to continue to reject neutral lures. This similarity would have to be used despite stark differences between the target and special lure in meaning (Experiment 2) or in phonology and lexical identity (Experiment 4). The literature we consider led us to expect that it might be difficult to ignore phonology in order to use semantics for a working memory task.

Our reject-similar and accept-similar tasks resemble the *accept targets only* and *accept targets plus related distractors* conditions, respectively, of conjoint recognition tasks (e.g., Brainerd et al., 1999). However, our decision rules apply to different types of similarity codes (phonological or semantic, but not both) and in working memory procedures, as opposed to long-term memory procedures typical in conjoint recognition tasks.

Putting four experiments together, we endeavored to judge the reject-similar and accept-similar use of phonological and semantic codes, allowing a comparison of the range of roles of

the two kinds of codes. We did so across a range of list lengths so that we could examine the degree to which the uses of each code depended on free working memory capacity.

Phonological and Semantic Codes in Working Memory: A Brief Review

Predominance of Phonological Codes

We asked, essentially: how much does phonological (or semantic) information get in your way in situations in which it might be better excluded? How well can you categorize probe items on the basis of only phonological (or only semantic) information if the task requires it? The most relevant research motivating the work examined the role of phonological and semantic codes in memory tasks. Much of this work indicates that phonological information predominates in working memory procedures (for early evidence of the use of phonological codes even for printed materials see Conrad, 1964; Wickelgren, 1965) and that semantic information dominates in long-term procedures. For example, Baddeley (1966a) showed enormous detrimental effects of phonological similarity versus tiny effects of semantic similarity in the immediate recall of word lists whereas, in delayed recall, it was semantic similarity that had a detrimental effect (Baddeley, 1966b; cf. Matzen et al., 2011). Craik and Lockhart (1972) proposed that shallow, orthographic and phonological codes can suffice for immediate recall but that deep, semantic encoding is needed for longer-term recall.

Presence of Semantic Codes in Immediate Memory

Other research shows that a semantic trace is not completely absent in the short-term representation. It is rapidly generated (Potter, 1993) and can even cause false memories in working memory procedures (Flegal et al., 2010). In a procedure that is perhaps the closest precursor of ours, Shulman (1970) presented a list of 10 words at a rate of one word every 350 ms, 700 ms or 1400 ms. Following the presentation of the last word, participants received a cue to refer to the test probe condition, which was *identical* (is the probe identical to any

presented words?), *homonym* (does the probe sound like any presented word?), or *synonym* (does the probe have the same meaning as any presented word?). For all conditions, participants had to indicate whether the test probe matched the stated condition (identical, homonym, or synonym of a list item). Overall, participants were less accurate for the condition *synonym* (e.g., *leap* versus *jump*) compared to the conditions *identical* and *homonym* (e.g., *board* versus *bored*), which did not differ one from another. For accurate trials, participants were faster to respond in the condition *identical* compared to the condition *homonym*, in which participants were in turn faster relative to the condition *synonym*. Performance was lower for synonyms than for the other two kinds of probes, but all three probe types showed comparable recency effects across 10 serial positions (and comparable, slight primacy effects), suggesting that the information is typically present. Similar to Shulman, McElree (1996) used a modified version of probe recognition task to measure speed-accuracy trade-off for lists of 5 words. Participants had to identify if a probe word was in the list, a word rhyming with a word in the list, or similar in meaning to a word in the list (i.e., *same*, *rhyming*, or *synonym*). Participants were better in the *same* judgment compared to the other judgments, *synonym* and *rhyme*, which did not differ one from another. Participants also responded faster for correct responses in the "same" condition compared to the other conditions, which did not differ one from another.

The results of Shulman (1970) and McElree (1996) suggest that there is no advantage for semantic or phonological information considered alone. They can be interpreted as follows. In the identical condition, for a correct response, participants are required to identify that the probe is both identical in sound (phonological information) and meaning (semantic information), whereas the other conditions involve one sameness and one difference. The results do not suggest that people fail to use those codes, but rather that people may ordinarily use or consider them together. In a review, Shulman (1971) concluded that for the random lists of words typically used

in working memory tasks, semantic codes tended not to be used unless the task required it or the list was presented at an unusually slow pace. Shulman suggested that phonological coding occurs much more quickly than semantic coding. In another review, Baddeley (1972) concluded that the use of semantic codes could occur but that these were useful only when there were retrieval rules stored in long-term memory that could be applied. There has since been further research showing the importance of semantic codes in working memory for lists (e.g., McElree, 1996; Potter, 1993).

Comparison of Code Use

In the phonological similarity effect, words that are phonologically similar are more likely to be recalled incorrectly, most often because they are recalled in the wrong order (e.g., Poirier & Saint-Aubin, 1996). However, when stimulus lists rhyme, a beneficial effect for item recall is observed (Nimmo & Roodenrys, 2004). There has been comparable research on semantic, as well as phonological, similarity effects (e.g., Chubala et al., 2019; Crowder, 1979; Guerard & Saint-Aubin, 2012; Murdock, 1976; Neale & Tehan, 2007; Poirier & Saint-Aubin, 1995; Saint-Aubin & Poirier, 1999; Saint-Aubin et al., 2005; Tse, 2009; Tse et al., 2011). The predominant semantic effect is facilitation as a result of similarity between list items. In a recent review (Ishiguro & Saito, 2020 online ahead of print), it was proposed that "semantic similarity has a detrimental effect on both serial reconstruction and serial recall, while semantic association, which is correlated with semantic similarity, contributes to an apparent facilitative effect." Thus, studies on the semantic similarity effect, like the phonological similarity effect, have predominantly shown a detrimental effect on order and a beneficial effect on item information (e.g., Murdock, 1976; Saint-Aubin et al., 2005; Tse, 2009; Tse et al., 2011; but for conflicting results see Neale & Tehan, 2007; Saint-Aubin & Poirier, 1999; Tehan, 2010). The comparability of how phonological and semantic codes are used, combined with the predominance of the phonological code in

working memory for word lists, motivated this study, which was designed to explore some important, remaining uncertainties about how phonological and semantic codes are used together.

Rationale of the Present Study

Experimental Manipulations

No prior study has examined both reject-similar and accept-similar situations for both phonological and semantic codes. Here we do so with a focus on item information, using a probe recognition task based on the seminal work of Sternberg (1975). In the reject-similar case, one might expect that it is helpful to ignore or exclude the similarity, and doing so might be easier when semantic information is to be excluded, given that each list is a printed word sequence that can be easily articulated but is not semantically organized.

It is also possible for similarity to have a helpful effect in the reject-similar situation. This was unanticipated but did occur in Experiment 3. The way that this could occur is if the list length is long enough that not all list items are remembered. In that case, a similar probe could serve as a reminder of the target item in the list. In Experiment 3, for example, the probe word *conviction* could serve as a retrieval cue for *belief*, and knowledge that words in a list were semantically diverse could facilitate the correct judgment that *conviction* was not in the list. This kind of process is called *recall-to-reject* (e.g., Rotello et al., 2000), a process "in which mismatching information that is retrieved from memory is used to reject test foils that are similar to studied items" (Rotello et al., p. 67).

The accept-similar use of phonological and semantic codes is examined in the present Experiments 2 and 4, respectively. In those experiments, the instructions were altered so that a "yes" response would indicate that the probe was either identical to the target or similar to it (phonologically similar in Experiment 2; semantically similar in Experiment 4). If one code were

used exclusively, it should be nearly as easy to judge that code to be present as it would be from an identical item. For example, it should be almost as easy to judge that *relief* and *belief* are similar or identical as it is to judge that *belief* and *belief* are similar or identical. If, however, participants inevitably use both phonological and semantic codes, it should be much more difficult to say "yes" to *relief* and *belief* being similar because of their different meanings. A comparable logic applies to Experiment 4: if semantic codes can be used exclusively, it should be easy to judge *belief* and *conviction* to be semantically similar or identical, but not if phonological codes are inevitably considered along with the semantic codes.

In sum, we used a series of 4 experiments with a probe recognition task to investigate separately the use of phonological codes (Experiment 1 and Experiment 2) and semantic codes (Experiment 3 and Experiment 4) under reject-similar (Experiments 1 and 3) or accept-similar (Experiments 2 and 4) task conditions. In all the experiments participants study list of 1, 2, 3, 4, 6, or 8 words. The task is illustrated in Figure 1.

Expectations for Phonological versus Semantic Code Use

Overall, we found no a priori reason to expect grossly different patterns of results for phonological versus semantic information. As noted, there are phonological and semantic similarity effects in serial recall, in which the task is to recall items in the presented order rather than attending to any kind of similarity. In keeping with Baddeley (1972), however, we suggest that the predominant use of phonological codes in short-term recall may have to do with the typical practice of using random word lists, making semantic encoding incoherent whereas phonological coding is well-suited to sequences of words, semantically random or otherwise. Thus, in contrast to random word lists, immediate recall of coherent sentence information is typically much richer in semantic information (Begg, 1971; Gilchrist et al., 2008, 2009; Sachs, 1967).

Rejecting rhymes and synonyms. Based on the aforementioned information, we expected similar patterns for reject-rhyme and reject-synonym situations but lower performance levels for reject-rhyme than for reject-synonym because it could be more difficult to ignore and exclude the phonological information shared with the target, when phonological information is needed for short-term maintenance of the list during the decision process (e.g., Baddeley, 1966a, 1966b; Craik, 2020; Matzen et al., 2011). The semantic information is presumably needed less for list maintenance, which also could contribute to the ability to use semantics for the aforementioned recall-to-reject process.

Accepting rhymes and synonyms. Based on results of Shulman (1970), McElree (1996), and others discussed earlier, we expected comparable patterns of responses in the accept-rhyme and accept-synonym situations, albeit with better performance in the accept-rhyme condition because only semantic information had to be excluded and it is considered weaker than phonological information within random word lists.

Finally, it would also be possible to discuss the role of familiarity and recollection (e.g., Jacoby, 1991) in our procedure. That dissection of memory strength, however, is not central to our approach and we save it for a section within the General Discussion.

Experiment 1

Experiment 1 was designed to explore the reject-similar use of phonological codes in working memory. Participants had to identify as quickly and as accurately as possible if a probe was identical to or different from one of the words previously presented in a study list of 1, 2, 3, 4, 6, or 8 words. The probe was either a word presented before in the list, a word rhyming with a word in the list, or a different word not rhyming with a word in the list.

Method

Participants. The final sample was composed of 36 undergraduate students who

volunteered from University of Missouri to participate for research credits. The mean age of the participants was 18.75 (*SD* = 1.05, range 17–21); 24 self-identified as female and 12 as male. Four participants were removed and replaced for not following properly the instructions of the experiment.

Materials. All experiments were programmed with E-Prime 2.0 (Schneider et al., 2012). The stimuli were 64 word pairs varying between one and four syllables, taken from the 100 triplets of McElree (1996). The triplets of McElree correspond to an item with a corresponding rhyme and synonym. The stimuli for this and all subsequent experiments are presented in Appendix A. In this experiment, we only used pairs which were composed of a study item and a corresponding rhyme that served as a probe and was never used as a study item. For instance, the study item could be “alone” and the corresponding rhyme “phone”. All words and texts, unless otherwise mentioned, were presented in black, uppercase, 20 points Times New Roman font, at the center of a computer screen on a silver background.

Design. A 3 × 6 repeated-measure design was implemented with the following two repeated-measure factors: probe type (same, rhyme, different) and memory set size (1, 2, 3, 4, 6, 8). The experiment was divided in six blocks of 72 trials, each corresponding to a memory set size. In each block, there was an equal number of trials for each of the three possible probe types (24 same trials, 24 rhyme trials, 24 different trials). The study items were randomly drawn on each trial from the 64 possible words. The same and rhyme probes were drawn equally often for each serial position of each memory set size. The different probe was randomly drawn from the remaining study items that were not presented in the current trial. The order of the memory set size was counterbalanced across participants. The probe type conditions were randomized within each memory set size block and for each participant.

It is noteworthy that the frequent re-use of stimuli in this experiment ensures that a high level of proactive interference occurs, which should minimize answers on the basis of familiarity of the items and maximize responding on the basis of recollective aspects of working memory. Indeed, it has been demonstrated that the finding of severe capacity limits in working memory depends on the presence of this proactive interference from trial to trial (e.g., Endress & Potter, 2014).

Procedure. All participants were tested in one experimental session lasting approximately 45 minutes in a sound attenuated booth. Participants were informed that they should try to be as accurate and as fast as possible for each trial. They were further informed that there would be six blocks of 72 trials in the experiment. The participants were able to take a short break between each block and each trial. Before each block, participants were instructed to put their left index on the "z" key and their right index on the "m" key of the keyboard.

The progression of a typical trial is shown in Figure 1. The participants initiated each block and each trial by pressing the "space bar" key. After the initiation of the trial, participants first saw a fixation cross "+" for 500 ms on the center of the screen. Immediately after the fixation cross, the to-be-remembered words were presented at a rate of one word per 450 ms (400 ms on, 50 ms off) at the center of the screen. The presentation of the last word was immediately followed by a visual mask composed of random characters that was presented on one line at the center of the screen (e.g., $ _? & @ + - & & _) that was accompanied by a 20 ms tone signalling that a probe would soon be presented for the test (see Figure 1). Twenty-four visual masks were created with 19 characters in each of them. These masks were each presented three times per block, with the same masks across conditions. Immediately after the mask, a probe was presented that was either a study word (same-probe condition), a word that rhymed with a study word (rhyme-probe condition) or a word that was not presented and not rhyming

with a study word (different-probe condition). The test probe was presented until the participant's response. If the word was the same as a studied word, the participants had to press the "z" key with their left index finger and if the word was not the same as a studied word, the participants had to press the "m" key with their right index finger. After their response, the participant received a reaction time feedback in blue for 1000 ms (see Figure 1). More specifically, the feedback corresponds to the reaction time of the participant in seconds and was identical for correct and incorrect answers. Participants did not receive feedback regarding the accuracy of their response. The trials were identical for each memory set size block (1, 2, 3, 4, 6, 8) except for the number of words presented in the sequence.

Data Analysis

In all experiments, we used Bayesian inferential statistics. Several reasons motivated the use of this statistical framework for analyzing our data. It can express results in terms of the probability distribution of a parameter value given prior assumptions and new data, combined to yield what is termed a posterior distribution, which is relatively straightforward to interpret. One can observe the relative probabilities of null and non-null hypotheses under these conditions, with Bayes factors. This situation is unlike the frequentist, null hypothesis statistical testing approach, in which one cannot obtain positive evidence favoring the null. Alternatively, using a posterior distribution of each parameter value as we do, we can quantify the range of values of the parameter best supported by the data, whether or not it includes zero (Kruschke & Liddell, 2018).

There are additional reasons why a Bayesian approach is especially useful for the present project. First, our participants were measured multiple times on the same variables that were not normally distributed, an issue addressed through complex multilevel models for non-Gaussian distributions. Second, our sample sizes were relatively small regarding these multilevel models (see Atkins, Baldwin, Zheng, Gallop, & Neighbors, 2013). Under these circumstances, it is well-

know that the Bayesian approach gives regression estimates that are more accurate than those used by other approaches (Bolker et al., 2009; McElreath, 2016). Finally, the fact that a Bayesian framework provides results that are intuitive and straightforward to interpret is especially welcome when interpreting complicated multilevel models like ours (Andrews & Baguley, 2013; Etz & Vandekerckhove, 2016; Kruschke, 2015; Wagenmakers et al., 2018).

All statistical analyses were performed using R, version 3.6.1 or 3.6.3 (R Core Team, 2019, 2020). All Bayesian models were run using the R package brms (Bürkner, 2017), which interfaces with Stan (Carpenter et al., 2017) as a programming language for running the analyses. For each analysis and each experiment, the data, the Bayesian models, and the R scripts were integrated into different R markdown files, which are openly available on the open science framework (see author notes).

In the current study, we express the results in terms of probabilistic estimates rather than ratios between null and non-null hypotheses, or Bayes factors, as is often done (for justification, see Kruschke, 2015; van der Linden & Chryst, 2017). We report 95% highest density intervals (HDI) of the parameter estimates, and we also included the posterior probability that the difference between the conditions is larger than 0, as denoted by $\Pr_{>0}$. Whenever the 95% HDI does not include 0 and the $\Pr_{>0}$ is superior to 97.5 % (because we used two-tailed hypothesis tests), we conclude that we have at least a 95% chance that our estimates differ from 0 and reported that we have observed an effect. The HDI is more exactly the desired probabilistic expression. In comparison, a traditional confidence interval is premised on repeated sampling and does not express the probability that the true value is contained within the limits (Kruschke, 2015).

The statistical models we developed were fitted using Bayesian Markov chain Monte Carlo (MCMC) methods, which are efficient means of sampling from a probability distribution.

Two outcome variables were analyzed: accuracy and reaction time. Accuracy was a binary outcome (correct or incorrect response). To model this binary response variable, we used a Bernoulli probability density function as likelihood. Reaction time was measured in milliseconds (a continuous positive response). Reaction time was restricted to accurate trials, and reaction time responses faster than 150 ms were presumed to be anticipatory responses and were removed (see Saint-Aubin et al., 2018 for similar justification). Whenever we report trials removed from the analysis of reaction times, they include both inaccurate and anticipatory responses. As is typical, reaction times were distributed with a positive skew. To do analyses with that skew made more normal, we modeled this time response variable using a lognormal probability density function as likelihood (Hilbe, de Souza, & Ishida, 2017).

Model selection. Our independent variables of most interest were used as predictors of performance within mathematical models of the experimental situation. In all four experiments, we ran several models using both outcome variables. Given that all participants were measured multiple times in the different conditions, to determine the structure of random effects that best fit our data, we ran two null models with no predictors, only random effects. In the first null model, we added subjects as group-level effects (random effects). Thus, each participant had a unique intercept, allowing us to take into consideration the variability associated with each subject. For the second null model, we upgraded the previous model by adding a random slope effect on each participant as a function of memory set size. In this model, as suggested by McElreath (2016), we also estimated the intercept, the slope, and the covariance between the two. To determine the best random effect structure for our data, we compared these two null models using the package LOO (leave-one-out cross-validation; Vehtari, Gelman & Gabry, 2017). As a criterion, we selected the model with the highest probability of making superior predictions on new data. To select the best predictive model, we used a method termed Bayesian stacking weights (Yao, Vehtari, Simpson &

Gelman, 2018), which allows a comparison of multiple competing models. In the results section, we used $\Pr_{weights}$ to label the probability of the model with the highest predictive weight, i.e., the preferred model.

To test the main hypotheses, for each experiment, two additional models were run. In these models, we examined if our outcome variables varied as a function of probe type (different, rhyme, same) and memory set size (1, 2, 3, 4, 6, 8). In the first model, probe type (a categorical variable), memory set size (treated as a continuous variable) and their interaction were integrated as population-level effects (fixed effects). In the second model, the interaction term was omitted. These models included a random effect structure also, in a way that was the same as in null models that omitted effects. In these regression models, it was necessary to consider one level of probe type as a base level and we did so with the *different* probes. With regards to the memory set size variable, to facilitate the interpretation of its regression coefficient, the baseline level was set at 0. We also used LOO to compare the models and examine if the interaction term should be kept in the final model or not.

To replicate our analyses, one must observe some technical decisions we made. We evaluated each model with four different MCMC chains and pooled them for the final estimation of the parameters. We also used a minimum of 2000 iterations. However, given that the parameters of the lognormal models were often highly correlated, we applied some thinning (1 out of 10) to reduce the autocorrelation, resulting in a much higher number of iterations for these models. We applied weakly informative priors on the different parameters (see Gelman et al., 2013). We present the selection and justification of the priors in the R markdown documents available on the open science framework (see author notes). To ensure the credibility of our Bayesian estimations, we conducted many verifications. For instance, all $\hat{R}$ values were at 1.0,

providing support in favour of the convergence of the MCMC chains. We also used autocorrelation plots to check for the progression of Gelman and Rubin's shrink factor as a function of the number of iterations, and we ran posterior predictive checks to compare the observed data with the simulated data from the posterior predictive distributions.

Results

Figure 2 illustrates the observed accuracy (left panel) and reaction time for correct responses (right panel) for each probe type and across set sizes. The striking result for accuracy is its decline across memory set sizes for *same* trials, much more than for the other two trial types, with a slight advantage for *different* trials over *rhyme* trials, i.e., a reject-similar drawback of phonological information. Similarly, in the reaction times, faster responses for *different* trials than for the other two types support the notion that added rhyme information was disadvantageous. The statistical analyses support and elaborate upon these key observations.

Accuracy. *Model selection*. The null model with varying intercepts and slopes was better than the null model with varying intercepts only (expected log-predictive density or elpd diff = -115.0, se diff. = 15.5, $\Pr_{weights} = 0.99$). Thus, we fit our full models with varying intercepts and slopes. The full model with the interaction term had a higher probability of making superior predictions than the model without the interaction term (elpd diff. = -11.0, se diff. = 5.0, $\Pr_{weights} = 0.95$). Therefore, we concluded that there was an interaction between the fixed effects probe type and memory set size.

Group level effects (random effects). These effects, though not the main object of the analysis, provide important information about the patterns of variation that the data display. At the group level, as one can see in Table 1, our model suggests that the predicted performance of participants (random intercept) varied considerably, especially for *same* probes (see

Supplementary Figure 1A), but that there was little variation between the slope for memory set size among the participants (random slope).

Population level effects (fixed effects). Overall, as illustrated in Figure 2, participants' accuracy was better for *different* probes (*M* = .96, *SD* = .20) than *rhyme* probes (*M* = .93, *SD* = .25) and *same* probes (*M* = .86, *SD* = .35). Furthermore, performance declined as the number of words in the memory set increased. The results of our Bayesian model confirmed those trends. As presented in Table 1, when all fixed effects were at the baseline, the analysis revealed that performance was superior for *different* probes relative to *rhyme* probes ($\Pr_{>0} = 100\%$) and *same* probes ($\Pr_{>0} = 100\%$). Performance was marginally superior for *rhyme* probes relative to *same* probes in the two-tailed hypothesis tests ($\Pr_{>0} = 96.9\%$). In addition, performance declined as the number of words in the memory set size increased. However, due to the interaction between probe type conditions and memory set size, the decrease was not constant for all conditions. The decline of the performance for *different* probes and *same* probes was similar ($\Pr_{>0} = 72.7\%$) but differed from *rhyme* probes (both $\Pr_{>0} = 100\%$), which was less affected by the increase of memory set size. In other words, the performance for *same* and *different* probes declined more rapidly as a memory set size increased compared to *rhyme* probes. Supporting the latter interaction, whereas accuracy to *different* and *rhyme* probes differed when memory set size was set at the baseline (0), they did not credibly differ when memory set size was set at 8 words ($\Pr_{>0} = 63.7\%$). However, when memory set size was set at 8 words, participants were more accurate for both *different* and *rhyme* probes compared to *same* probes (both $\Pr_{>0} = 100\%$).

Reaction Time. *Model selection*. The null model with varying intercepts and slopes was better than the null model with varying intercepts only (elpd diff = -1004.5, se diff. = 50.3, $\Pr_{weights} = 0.94$). Therefore, we integrated a random intercept and slope structure to our full

models. Next, the full model with the interaction term had a higher probability of making superior predictions than the model without the interaction term (elpd diff. = -5.56, se diff. = 4.11, $\Pr_{weights} = 0.83$). Hence, we selected this model to analyze our data and concluded that there was an interaction between the fixed effects probe type and memory set size.

Group level effects (random effects). At the group-level effects, as shown in Table 1, our model suggests that the variability between participants (random intercept) and their corresponding slope for memory set size (random slope) was small (see Supplementary Figure 1B).

Population level effects (fixed effects). Overall, as shown in Figure 2, for accurate trials, participants were faster in responding to *different* probes (inaccurate and anticipatory trials removed = 5.27%, *M* = 559.40, *SD* = 241.26) than *same* probes (trials removed = 7.68%, *M* = 579.26, *SD* = 211.09) and *rhyme* probes (trials removed = 14.29%, *M* = 585.85, *SD* = 247.58), which was associated with the highest reaction time. As expected, participants' reaction time increased as the number of words in the memory set increased. Our Bayesian model confirmed this pattern of results. As shown in Table 1, when all fixed effects were at the baseline, the predicted reaction time was faster for *different* probes compared to *same* probes ($\Pr_{>0} = 99.8\%$) and *rhyme* probes ($\Pr_{>0} = 100\%$). The predicted reaction time was also faster for *same* probes compared to *rhyme* probes ($\Pr_{>0} = 100\%$). In addition, the predicted reaction time of participants increased as the number of words in the memory set size increased. However, due to the interaction between probe type conditions and memory set size, the increase in reaction time was not constant for all conditions. The increase of predicted reaction time as a function of set size for *different* and *same* probes was similar ($\Pr_{>0} = 87.6\%$) and was higher than the increase predicted for *rhyme* probes (respectively, $\Pr_{>0} = 99.7\%$, $\Pr_{>0} = 100\%$). Thus, *rhyme*

probes were less influenced by the increase of memory set size than the two other probes. In other words, the speed of response for the participants to *same* and *different* probes decreased more rapidly as memory set size increased compared to the speed of responses to *rhyme* probes.

Discussion

Experiment 1 investigated the reject-similar use of phonological information. The main finding of interest was that the reject-similar use of phonological information (*rhyme*) was disadvantageous. Specifically, participants were more accurate and faster for correct responses to *different* probes (i.e., neutral lures) relative to *rhyme* probes (which share phonological cues with target items), the condition with which it should be compared because the desired response ("reject") was the same. Thus, the reject-similar use of phonological cues in working memory results in less accurate responses and slower reaction times for item recognition.

Experiment 2

Experiment 2 was designed to explore the accept-similar use of phonological codes in working memory. In this experiment, participants had to press one key if the test probe was the same or rhyming with a study words, and a different key if the word was not the same and not rhyming with the study words. The test probe was either identical, rhyming with, or a different word not rhyming with one of the words in a previously study lists of 1, 2, 3, 4, 6, or 8 words.

Method

Participants. The final sample was composed of 36 undergraduate students who volunteered from University of Missouri to participate for research credits. The mean age of the participants was 18.69 (SD = 1.06, range 17–22); 24 self-identified as female and 12 as male. One participant was removed and replaced for not following properly the instructions of the experiment. None of the participants took part in the previous experiment.

Materials, Design, Procedure, and Data Analysis. The materials, the design, the procedure, and the data analysis were identical to Experiment 1 except for the following changes. For the experiment, if the test probe was the same or rhyming with a study word, the participant had to press with their left index finger the "z" key and if the word was not the same and not rhyming with the study words, the participants had to press with their right index finger the "m" key.

Results

Figure 3 shows the accuracy (left panel) and reaction times for correct responses (right panel) for probe type. In this experiment, although the phonological relation between the list and probe items was identical to Experiment 1, the instructions differed. In particular, in the present experiment, rhyme probes were to be classified as same as one of the list items in that the phonological form was shared. Thus, there was a disadvantage for *rhyme* probes compared to *different* probes, both in accuracy and in reaction time. These findings are elaborated in the statistical analyses below.

Accuracy. *Model selection*. As in Experiment 1, the null model with varying intercepts and slopes was better than the null model with varying intercepts only (elpd diff = -201.1, se diff. = 19.6, $\Pr_{weights} = 1.00$). Thus, we fit our full models using varying intercepts and slopes. The full model with the interaction term had a higher probability of making superior predictions than the model without the interaction term (elpd diff. = -7.8, se diff. = 4.0, $\Pr_{weights} = 0.98$). Therefore, we selected the model with the interaction term and concluded that there was an interaction between the fixed effects probe type and memory set size.

Group level effects (random effects). At the group-level effects, as one can see in Table 2, our model suggests that there was a large variation between the participants (random intercept),

especially for a few participants' accuracy to *rhyme* and *same* probes (see Supplementary Figure 2A), but the variation between the slope for memory set size among the participants (random slope) was small.

Population level effects (fixed effects). Overall, as illustrated in Figure 3, participants were more accurate to *same* probes (M = .88, SD = .33) than *different* probes (M = .84, SD = .37) and *rhyme* probes (M = .74, SD = .44). Furthermore, as expected, the performance for all three probe types declined as the number of words in the memory set increased. The results of our Bayesian model confirmed those trends. As presented in Table 2, when all fixed effects were at the baseline, performance was superior for *same* probes compared to *different* probes ($\Pr_{>0} = 100\%$) and *rhyme* probes ($\Pr_{>0} = 100\%$). Performance was also superior for *different* probes relative to *rhyme* probes ($\Pr_{>0} = 100\%$). Also, performance of participants declined as the number of words in the memory set size increased. However, due to the interaction between probe type conditions and memory set size, the decrease was not constant for all conditions. The decline of performance with increasing set size for *rhyme* probes was lower relative to that for *different* probes ($\Pr_{>0} = 99.6\%$) and *same* probes ($\Pr_{>0} = 100\%$). In other words, the *rhyme* probes were less affected by the increase in the memory set size compared to the other probes. In addition, decline of performance for *same* probes was marginally larger relative to the predicted decline for *different* probes ($\Pr_{>0} = 95.6\%$).

Reaction Time. *Model selection*. The null model with varying intercepts and slopes was better than the null model with varying intercepts only (elpd diff = -739.3, se diff. = 42.2, $\Pr_{weights} = 0.94$). Consequently, we integrated a random intercept and slope structure to our full models. The full model without the interaction term had a higher probability of making superior predictions than the model with the interaction term (elpd diff. = -0.6, se diff. = 1.9, $\Pr_{weights} =$

0.67). Therefore, we selected this model and concluded that there was no interaction between the fixed effects probe type and memory set size.

Group level effects (random effects). At the group-level effects, as shown in Table 2, our model suggests that there was little variability between participants (random intercept) and their corresponding slope for memory set size (random slope) (see Supplementary Figure 2B).

Population level effects (fixed effects). Overall, as illustrated in Figure 3, for accurate trials, participants were faster for *same* probes (trials removed = 12.83%, *M* = 546.95, *SD* = 236.70) relative to *rhyme* probes (trials removed = 27.35%, *M* = 664.74, *SD* = 366.28) and *different* probes (trials removed = 17.86%, *M* = 715.64, *SD* = 355.39). As illustrated in Figure 3, participants were taking more time to respond to *different* probes compared to the other probes. As expected, participants' reaction time also increased as the number of words in the memory set increased. Our Bayesian model presented in Table 2 confirmed those patterns of results. As shown in Table 2, when all fixed effects were at the baseline, the predicted reaction time was faster for *same* probes compared to *rhyme* probes ($\Pr_{>0} = 100\%$) and *different* probes ($\Pr_{>0} = 100\%$). The predicted reaction time was also faster for *rhyme* probes ($\Pr_{>0} = 100\%$) relative to *different* probes ($\Pr_{>0} = 100\%$). Furthermore, the predicted reaction time of participants increased as the number of words in the memory set size increased. Due to the absence of interaction between probe type conditions and memory set size, the increase in reaction time was constant for all probes.

Discussion

In Experiment 2 we investigated the accept-similar use of phonological information. In this experiment rhyme probes were to be classified as same as one of the list items for which the phonological form was shared. The accept-similar use of phonological information, like the

reject-similar use in Experiment 1, was disadvantageous. More specifically, participants were less accurate and slower for correct responses to *rhyme* probes relative to *same* probes, the condition with which it should be compared because the desired response ("accept") is the same. These findings are in line with the notion that individuals may inevitably use both phonological and semantic cues, even when the task demands the use of just phonological cues. In Experiment 3 and Experiment 4, we explored if the same pattern of results observed with phonological information will be observed with semantic information.

Experiment 3

Experiment 3 was designed to explore the reject-similar use of semantic codes in working memory. In this experiment, participants had to identify as quickly and as accurately as possible if a probe was identical or different from one of the words previously presented in a study list of 1, 2, 3, 4, 6, or 8 words. The test probe was either a word presented in the list, a synonym of a word in the list or a different word that was not a synonym with a word in the list.

Method

Participants. The final sample was composed of 36 undergraduate students who volunteered from University of Missouri to participate for research credits. The mean age of the participants was 18.72 (SD = 0.85, range 18–20); 23 self-identified as female and 13 as male. Four participants were removed and replaced for not following properly the instructions of the experiment. None of the participants took part in the previous experiments.

Materials. The material was identical as in the previous experiments except for the following changes. In this experiment we only used pairs which were composed of a study item and a corresponding synonym that served as a probe and was never used as a study item (see Appendix A).

Design, Procedure, and Data Analysis. The design, the procedure, and the data analysis were identical to Experiment 1.

Results

Figure 4 shows the accuracy (left panel) and reaction times for correct responses (right panel) to each probe type. The results here for the reject-similar use of semantic information are strikingly different from the results of the reject-similar use of phonological information in Experiment 1 (see Figure 2). In particular, Figure 4 shows that at larger memory set sizes, there was an advantage of *synonym* information relative to the *different* probe condition in both accuracy and reaction times, compared to a disadvantage of *rhyme* information in Experiment 1. This advantage suggests that synonyms provide cues that allow a recall-to-reject process to occur (Rotello et al., 2000) at larger set sizes (e.g., 6 and over), when some list items might have been forgotten. These findings are elaborated in the statistical analyses below.

Accuracy. *Model selection*. The null model with varying intercepts and slopes was better than the null model with varying intercepts only (elpd diff = -259.6, se diff. = 21.4, $\Pr_{weights} = 0.97$). Consequently, we included varying intercepts and slopes in our full models. The full model with the interaction term had a higher probability of making superior predictions than the model without the interaction term (elpd diff. = -15.1, se diff. = 5.4, $\Pr_{weights} = 0.99$). Therefore, we selected this model and concluded that there was an interaction between the fixed effects probe type and memory set size.

Group level effects (random effects). At the group-level effects, as shown in Table 3, our model suggests that the predicted performance between the participants (random intercept) was highly variable, especially for *different* and *same* probes (see Supplementary Figure 3A). With regards to the slope for memory set size (random slope), however, the variability among the

participants was small.

Population level effects (fixed effects). Overall, as illustrated in Figure 4, participants' accuracy was higher for *synonym* probes (*M* = .96, *SD* = .20) and *different* probes (*M* = .92, *SD* = .27) compared to *same* probes (*M* = .85, *SD* = .36). In addition, the performance declined for *different* and *same* probes as the number of words in the memory set increased. Those trends were confirmed by our Bayesian model. As presented in Table 3, when all fixed effects were at the baseline, performance was inferior for *same* probes relative to *synonym* probes ($\Pr_{>0} = 100\%$) and *different* probes ($\Pr_{>0} = 100\%$). Accuracy to *synonym* probes was about equivalent to *different* probes ($\Pr_{>0} = 78.2\%$). In addition, performance of participants declined as the number of words in the memory set size increased. However, due to the interaction between probe type conditions and memory set size, the decrease was not constant for all conditions. Importantly, *synonym* probes were not credibly affected by the increase of memory set size ($\Pr_{>0} = 90.6\%$). Relative to *same* probes, the decline of the predicted performance in for *different* probes was larger ($\Pr_{>0} = 99.9\%$). In other words, the performance for *different* probes was the most affected as memory set size increased, follow by *same* probes, and *synonym* probes were not credibly affected.

Reaction Time. *Model selection.* As in the previous experiments, the null model with varying intercepts and slopes was better than the null model with varying intercepts only (elpd diff = -868.9, se diff. = 45.5, $\Pr_{weights} = 0.94$). Therefore, we integrated a random intercept and slope structure to our full models. The full model with the interaction term had a higher probability of making superior predictions than the model without the interaction term (elpd diff. = -16.8, se diff. = 6.0, $\Pr_{weights} = 0.96$). We selected this model to analyze our data and concluded that there was an interaction between the fixed effects probe type and memory set size.

Group level effects (random effects). At the group-level effects, as shown in Table 3, our model suggests that there was a relatively large variability between participants (random intercept) but minimal variability between their corresponding slope for memory set size (random slope) (see Supplementary Figure 3B).

Population level effects (fixed effects). Overall, as shown in Figure 4, for accurate trials, participants were faster at responding to *synonym* probes (trials removed = 4.82%, *M* = 518.82, *SD* = 226.21) compared to *different* probes (trials removed = 8.82%, *M* = 544.53, *SD* = 302.33) and *same* probes (trials removed = 15.34%, *M* = 545.09, *SD* = 208.79), and these latter two did not differ. As in the previous experiments, participants' reaction time increased as the number of words in the memory set increased. Our Bayesian model revealed, when all fixed effects were at the baseline, that the predicted reaction time did not differ between the three probe types (see Table 3). The predicted reaction time of participants increased as the number of words in the memory set size increased. However, due to the interaction between probe type conditions and memory set size, the increase in reaction time was not constant for all conditions. The predicted increase of reaction time as a function of set size for *same* probes was marginally larger than that for *different* probes ($\Pr_{>0} = 96.9\%$), which in turn was larger than the increase for *synonym* probes ($\Pr_{>0} = 100\%$). Supporting the latter interaction, whereas the three conditions did not differ one from each other when memory set size was set at the baseline (0), they did when memory set size was at 8 words. More specifically, when memory set size was at 8 words, participants were faster at correctly responding to *synonym* probes relative to *different* probes ($\Pr_{>0} = 100\%$), and were faster at responding to *different* probes than *same* probes ($\Pr_{>0} = 99.58\%$).

Discussion

In Experiment 3, we investigated the reject-similar use of semantic information. When participants had to identify if a probe was identical or different from one of the words in the previously study list, they were faster and more accurate in the *synonym* trials relative to *different* trials, the condition with which it should be compared because the desired response ("reject") is the same. Evidence for the latter advantage in accuracy and reaction time was observed for larger memory set sizes. This suggests that the reject-similar use of semantic information is advantageous and provides cues that allow a recall-to-reject process to occur (Rotello et al., 2000) at larger set sizes (e.g., 6 and larger), when some list items might have been forgotten.

Experiment 4

Experiment 4 was designed to explore the accept-similar use of semantic codes in working memory. In this experiment, if the test probe was the same or similar in meaning with a study word, the participant had to press one key and if the word was not the same and not similar in meaning with the study words, the participants had to press another key. The probe was either a word presented in the list, a synonym of a word in the list or a different word that was not a synonym with a word in the previously study lists of 1, 2, 3, 4, 6, or 8 words.

Method

Participants. The final sample was composed of 36 undergraduate students who volunteered from University of Missouri to participate for research credits. The mean age of the participants was 18.83 (*SD* = 0.88, range 18–21); 24 self-identified as female and 12 as male. None of the participants took part in the previous experiments.

Materials, Design, Procedure, and Data Analysis. The materials, the design, the procedure, and the data analysis were identical to Experiment 3 except for the following changes. For this experiment, if the test probe was the same or similar in meaning with a study words, the participant had to press the "z" key with their left index finger. If the word was not the same and

not similar in meaning with the study words, the participants had to press the "m" key with their right index finger.

Results

The results of this experiment are shown in Figure 5 for accuracy (left panel) and reaction times for correct responses (right panel). Basically, the pattern of results for the accept-similar use of *synonyms* in the probes in this experiment was quite similar to the pattern for the accept-similar use of *rhyme* information in Experiment 2 (see Figure 3). Specifically, accuracy was lowest for *synonym* probes, and reaction time for correct responses to *synonym* probes was, on average, much slower than for *same* probes, i.e., the condition with which it should be compared because the desired response is the same. The statistical analyses below document the pattern of results.

Accuracy. *Model selection*. The null model with varying intercepts and slopes was better than the null model with varying intercepts only (elpd diff = -169.1, se diff. = 18.2, $\Pr_{weights} = 1.00$). Therefore, as in previous experiments, we included varying intercepts and slopes in our full models. The full model with the interaction term had a higher probability of making superior predictions than the model without the interaction term (elpd diff. = -5.4, se diff. = 3.7, $\Pr_{weights} = 0.90$). Consequently, we selected the model and concluded that there was an interaction between the fixed effects probe type and memory set size.

Group level effects (random effects). At the group-level effects, as shown in Table 4, our model suggests that there was a large variability between the participants (random intercept), especially for *different* and *synonym* probes (see Supplementary Figure 4A). In contrast, however, the individual performance as a function of memory set size (random slope) was very similar for all participants.

Population level effects (fixed effects). Overall, as illustrated in Figure 5, participants' accuracy was higher for *same* probes (M = .93, SD = .26) than *different* probes (M = .80, SD = .40) and *synonym* probes (M = .55, SD = .50). As can be seen in Figure 5, the performance also declined as the number of words in the memory set increased. Our Bayesian model confirmed those trends. As presented in Table 4, when all fixed effects were at the baseline, performance was higher for *same* probes relative to *different* probes ($\Pr_{>0} = 100\%$), which was superior to *synonym* probes ($\Pr_{>0} = 100\%$). Performance of participants declined as the number of words in the memory set size increased. However, the decrease was not constant for all conditions due to the interaction between probe type conditions and memory set size. More specifically, the decline of performance for *same* probes was smaller than the predicted decline of performance for *different* probes ($\Pr_{>0} = 99.7\%$) and *synonym* probes ($\Pr_{>0} = 100\%$), and the decline of performance for the latter two probes did not credibly differ ($\Pr_{>0} = 87.3\%$).

Reaction Time. *Model selection*. The null model with varying intercepts and slopes was better than the null model with varying intercepts only (elpd diff = -548.0, se diff. = 37.2, $\Pr_{weights} = 0.93$). For the full models, we integrated a random intercept and slope structure. The full model without the interaction term had a higher probability of making superior predictions than the model with the interaction term (elpd diff. = -1.3, se diff. = 1.9, $\Pr_{weights} = 0.71$). We selected this latter model to analyze our data and concluded that there was no interaction between the fixed effects probe type and memory set size.

Group level effects (random effects). At the group-level effects, as can be seen in Table 4, our model suggests that there was a credible variability between participants (random intercept), which was more apparent for *different* and *synonym* probes (see Supplementary Figure 4B). Variability among the participants as a function of memory set size (random slope), however,

was small.

Population level effects (fixed effects). Overall, as shown in Figure 5, for accurate trials, participants were faster to respond to *same* probes (trials removed = 8.91%, *M* = 551.91, *SD* = 451.57) compared to *synonym* probes (trials removed = 46.24%, *M* = 793.61, *SD* = 748.67) and *different* probes (trials removed = 19.97%, *M* = 798.51, *SD* = 449.89). As in the previous experiments, participants' reaction time increased as the number of words in the memory set increased. Our Bayesian model revealed, when all fixed effects were at the baseline, that the predicted reaction time was slower for *different* probes relative to *synonym* probes ($\Pr_{>0} = 100\%$), which was in turn slower than *same* probes ($\Pr_{>0} = 100\%$). The predicted reaction time of participants increased as the number of words in the memory set size increased. Due to the absence of interaction between probe type conditions and memory set size, the predicted increase in reaction time was constant for all conditions.

Discussion

In Experiment 4, we explored the accept-similar use of semantic information. Consistent with accept-similar use of phonological information in Experiment 2, the accept-similar use of semantic information was disadvantageous. Specifically, participants were slower and less accurate in at responding to *synonym* probes relative to *same* probes, the condition with which it should be compared because the desired response is the same. Thus, as with the accept-rhyme situation of Experiment 2, findings from the accept-similar task of Experiment 4 are in line with the notion that individuals do not restrict their basis of selection to a single type of feature that defines both the identical and the similar probe (phonological features in Experiment 2; semantic features in Experiment 4). Instead, they make use of the identity of the probe to a list item and more quickly accept identical probes, compared to similar probes.

Summary and Analyses Across all Four Experiments

Empirical Summary

Our focus in the four experiments just reported was to examine whether using phonological or semantic cues in an accept-similar or reject-similar way affects performance on a working memory probe recognition task. To recapitulate the results, when phonological cues were used in a *reject-similar* manner (i.e., a situation in which using phonological cues could not always differentiate a word as "same" or "different," as in Experiment 1, when only target items were to be classified as "same"), the most noticeable effect was that accuracy to target items declined precipitously at the largest set sizes and that accuracy to rhyming lures was worse than accuracy to novel distractors at almost all set sizes (see Figure 2). These findings suggest that the reject-similar use of phonological cues was disadvantageous because the overlapping phonological representations of rhyming lures and target words resulted in more confusability among these items and worsened performance.

Interestingly, however, when semantic cues were used in a reject-similar manner (as in Experiment 3, where synonyms were to be classified as "different"), although there was a decline in performance for target items with increasing set size, accuracy for synonyms was *higher* than accuracy to novel items at larger memory set sizes (see Figure 4). Such findings suggest that the reject-similar use of semantic cues can be advantageous, presumably by a recall-to-reject process for synonyms that results, correctly, in classifying synonyms as "different" from originally-studied words, at least at larger set sizes. An example is receiving the probe word *ship*, having it prime memory for the list word *boat*, and realizing that no synonyms were to be found in the list so that *ship* should be judged absent from the list.

Phonological cues used in an accept-similar manner were not fully effective (i.e., in Experiment 2, actively using phonological information about a word to classify the probe into one response category if it shared phonemes with an originally-studied word, where both targets

and rhyming lures were to be classified as "same"). This was the case especially with increasing set size, as response accuracy to rhyming lures was worse than that of the other trial types and response speed was slower than for probes identical to a list item (Figure 3). Similarly, in Experiment 4, the accept-similar use of semantic cues was disadvantageous. This effect was found at all set sizes and especially at the largest set sizes (see Figure 5).

Cross-Experiment Analysis

So far, we have not yet examined the different tasks and materials using a common metric. We applied a signal detection framework (SDT; Green & Swets, 1966) to get a better idea of sensitivity free of bias that might differ between conditions or experiments. In recognition memory studies in which participants must decide if a probe item (*e.g.,* a word) is included in a previously memory set of words or not, SDT assumes that the presentation of a probe (signal in noise or just noise) is represented in memory as a point along an underlying probability distribution of memory strength (Snodgrass & Corwin, 1988).

Analytic Method

In SDT, responses on each trial of a choice discrimination task depend on the strength of signal provided by the test probe and the participant's decision criterion along a strength continuum (cf. Criss, 2009). If the signal exceeds the participant's criterion, the response "old" is given; otherwise, the participant responds "new." Given that previously-studied items have been more recently encountered, thereby resulting in presumably stronger memory traces than new items, the general assumption of SDT is that old items elicit stronger signals than new items. However, new items can elicit erroneous signals as well, such that the decision process is not perfect. SDT assumes that old and new items form their own distributions, which reflects this uncertainty, and that each distribution is normal, with the distribution for old items shifted farther

to the right along the strength continuum (e.g., Green & Swets, 1966). The difference in the peaks of the two distributions corresponds to the discrimination (or memory strength) metric, *d'*.

Traditionally, SDT has been used for yes-no and ratings task experiments comparing recognition judgments to old and new items. However, in our experiments, test probes varied in how similar they were to originally-studied items, and although there were only two response options in each experiment, there were all together three types of probes per experiment, which we can classify as Old (targets), Similar (rhymes or synonyms), or New (different items). We first examined the Old versus New sensitivity and compared them in all experiments for every list length to determine whether these judgments were influenced by the presence of the third, Similar condition and by the instructions as to how it was to be treated.

The presence of highly similar distractor items means that the discrimination of targets from different items theoretically could be influenced by both verbatim and gist memory processes (e.g., Brainerd et al., 1999, 2014), which are central to fuzzy-trace theory (Reyna & Brainerd, 1995). Accordingly, verbatim memory describes memory for surface-level details and specific information about the encoding context, while gist memory pertains to higher level or meaning-based representations which provide a less fine-grained discrimination of targets from related distractors. The ability to discriminate old and similar items depends, at least in part, on the ability to retrieve verbatim memory representations of old items, as gist memory retrieval alone would be insufficient given the high amount of representational overlap between these items (Brainerd et al., 2014). To derive an approximate index of verbatim memory retrieval, we measured participants' ability to discriminate Old and Similar items in the reject-similar conditions (Experiment 1 and 3) using SDT models. This discrimination metric has been shown to correspond to verbatim memory parameters of a multinomial processing tree model (Greene & Naveh-Benjamin, 2020) and requires the retrieval of a sufficient number of specific details to

correctly discriminate between target items and similar foils (e.g., Loiotile & Courtney, 2015). These models address whether participants were able to remember enough specific details about originally-studied items to discriminate between target items and rhymes or synonyms in the conditions in which only targets were to be judged "same."

Last, we measured whether the reinstatement of both phonological and semantic cues at retrieval (i.e., in the case of old items) resulted in better discrimination from novel items than the reinstatement of just one cue (as in rhymes, with just phonological cues; or synonyms, with just semantic cues), as in the accept-similar conditions of Experiments 2 and 4.

All models were set up as hierarchical Bayesian probit-regression models, which are equivalent to the equal variance SDT model (DeCarlo, 1998; Rouder & Lu, 2005). In all models, the outcome was whether the response on a given trial was "same" (coded as 1) or "different" (coded as 0). The response on trial i for subject j was assumed to be Bernoulli distributed, with probability p_{ij} that $y_{ij} = 1$. We used a generalized linear model with a probit link function to map the probabilities to the real line, such that:

$$p_{ij} = \Phi(\beta_{0j} + \beta_{1j}*\text{Probe}_{ij})$$

where Φ is the cumulative normal density function. The predictor Probe codes for whether the test probe is an old item or related distractor for Experiments 1 and 3 (coded as 1 = Old, 0 = Rhyme/Synonym). In Experiments 1 and 3, we also compared Old items with New (i.e., different) items (coded as 0) in separate models for comparison purposes of Old/New discrimination in Experiments 2 and 4. For Experiments 2 and 4, we computed two SDT models, one in which the effect of Probe was coded as 1 = Old, 0 = New, and the other for which 1 = Rhyme/Synonym, 0 = New, to compare discrimination when both cues were available (old items) to when only one cue was available (rhymes or synonyms). The intercept in each model corresponds to the standardized false alarm rate (zFA), and inverting the sign of the intercept

yields SDT's response bias parameter, that is $c_j = - \beta_{0j}$. The slope of the model corresponds to the increase in the probability of "old" responses for probes coded as 1 relative to probes coded as 0, and is thus equal to SDT's estimate of discrimination, *d'* (DeCarlo, 2010; Rouder & Lu, 2005).

For each experiment, we computed the SDT model separately at each set size. Models were estimated using the brms package for R (Bürkner, 2017, R Core Team, 2020) with weakly informative normal priors specified. For the effect of Probe (corresponding to *d'*), we used a Normal(0.5, 1) prior, centering our prior belief of the value of *d'* at 0.5, a reasonably small and non-informative estimate of *d'*, sufficient to regularize the posterior distribution to avoid incalculable estimates of *d'* which can sometimes arise when performance is perfect (e.g., Stanislaw & Todorov, 1999). At the group-level effects, we used an LKJ(2) prior on the correlation between the intercept and slope. Models were estimated from four independent MCMC chains for 1000 iterations each (the first 500 of which were warm-up samples), with convergence monitored by the scale-reduction factor statistic $\hat{R}$. All $\hat{R} < 1.03$, indicating the chains converged.

Cross-Experiment Results and Discussion

Estimates of *d'* are reported in Table 5, and Figure 6 shows how *d'* changed for each contrast as a function of set size. First, examination of the left panel of Figure 6 reveals that the ability to discriminate Old items from New, dissimilar items was relatively constant across experiments, and this discrimination declined in a monotonic fashion at higher set sizes, being about constant for set sizes 1, 2, and 3 before dropping at each successive set size for 4, 6, and 8 item arrays (see Table 5 for exact means of *d'*). Thus, Old/New discrimination was unaffected by instructions to accept only target items (as in Experiments 1 and 3; i.e., the reject-similar conditions) or to accept both target items and similar lures (as in Experiments 2 and 4; i.e., the

accept-similar conditions). However, this discrimination was affected by our set size manipulation.

Regarding the ability to discriminate old items from similar distractors (rhymes or synonyms), *d'* was relatively high (see middle panel of Figure 6), indicating good discrimination, and there were only very modest changes with increasing set size, which mostly appeared restricted to Old/Rhyme discrimination. In fact, at lower set sizes, Old/Similar item discrimination (for both rhymes and synonyms) was essentially identical to Old/New discrimination. Interestingly, at the highest set sizes (6 and 8), Old/Synonym discrimination in the reject-similar condition of Experiment 3 was *higher* than Old/New discrimination in this same experiment (see Table 5 for means). This is evident in Figure 6 by comparing the HDIs of Old/Synonym in the middle panel with those of Old/New for Experiment 3 in the leftmost panel. It supports the notion of a recall-to-reject process.

Also, as evidenced by the overlapping HDIs in Figure 6, middle panel, the ability to discriminate old items from similar distractors was generally equal for both rhymes and synonyms, suggesting that participants were about equally good at discriminating old items from distractors that were either phonologically or semantically related. However, at set size 8, Old/Rhyme discrimination was worse than Old/Synonym discrimination (with non-overlapping HDIs), indicating that at the largest set size, participants were worse at remembering specific enough information to discriminate old items from phonologically-similar items than to discriminate old items from semantically-similar items, consistent with prior research on the predominance of phonological cues in working memory (e.g., Baddeley, 1966a, 1966b; Craik, 2020).

Finally, the rightmost panel of Figure 6 shows how *d'* changed across set sizes for the contrast of similar lures (rhymes in Experiment 2, and synonyms in Experiment 4) from new

items in the accept-similar conditions. Notably, *d'* was generally lower for the similar/new contrast than for the old/new contrasts in these experiments, and for the old/similar contrasts, and was especially low at the highest set sizes. Thus, when a test probe contained *both* accurate phonological and accurate semantic cues (as was the case for old items), discrimination from different items in the accept-similar condition was much better than when a test probe contained a relevant cue (a rhyme or a synonym of a target item) along with a mismatch in the other feature. Also of note, Rhyme/New discrimination was superior to Synonym/New discrimination at all set sizes, as evident by the non-overlapping HDIs in the right panel of Figure 6 for the two types of contrasts. The fact that it was somewhat harder to carry out the accept-synonym judgment, despite conflicting phonological and orthographic information, is in keeping with the notion that it is difficult to ignore the phonological information in a working memory task (e.g., Baddeley, 1966a, 1966b).

In sum, the signal detection analyses converge with the accuracy and RT analyses to show, now without the contaminating effects of bias, how sensitive participants are to phonological and semantic information similar to list item information in reject-similar and accept-similar conditions. One can examine the fate of information with increasing set size to get an indication of whether capacity limits are related to other processes. Most of the functions in Figure 6 show a fairly similar (though not identical) decline in sensitivity across set sizes. The one exception is that the reject-synonym judgment (middle panel, Experiment 3), which seems to benefit from a recall-to-reject process, also seems protected by that process from list length effects compared to other trial types. The recall-to-reject process may allow a use of long-term memory representations that does not seem to occur in other conditions. The error bars are wider than in other conditions, suggesting that the benefit of a recall-to-reject process was present more in some participants than in others.

General Discussion

Here, we examined the reject-similar and accept-similar use of phonological and semantic information using a probe recognition task based on the seminal work of Sternberg (1975). Participants had to memorize a sequence of 1, 2, 3, 4, 6, or 8 words and had to indicate whether a test probe was part of the list or not. Lures sometimes rhymed with a list item (Experiment 1) or were similar in meaning to the list item (Experiment 3). In Experiment 2, participants had to press one key if the test probe was either the same or rhymed with a study word, and a different key otherwise; Experiment 4 was comparable, but with semantic as opposed to phonological similarities to be judged. Therefore, Experiments 1 and 3 investigated the reject-similar use of phonological and semantic codes, respectively, and Experiments 2 and 4 investigated the accept-similar use of phonological and semantic codes.

Use of Phonological/Orthographic versus Semantic Codes in the Present Experiments

The main findings of interest for phonological information were that the reject-similar and the accept-similar use of phonological information were disadvantageous for both accuracy and speed when compared with the control probe with the same desired response (Experiment 1: *different* probes; Experiment 2: *same* probes). For semantic information, in contrast, the reject-similar use of semantic information was inconsequential for smaller set sizes and advantageous for larger memory set sizes (e.g., 6 and larger), in both accuracy and speed. The accept-similar use of semantic information was once again disadvantageous for both accuracy and speed. These outcomes, in each case, are in comparison to the probes with the same desired response (Experiment 3: *different* probes; Experiment 4: *same* probes).

The accept-similar use of phonological information (Experiment 2) and of semantic information (Experiment 4) was disadvantageous relative to the condition with the same desired response (*same* probes). In other words, it was more difficult for the participants to call a probe

"same" when only one dimension (Experiment 2: phonological; Experiment 4: semantic) was available relative to *same* probes, in which both phonological and semantic information shared with the target were available. These results suggest that phonological and semantic information are used together to make deliberate similarity judgments. These results support the implication that phonological and semantic codes are used in working memory.

However, for the reject-similar use of phonological (Experiment 1) and of semantic information (Experiment 3) the patterns of results differ. In particular, when compared to the probe type with the same desired response (*different* probes), the reject-similar use of phonological information was disadvantageous, but the reject-similar use of semantic information was advantageous for larger set size. It was more difficult for the participants to call a probe "different" when the probe shared some phonological information with the target item, relative to the *different* probes in which no information was shared with the target. However, it was easier for the participants to call a probe "different" when the probe shared some semantic information with the target item, relative to *different* probes in which no information was shared with the target. Importantly, the latter advantage was only observed for larger memory set sizes (6 and 8 items). These results suggest that phonological information does not serve as an efficient reminder of the target item in the list, but semantic information can serve as an efficient reminder.

To take the example in the introduction, imagine that the word *belief* is presented in the list and then forgotten. Then a probe is presented and is phonologically similar to the target item (e.g., *relief*). Based on our results, this phonologically similar probe does not serve as an efficient recall cue. Indeed, rather than rejecting the probe the participant often misidentifies the phonologically similar probe as the target item on the basis of the similarity. However, based on our results, when the probe is semantically related (e.g., *conviction*), it does serve as an efficient

retrieval cue for *belief*. Combined with a realization that words in a list did not contain synonyms, it allows rejection of *conviction* as absent from the list. The results for reject-similar use information support the *recall-to-reject* process only for semantic information "in which mismatching information that is retrieved from memory is used to reject test foils that are similar to studied items" (Rotello et al., 2000, p. 67).

Overall, the results support some notion put forward by Shulman (1971) and Baddeley (1972), but in a new manner, by distinguishing between reject-similar and accept-similar use of phonological and semantic information with a common paradigm. Baddeley suggested that the use of semantic codes could occur, but that these were useful only when there were retrieval rules stored in long-term memory that could be applied. In this case, the advantage for the reject-similar use of semantic codes observed with larger memory set size can only occur if participants stored the rule that the words in a list were semantically diverse. Somewhat consistent with the conclusions of Shulman (1971), when the task encouraged the use of semantic codes, they could be used especially in a reject-similar manner (Experiment 3). When the task required the accept-similar use of semantic codes, however, this use was still detrimental, for *synonym* probes, compared to *same* probes in which both semantic and phonological codes were available for use together (Experiment 4).

One unanswered question is why a recall-to-reject advantage was only observed in the reject-synonym case (Figure 6, middle panel, Experiment 3) and not in the reject-rhyme, Experiment 1 case. The recall-to-reject advantage presumably occurs through retrieval from long-term memory, which can play a role in immediate memory tasks (Unsworth & Engle, 2007). Therefore, if semantic information serves as a better long-term memory retrieval cue, this could explain the difference. It is also likely that there is much more phonological overlap between

materials compared to the amount of semantic overlap, although there may not presently be a completely clear metric for comparing the two. We now suggest avenues for further research.

Phonological versus Orthographic Codes

Note that we have not distinguished between phonological and orthographic codes. Although it is clear that phonological codes prevail (Conrad, 1964), there is also an important visual or orthographic similarity effect (e.g., Guitard & Cowan, 2020; Lin et al., 2015; Logie et al., 2000). This could be examined in future extensions of the present work if enough rhyming word pairs can be found that substantially differ orthographically (e.g., *rhyme* vs. *climb*).

Familiarity and Recollection Processes

One way to think of the reject-similar situation is in terms of familiarity and recollection (e.g., Jacoby, 1991). The similar feature is familiar from the list so it takes recollection to indicate that the familiarity is not to be trusted. Unsworth and Brewer (2009) showed that although familiarity and recollection are separate, both can be involved in recognition tasks. However, Matzen et al. (2011) examined recognition using probes that included lures with phonological features similar to list items (e.g., tailgate when the participant saw tailspin and floodgate) and other lures with semantic features (e.g., bunny when the participant saw rabbit). For words that participants thought they had studied, they were able to respond “remember” or “familiar” and for words that participants did not think they had studied, “unfamiliar” or “different”. Results varied by lag, but at the shortest lag, most relevant to our immediate-recognition procedure, phonologically similar lures produced more incorrect remember responses than semantic lures, the same proportion of incorrect familiar responses, and fewer correct rejections than semantic lures. This result suggests that participants did not make more phonological errors by familiarity alone, but by mistaking the phonologically similar items in a faulty recollection process.

In the similar-accept situation, the dissimilarity from any targets in the non-similar feature can harmfully lead to a different answer than correct recollection. For example, in the accept-rhyme situation, the rhyming lure is semantically different from any list item and this difference has to be overlooked to make the correct response. The results suggest that familiarity with both phonological/orthographic and semantic information had detrimental effects. In the reject-similar situation, the poorer performance for rhymes than for control lures shows this. In the accept-similar situation, the poorer performance for both rhymes and synonyms compared to control lures shows this. Given that performance is poorer for the accept-synonym situation, it appears that conflicting phonological information is most difficult to overcome. More work would be helpful to distinguish more clearly between the use of familiarity and recollection and its involvement in the unexpected recall-to-reject process that governed reject-synonym responses.

Semantics and Phonology During Maintenance versus Retrieval

Although the typical assumption in a working memory task is that successful responses are based on a process of encoding, continual maintenance, and retrieval of the information from working memory, there is research suggesting that, sometimes, what actually happens is encoding, inactivation, and later reactivation of the information. That later reactivation may refer to retrieval from long-term memory into an activated state and/or re-entry of activated information into the focus of attention to allow a deliberate response (for reviews see Cowan, 2017, 2019). It will take further work to determine which processes must take place for the present phenomena to occur. One way to disentangle the possibilities comes from a study by Shivde and Anderson (2011). They had participants retain a word for subsequent comparison with another word; in different experiments, the comparison was based on semantic or phonological similarity between the words. During the retention interval, there were multiple trials of a lexical decision task, which included probes that were semantically or phonologically

similar to the word that was supposed to be in memory. Effects of the to-be-retained word on the lexical decision latency provided an indication that the word was indeed retained in an active form capable of causing interference. In principle, a check like that from a lexical decision task could be interpolated between the stimulus set and probe of the present task to learn more about the state of maintenance of the memory set items at the time that the probe item is presented.

Implications of the Present Findings for Theories of Working Memory

The present results can be assessed with respect to two areas of research and has theoretical implications for both. The first is cognitive behavioral research, and the second is brain research, and especially brain imaging, directed at the neural representation of functions underlying the cognitive models.

Cognitive Behavioral Implications

Shivde and Anderson (2011) noted that there was very little evidence for the use of semantic codes in working memory, and they provided some evidence. The present work goes further in not only establishing another method to index phonological and semantic codes in working memory, but also documenting important differences in how these codes are used. There is an ability to retain phonological information (Experiments 1 & 2) and semantic information (Experiments 3 & 4) about word lists to be compared to a probe item using these codes. It makes sense that it was difficult to consider a probe item "different" from a list item while in other ways it is similar (Figure 6, right-hand panel) compared to considering a probe item "same" as a list item in a critical manner even though it is not identical (Figure 6, middle panel). Thus, more difficulty occurs when the probe item must be classified in a manner that contrasts with the critical phonological or semantic features (reject-similar, Experiments 1 & 3), with less difficulty when the probe item is to be classified as the same as the target sharing those features (accept-similar, Experiments 2 & 4).

Although the use of phonological and semantic information about a particular item decreases as a function of the set size, it does so less for semantic information that can be used in a recall-to-reject process in Experiment 3. The reason why this process comes into play only for semantic information could have to do with the arrangement of stimuli. There are only a limited number of phonemes in the language that can make up all of the words in the stimulus set. They do, however, most likely include a larger set of semantic features (see Appendix A). If this account is correct then, in a subsequent experiment, drawing stimuli from a more crowded semantic space could remove the recall-to-reject process.

The results are consistent with theories that allow both phonological and semantic maintenance during working-memory tasks. This would apply to the most recent multicomponent model that includes not only a phonological buffer and a visuo-spatial buffer, but also an episodic buffer capable of holding semantic information (Baddeley, 2000). It would also apply to the embedded-processes model (Cowan, 1988, 2019) that includes activated phonological and semantic features. The latter approach makes heavier use of general learning principles, inasmuch as there must be rapid learning of information so that the order of items is preserved when it leaves the focus of attention, as was emphasized by Cowan (2019). From that perspective, the evidence for a recall-to-reject process that previously was invoked primarily within a long-term memory type of paradigm (Rotello et al., 2000) is favorable to the model. Similarly, Cowan and Hardman (in press) presented lists of digits for recall in which there can be multiple repetitions of a digit in a list, and found that another long-term memory principle, fan effects, applies to this short-term recall situation. These long-term learning factors do not contradict a multicomponent approach but seem more directly relevant to the embedded processes approach.

Brain Basis of Working Memory

Future progress could also come from a neuroscientific investigation of the distinction between the use of phonological and semantic information in reject-similar and accept-similar situations in immediate probe recognition. According to an embedded-processes view, in all cases, the hippocampal system, guided by deliberate search using the prefrontal cortex (Nee & Jonides, 2011), might produce the retrieval of recently presented information from activated long-term memory into the focus of attention for episodic information about the list (Cowan, 1988, 2019). The recall-to-reject windfall would be explained as resulting from semantic priming that is more effective and specific than phonological priming. Alternatively, according to a multicomponent model of working memory, phonological information would be automatically activated in the buffer, whereas central executive processes should be more highly involved in semantic search to produce a representation in the episodic buffer (Baddeley, 2000). The differential use of the prefrontal system for phonological versus semantic information then might underlie the recall-to-reject process limited to semantics. The models thus appear to differ in the predicted levels of involvement of frontal processes for phonological retrieval in our tasks.

Conclusion

By distinguishing between reject-similar and accept-similar use of phonological and semantic information with a common paradigm we have found theoretically distinct patterns of results. We have found evidence that the reject-similar use of semantic codes is advantageous to making dissimilarity judgments, as it provides a distinctive cue to *recall-to-reject* the probe. In contrast, phonological codes are apparently not sufficiently distinct for this process and lead to confusion of the probe with the target item with which it rhymes. For the accept-similar use of phonological and semantic information, we have found evidence for a disadvantage to make similarity judgments. Here we have addressed the relative paucity of work on the way semantic information is used in working memory, and we have found that its use differs in both the

strength of its use (weaker than phonological information for accept-similar purposes) and manner of its use (stronger than phonological information for reject-similar purposes). The results might well be different for retention of semantically coherent text. We encourage researchers to also focus on various codes (e.g., phonological, semantic, visual, and orthographic codes) as it will improve our understanding of working memory.

References

Andrews, M. and Baguley, T. (2013), Prior approval: The growth of Bayesian methods in psychology. *Br J Math Stat Psychol, 66*, 1-7. doi:10.1111/bmsp.12004

Shivde, G., & Anderson, M. C. (2011). On the existence of semantic working memory: Evidence for direct semantic maintenance. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(6), 1342–1370. https://doi.org/10.1037/a0024832

Atkins, D. C., Baldwin, S. A., Zheng, C., Gallop, R. J., Neighbors, C. (2013). A tutorial on count regression and zero-altered count models for longitudinal substance use data. *Psychology of Addictive Behaviors*, *27*, 166–77. doi: 10.1037/a0029508

Baddeley, A. D. (1966a). Short-term memory for word sequences as a function of acoustic, semantic and formal similarity. *Quarterly Journal of Experimental Psychology, 18,* 362–365. doi: 10.1080/14640746608400055

Baddeley, A.D. (1966b). The influence of acoustic and semantic similarity on long term memory for word sequences. *Quarterly Journal of Experimental Psychology, 18*, 302-309. DOI: 10.1080/14640746608400047

Baddeley, A.D. (1972). Retrieval rules and semantic coding in short-term memory. *Psychological Bulletin*, 78, 379-385. doi: 10.1037/h0033477

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences, 4*, 417 423. DOI: 10.1016/S1364-6613(00)01538-2

Begg, I. (1971). Recognition memory for sentence meaning and wording. *Journal of Verbal Learning & Verbal Behavior, 10*, 176–181. Doi: 10.1016/S0022-5371(71)80010-5

Brainerd, C. J., Gomes, C. F. A., & Moran, R. (2014). The two recollections. *Psychological Review, 121*(4), 563-599. https://doi.org/10.1037/a0037668

Brainerd, C. J., Reyna, V. F., & Mojardin, A. H. (1999). Conjoint recognition. *Psychological Review, 106*, 160-179. https://doi.org/10.1037/0033-295X.106.1.160

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software, 80*(1), 1-1. doi:10.18637/jss.v080.i01

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, *24*, 127 - 135. doi: 10.1016/j.tree.2008.10.008

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., . . . Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software, Articles*, *76*(1), 1-1. doi:10.18637/jss.v076.i01

Chubala, C. M., Neath, I., & Surprenant, A. M. (2019). A comparison of immediate serial recall and immediate serial recognition. *Canadian Journal of Experimental Psychology, 73*, 5–27. doi: 10.1037/cep0000158

Conrad, R. (1964). Acoustic confusions in immediate memory. *British Journal of Psychology, 55*, 75 –84. doi: 10.1111/j. 2044-8295.1964.tb00899.x

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information processing system. *Psychological Bulletin*, *104*, 163-191.

Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review*, *24*, 1158–1170. DOI: 10.3758/s13423-016-1191-6

Cowan, N. (2019) Short-term memory based on activated long-term memory: A review in response to Norris (2017). *Psychological Bulletin, 145*, 822-847.

Cowan, N., & Hardman, K.O. (in press). Immediate recall of grouped serial numbers with or without multiple item repetitions. *Memory*.

Craik, F. I. (2020). Remembering: An activity of mind and brain. *Annual review of psychology, 71*, 1-24. DOI: 10.1146/annurev-psych-010419- 051027

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671-684. DOI: 10.1016/S0022-5371(72)80001-X

Criss, A. H. (2009). The distribution of subjective memory strength: Foils and response bias. *Cognitive Psychology, 59*, 297–319. DOI: 10.1016/j.cogpsych.2009.07.003

Crowder, R.G. (1979). Similarity and order in memory. In G.H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory: Vol. 13* (p. 319-353). New York: Academic Press.

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods, 3*(2), 186-205. https://doi.org/10.1037/1082-989X.3.2.186

DeCarlo, L. T. (2010). On the statistical and theoretical basis of signal detection theory and extensions: Unequal variance, random coefficient, and mixture models. *Journal of Mathematical Psychology, 54*, 304-313. https://doi.org/10.1016/j.jmp.2010.01.001

Endress, A. D., & Potter, M. C. (2014). Large capacity temporary visual memory. *Journal of Experimental Psychology: General, 143*, 548-566. DOI: 10.1037/a0033934

Etz A. & Vandekerckhove, J. (2016) A Bayesian perspective on the reproducibility project: Psychology. *PLOS ONE*,*11*(2): e0149794. doi: 10.1371/journal.pone.0149794

Flegal, K. E., Atkins, A. S., & Reuter-Lorenz, P. A. (2010). False memories seconds later: the rapid and compelling onset of illusory recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 1331–1338. https://doi.org/10.1037/a0019903

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian data analysis* (3rd ed.). New York, NY: Taylor & Francis.

Gilchrist, A.L., Cowan, N., & Naveh-Benjamin, M. (2008). Working memory capacity for spoken sentences decreases with adult ageing: Recall of fewer, but not smaller chunks in older adults. *Memory, 16*, 773-787. doi: 10.1080/09658210802261124.

Gilchrist, A.L., Cowan, N., & Naveh-Benjamin, M. (2009). Investigating the childhood development of working memory using sentences: New evidence for the growth of chunk capacity. *Journal of Experimental Child Psychology, 104*, 252-265. doi:10.1016/j.jecp.2009.05.006

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Greene, N. R., & Naveh-Benjamin, M. (2020). A specificity principle of memory: Evidence from aging and associative memory. *Psychological Science, 31*(3), 316-331. https://doi.org/10.1177/0956797620901760

Guérard, K., & Saint-Aubin, J. (2012). Assessing the effect of lexical variables in backward recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 312–324. doi: 10.1037/a0025481

Guitard, D., & Cowan, N. (2020). Do we use visual codes when information is not presented visually? *Memory & Cognition, 48*(8), 1522–1536. https://doi.org/10.3758/s13421-020-01054-0

Guitard, D., Saint-Aubin, J., & Cowan, N. (2020, online ahead of print). Asymmetrical interference between item and order information in short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*

Hilbe, J. M., de Souza, R. S., & Ishida, E. E. O. (2017). *Bayesian models for astrophysical data: Using R, Jags, Python, and Stan*. Cambridge, UK: Cambridge University Press. doi: 10.1017/CBO9781316459515

Ishiguro, S., & Saito, S. (2020). The detrimental effect of semantic similarity in short-term memory tasks: A meta-regression approach. *Psychonomic Bulletin & Review*. Advance online publication. doi: 10.3758/s13423-020-01815-7

Jacoby, L.L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory & Language, 30*, 513-541. DOI: 10.1016/0749-596X(91)90025-F

Kruschke, J. K. (2015). Doing Bayesian data analysis: a tutorial with R, JAGS, and Stan. Retrieved from http://www.sciencedirect.com/science/book/9780124058880

Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review, 25*(1), 155–177. https://doi.org/10.3758/s13423-017-1272-1

Lin, Y.-C., Chen, H.-Y., Lai, Y. C., & Wu, D. H. (2015). Phonological similarity and orthographic similarity affect probed serial recall of Chinese characters. *Memory & Cognition, 43*, 538 –554. doi.org: 10.3758/s13421-014-0495-x

Logie, R.H., Della Sala, S., & Wynn, V., & Baddeley, A.D. (2000). Visual similarity effects in immediate verbal serial recall. *Quarterly Journal of Experimental Psychology, 53A*, 626-646.

Loiotile, R. E., & Courtney, S. M. (2015). A signal detection theory analysis of behavioral pattern separation paradigms. *Learning & Memory, 22*(8), https://doi.org/10.1101/lm.038141.115

Matzen, L. E., Taylor, E. G., & Benjamin, A. S. (2011). Contributions of familiarity and recollection rejection to recognition: Evidence from the time course of false recognition for semantic and conjunction lures. *Memory, 19*, 1-16. doi: 10.1080/09658211.2010.530271

McElree, B. (1996). Accessing short-term memory with semantic and phonological information: A time-course analysis *Memory & Cognition, 24*, 173–187. doi: 10.3758/BF03200879

McElreath, R. (2016). Statistical Rethinking: A Bayesian Course with Examples in R and Stan. CRC Press.

Murdock, B.B., Jr. (1976). Item and order information in short-term serial memory. *Journal of Experimental Psychology: General*, *105*, 191-216. doi: 10.1037/0096-3445.105.2.191

Neale, K., & Tehan, G. (2007). Age and redintegration in immediate memory and their relationship to task difficulty. *Memory & Cognition*, *35*, 1940–1953. doi: 10.3758/BF03192927

Nee, D.E. & Jonides, J. (2011). Dissociable contributions of prefrontal cortex and the hippocampus to short-term memory: Evidence for a 3-state model of memory, *Neuroimage. 54*, 1540–1548. doi:10.1016/j.neuroimage.2010.09.002.

Nimmo, L. M., & Roodenrys, S. (2004). Investigating the phonological similarity effect: Syllable structure and the position of common phonemes. *Journal of Memory and Language*, *50*, 245–258. doi: 10.1016/j.jml.2003.11.001

Poirier, M., & Saint-Aubin, J. (1995). Memory for related and unrelated words: Further evidence on the influence of semantic factors in immediate serial recall. *Quarterly Journal of Experimental Psychology*,*43A*, 384-404. doi: 10.1080/14640749508401396

Poirier, M., & Saint-Aubin, J. (1996). Immediate serial recall, word frequency, item identity and item position. *Canadian Journal of Experimental Psychology*, *50*, 408–412. doi: 10.1037/1196-1961.50.4.408

Potter, M. (1993). Very short-term conceptual memory. *Memory & cognition. 21*. 156-61. doi: 10.3758/BF03202727.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

R Core Team. (2020). R: A language and environment for statistical computing (Version 3.6.3) [Computer software]. Retrieved from http://www.R-project.org/

Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences, 7*, 1-75. https://doi.org/10.1016/1041-6080(95)90031-4

Rotello, C.M., Macmillan, N.A., & Van Tassel, G. (2000). Recall-to-reject in recognition: Evidence from ROC Curves. *Journal of Memory and Language, 43*, 67-88. doi: 10.1006/jmla.1999.2701

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review, 12*(4), 573-604. https://doi.org/10.3758/BF03196750

Sachs, J.S. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Perception & Psychophysics, 2*, 437-442. DOI: 10.3758/BF03208784

Saint-Aubin, J., Hilchey, M. D., Mishra, R., Singh, N., Savoie, D., Guitard, D., & Klein, R. M. (2018). Does the relation between the control of attention and second language proficiency generalize from India to Canada? *Canadian Journal of Experimental Psychology*, *72*, 208–218. doi: 10.1037/cep0000151

Saint-Aubin, J., Ouellette, D., & Poirier, M. (2005). Semantic similarity and immediate serial recall: Is there an effect on all trials? *Psychonomic Bulletin & Review*, *12*, 171–177. doi: 10.3758/BF03196364

Saint-Aubin, J., & Poirier, M. (1999). Semantic similarity and immediate serial recall: Is there a detrimental effect on order information? *The Quarterly Journal of Experimental*

Psychology A: Human Experimental Psychology, *52A*(2), 367–394. doi: 10.1080/027249899391115

Shivde, G., & Anderson, M. C. (2011). On the existence of semantic working memory: Evidence for direct semantic maintenance. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(6), 1342–1370. https://doi.org/10.1037/a0024832

Shulman, H. G. (1970). Encoding and retention of semantic and phonemic information in short-term memory. *Journal of Verbal Learning and Verbal Behavior, 9,* 499-508. doi: 10.1016/S0022-5371(70)80093-7

Shulman, H.G. (1971). Similarity effects in short-term memory. *Psychological Bulletin, 75*, 399-415. doi: 10.1037/h0031257

Snodgrass, J.G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*, 34–50. doi: 10.1037//0096-3445.117.1.34

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavioral Research Methods, Instruments, & Computers, 31,* 137-149. https://doi.org/10.3758/BF03207704

Tehan, G. (2010). Associative relatedness enhances recall and produces false memories in immediate serial recall. *Canadian Journal of Experimental Psychology*, *64*, 266–272. doi: 10.1037/a0021375

Tse, C.-S. (2009). The role of associative strength in the semantic relatedness effect on immediate serial recall. *Memory, 17*, 874–891. doi: 10.1080/09658210903376250

Tse, C.-S., Li, Y., & Altarriba, J. (2011). The effect of semantic relatedness on immediate serial recall and serial recognition. *The Quarterly Journal of Experimental Psychology*, *64*(12), 2425–2437. doi: 10.1080/17470218.2011.604787

Unsworth, N., & Brewer, G. A. (2009). Examining the relationships among item recognition, source recognition, and recall from an individual differences perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 1578–1585. https://doi.org/10.1037/a0017255

Unsworth, N., & Engle, R.W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review, 114*, 104-132. doi: 10.1037/0033-295X.114.1.104.

van der Linden, S., & Chryst, B. (2017). No need for Bayes factors: A fully Bayesian evidence synthesis. *Frontiers in Applied Mathematics and Statistics, 3*. doi: 10.3389/fams.2017.00012

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

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., … Morey, R. D. (2018). Bayesian inference for psychology Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25*, 35–57. doi: 10.3758/s13423-017 1343-3

Wickelgren, W. A. (1965). Short-term memory for phonemically similar lists. *American Journal of Psychology, 78*, 567-74. doi: 10.2307/1420917

Yao, Y., Vehtari, A., Simpson, D., Gelman, A. (2018). Using stacking to average Bayesian predictive distributions. *Bayesian Analysis, 13*, 917-1007. doi:10.1214/17-BA1091 (URL: https://doi.org/10.1214/17-BA1091).

Appendix A

Material for experiments 1 and 2 (item and rhyme) and material for experiments 3 and 4 (item and synonym).

Item	Synonym	Rhyme	Item	Synonym	Rhyme
ALONE	SINGLE	PHONE	INTERIOR	INSIDE	INFERIOR
APPEAL	PRAYER	SEAL	ISSUE	EDITION	TISSUE
BECKON	SUMMON	RECKON	JINX	HEX	LYNX
BELIEF	CONVICTION	RELIEF	LANK	GAUNT	SANK
BIBLE	SCRIPTURE	LIABLE	LIMB	BRANCH	BRIM
BOG	SWAMP	FOG	METAL	STEEL	PETAL
BREAK	SMASH	RAKE	NEW	MODERN	BLUE
BUNNY	RABBIT	FUNNY	NOOK	CORNER	BOOK
CARESS	FONDLE	REGRESS	PAIL	BUCKET	FAIL
CASTLE	PALACE	HASSLE	PARASOL	UMBRELLA	AEROSOL
CELLAR	BASEMENT	TELLER	PISTOL	REVOLVER	DISTAL
CHURCH	TEMPLE	LURCH	PLEDGE	PROMISE	HEDGE
CLAW	TALON	SAW	PORT	HARBOR	SORT
CLEAR	FREE	BEER	PROOF	EVIDENCE	SLEUTH
COACHED	INSTRUCTED	POACHED	ROCK	STONE	LOCK
COST	EXPENSE	LOST	SCOOP	LADLE	LOOP
DAWN	DAYBREAK	LAWN	SHREWD	ASTUTE	LEWD
DESCRIBED	RECOUNTED	PRESCRIBED	SIZE	MAGNITUDE	RISE
DOCTOR	PHYSICIAN	PROCTOR	SKIN	FLESH	BIN
DRAPES	CURTAIN	GRAPES	SOURCE	ORIGIN	FORCE
DUTY	OBLIGATION	BOOTY	SPEED	VELOCITY	SEED
ENTICE	ALLURE	CONCISE	STREET	ROAD	GREET
EXAM	TEST	CLAM	STYLE	FASHION	TILE
FARCE	MOCKERY	PARSE	TALE	STORY	WHALE
FEELING	SENSATION	REELING	TAVERN	SALOON	CAVERN
GEM	JEWEL	HEM	THOUGHT	IDEA	SOUGHT
GLOSS	SHEEN	TOSS	TOIL	DRUDGE	SOIL
GORED	PIERCED	BORED	TRADE	BARTER	CHARTER
GRIME	DIRT	CRIME	TROUBLE	DIFFICULTY	DOUBLE
HAMMER	MALLET	YAMMER	VILLAGE	TOWN	PILLAGE
HATCHET	TOMAHAWK	RATCHET	WOOD	LUMBER	COULD
HOME	HOUSE	COMB	YARD	FIELD	CARD

Table 1

Means estimates (and 95% highest density intervals (HDI)) of the posterior distributions of the regression coefficients for each selected Bayesian multilevel model used in Experiment 1

	Models		
Predicted variable	Reaction Time		Accuracy
Family distribution	Lognormal		Bernoulli
Link function (to map nonlinear data onto an unbounded line to allow regression)	Log		Logit (log odds)
Parameters	*M* [95% HDI]		*M* [95% HDI]
Group-level effects			
Random intercept (participants)	0.17 [0.13, 0.21]		0.82 [0.59, 1.09]
Random slope (memory set size)	0.02 [0.02, 0.03]		0.10 [0.07, 0.14]
Correlation (participants/memory set size)	-0.16 [-0.49, 0.17]		-0.75 [-0.92, -0.55]
Population level effects			
Intercept	6.11 [6.06, 6.17]		4.36 [3.95, 4.80]
Probe Type: Rhyme	**0.07 [0.05, 0.09]**		**-1.23 [-1.64, -0.86]**
Probe Type: Same	**0.03 [0.01, 0.05]**		**-1.51 [-1.87, -1.14]**
Memory Set Size	**0.04 [0.03, 0.05]**		**-0.24 [-0.30, -0.17]**
Probe Type: Rhyme x Memory Set Size	**-0.01 [-0.01, -0.00]**		**0.15 [0.08, 0.22]**
Probe Type: Same x Memory Set Size	0.00 [-0.00, 0.01]		0.02 [-0.05, 0.09]
Family Specific Parameters			
Sigma	0.26 [0.26, 0.26]		-

Note. bold = 95% HDI of the population-level effect excluded 0 (but intercept).

Table 2

Means estimates (and 95% highest density intervals (HDI)) of the posterior distributions of the regression coefficients for each selected Bayesian multilevel model used in Experiment 2

	Models		
Predicted variable	Reaction Time		Accuracy
Family distribution	Lognormal		Bernoulli
Link function (to map nonlinear data onto an unbounded line to allow regression)	Log		Logit (log odds)
Parameters	*M* [95% HDI]		*M* [95% HDI]
Group-level effects			
Random intercept (participants)	0.17 [0.12, 0.21]		0.89 [0.66, 1.11]
Random slope (memory set size)	0.03 [0.02, 0.04]		0.06 [0.04, 0.09]
Correlation (participants/memory set size)	0.20 [-0.12, 0.50]		-0.73 [-0.95, -0.50]
Population level effects			
Intercept	6.33 [6.28, 6.39]		2.67 [2.32, 2.99]
Probe Type: Rhyme	**-0.09 [-0.10, -0.08]**		**-0.94[-1.16, -0.74]**
Probe Type: Same	**-0.25 [-0.27, -0.24]**		**0.58 [0.33, 0.84]**
Memory Set Size	**0.04 [0.03, 0.05]**		**-0.20 [-0.24, -0.17]**
Probe Type: Rhyme x Memory Set Size	-		**0.05 [0.01, 0.09]**
Probe Type: Same x Memory Set Size	-		-0.04 [-0.09, 0.00]
Family Specific Parameters			
Sigma	0.29 [0.28, 0.29]		-

Note. bold = 95% HDI of the population-level effect excluded 0 (but intercept).

Table 3

Means estimates (and 95% highest density intervals (HDI)) of the posterior distributions of the regression coefficients for each selected Bayesian multilevel model used in Experiment 3

	Models		
Predicted variable	Reaction Time		Accuracy
Family distribution	Lognormal		Bernoulli
Link function (to map nonlinear data onto an unbounded line to allow regression)	Log		Logit (log odds)
Parameters	*M* [95% HDI]		*M* [95% HDI]
Group-level effects			
Random intercept (participants)	0.21 [0.16, 0.26]		1.08 [0.80, 1.39]
Random slope (memory set size)	0.02 [0.01, 0.02]		0.22 [0.16, 0.28]
Correlation (participants/memory set size)	-0.15 [-0.48, 0.21]		-0.72 [-0.88, -0.53]
Population level effects			
Intercept	6.08 [6.01, 6.15]		3.93 [3.52, 4.40]
Probe Type: Same	-0.00 [-0.02, 0.02]		**-1.19 [-1.48, -0.89]**
Probe Type: Synonym	-0.00 [-0.02, 0.02]		-0.15 [-0.52, 0.23]
Memory Set Size	**0.04 [0.03, 0.04]**		**-0.28 [-0.36, -0.20]**
Probe Type: Same x Memory Set Size	0.00 [-0.00, 0.01]		**0.09 [0.03, 0.14]**
Probe Type: Synonym x Memory Set Size	**-0.01 [-0.01, -0.00]**		**0.21 [0.14, 0.28]**
Family Specific Parameters			
Sigma	0.24 [0.24, 0.25]		-

Note. bold = 95% HDI of the population-level effect excluded 0 (but intercept).

Table 4

Means estimates (and 95% highest density intervals (HDI)) of the posterior distributions of the regression coefficients for each selected Bayesian multilevel model used in Experiment 4

	Models		
Predicted variable	Reaction Time		Accuracy
Family distribution	Lognormal		Bernoulli
Link function (to map nonlinear data onto an unbounded line to allow regression)	Log		Logit (log odds)
Parameters	*M* [95% HDI]		*M* [95% HDI]
Group-level effects			
Random intercept (participants)	0.18 [0.14, 0.23]		0.65 [0.48, 0.83]
Random slope (memory set size)	0.03 [0.02, 0.04]		0.05 [0.03, 0.08]
Correlation (participants/memory set size)	0.12 [-0.22, 0.45]		-0.72 [-0.96, -0.43]
Population level effects			
Intercept	6.41 [6.34, 6.46]		2.25 [1.99, 2.50]
Probe Type: Same	**-0.33 [-0.35, -0.32]**		**1.56 [1.27, 1.85]**
Probe Type: Synonym	**-0.03 [-0.05, -0.02]**		**-1.39 [-1.57, -1.19]**
Memory Set Size	**0.04 [0.03, 0.05]**		**-0.18 [-0.21, -0.14]**
Probe Type: Same x Memory Set Size	**-**		**-0.07 [-0.12, -0.02]**
Probe Type: Synonym x Memory Set Size	**-**		0.02 [-0.02, 0.06]
Family Specific Parameters			
Sigma	0.33 [0.32, 0.33]		-

Note. bold = 95% HDI of the population-level effect excluded 0 (but intercept).

Table 5

Population-level means [and 95% Highest Density Interval] of d' across set size from the probit regression models for Experiment 1 through 4.

Contrast	Set Size					
	1	2	3	4	6	8
Exp1 (Old vs Rhyme)	3.08 [2.81, 3.36]	3.15 [2.79, 3.51]	2.93 [2.64, 3.25]	2.82 [2.56, 3.09]	2.43 [2.15, 2.73]	2.08 [1.84, 2.31]
Exp1 (Old vs New)	3.32 [3.08, 3.54]	3.58 [3.19, 3.94]	3.38 [3.03, 3.77]	3.28 [2.94, 3.62]	2.60 [2.33, 2.90]	2.18 [1.96, 2.43]
Exp2 (Old vs New)	3.40 [2.79, 3.97]	3.09 [2.74, 3.44]	3.08 [2.56, 3.61]	2.58 [2.16, 2.99]	1.96 [1.60, 2.30]	1.55 [1.28, 1.80]
Exp2 (Rhyme vs New)	2.54 [2.19, 2.92]	2.20 [1.89, 2.48]	2.06 [1.74, 2.39]	1.67 [1.34, 1.98]	1.30 [1.06, 1.55]	1.11 [0.86, 1.34]
Exp3 (Old vs Synonym)	3.26 [2.91, 3.57]	3.49 [3.10, 3.89]	3.20 [2.85, 3.58]	3.41 [3.00, 3.90]	2.79 [2.22, 3.32]	2.76 [2.30, 3.26]
Exp3 (Old vs New)	3.30 [2.92, 3.66]	3.17 [2.85, 3.53]	3.16 [2.80, 3.56]	2.84 [2.59, 3.11]	2.25 [1.80, 2.67]	1.88 [1.53, 2.21]
Exp4 (Old vs New)	3.14 [2.74, 3.55]	3.46 [2.98, 3.97]	3.58 [3.06, 4.11]	2.68 [2.43, 2.94]	1.93 [1.61, 2.24]	1.77 [1.51, 2.05]
Exp4 (Synonym vs New)	1.90 [1.66, 2.16]	1.55 [1.25, 1.85]	1.54 [1.28, 1.82]	1.04 [0.79, 1.27]	0.59 [0.28, 0.88]	0.46 [0.25, 0.64]

Note. Exp = Experiment.

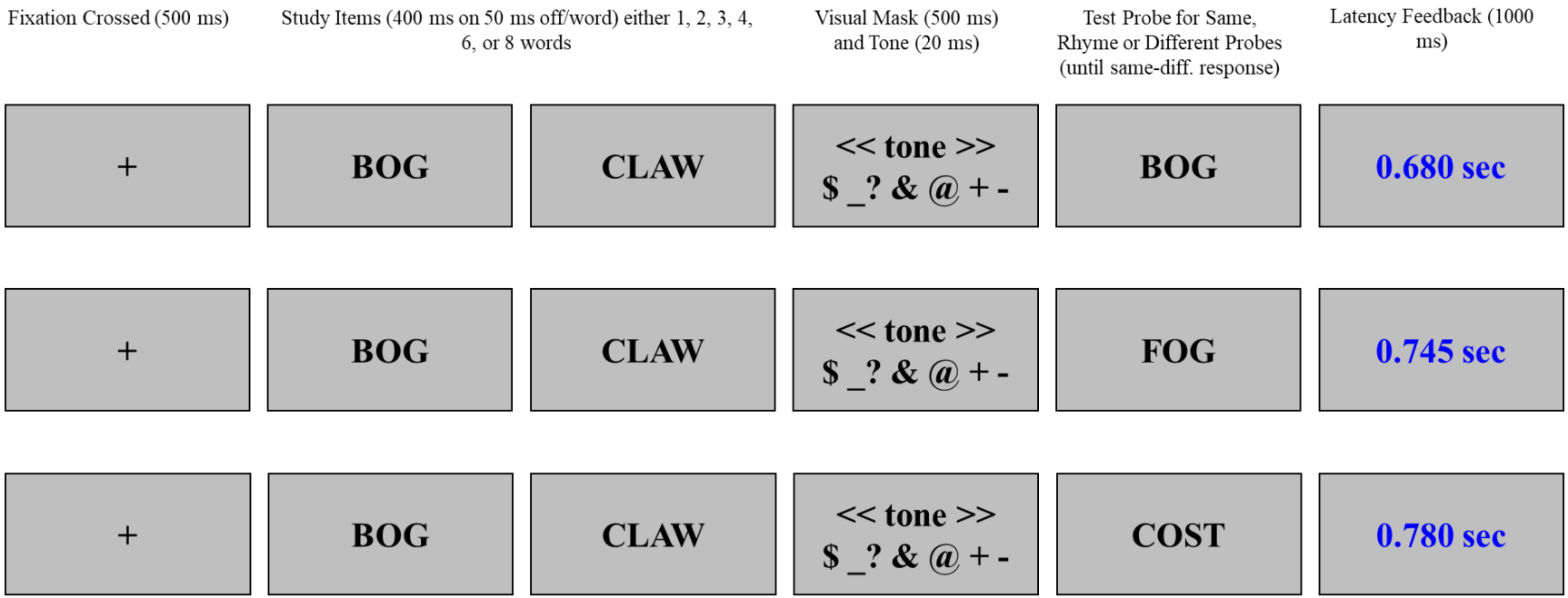

Figure 1. Illustration of a trial from left to right of the three different probe type conditions (same: top row, rhyme: middle row, different bottom row used in Experiment 1 and Experiment 2. In Experiment 1, rhyme and different probes were to be classified as "different;" in Experiment 2, rhyme and same probes were to be classified as "same or similar." Experiments 3 and 4 followed the same rules, but with semantic similarity instead of phonological similarity.

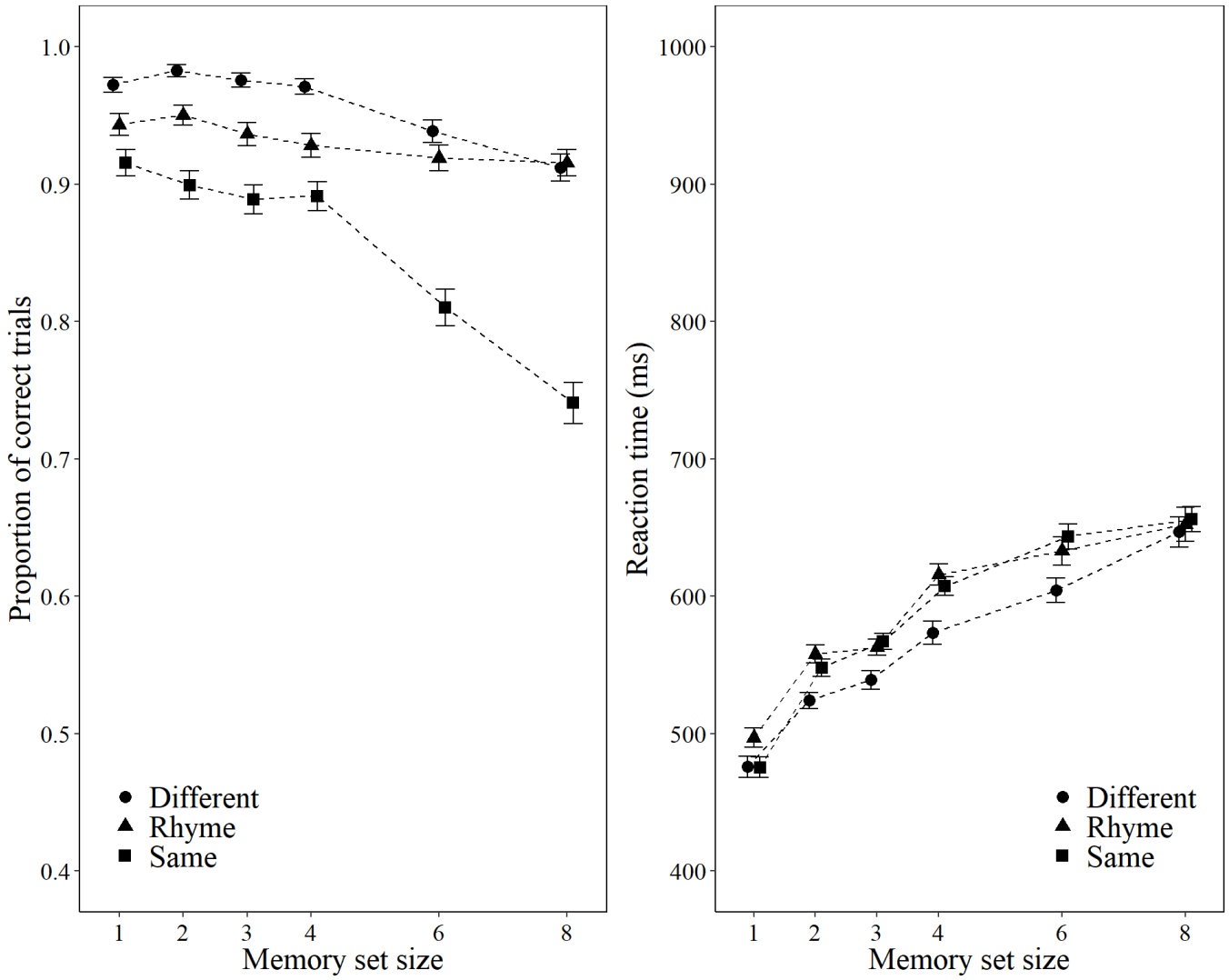

Figure 2. Mean accuracy (proportion of correct trials; left panel) and mean reaction time for correct responses (in milliseconds; right panel) as a function of probe type and memory set size for Experiment 1 (rhymes classified as different). Error bars show standard error of the mean.

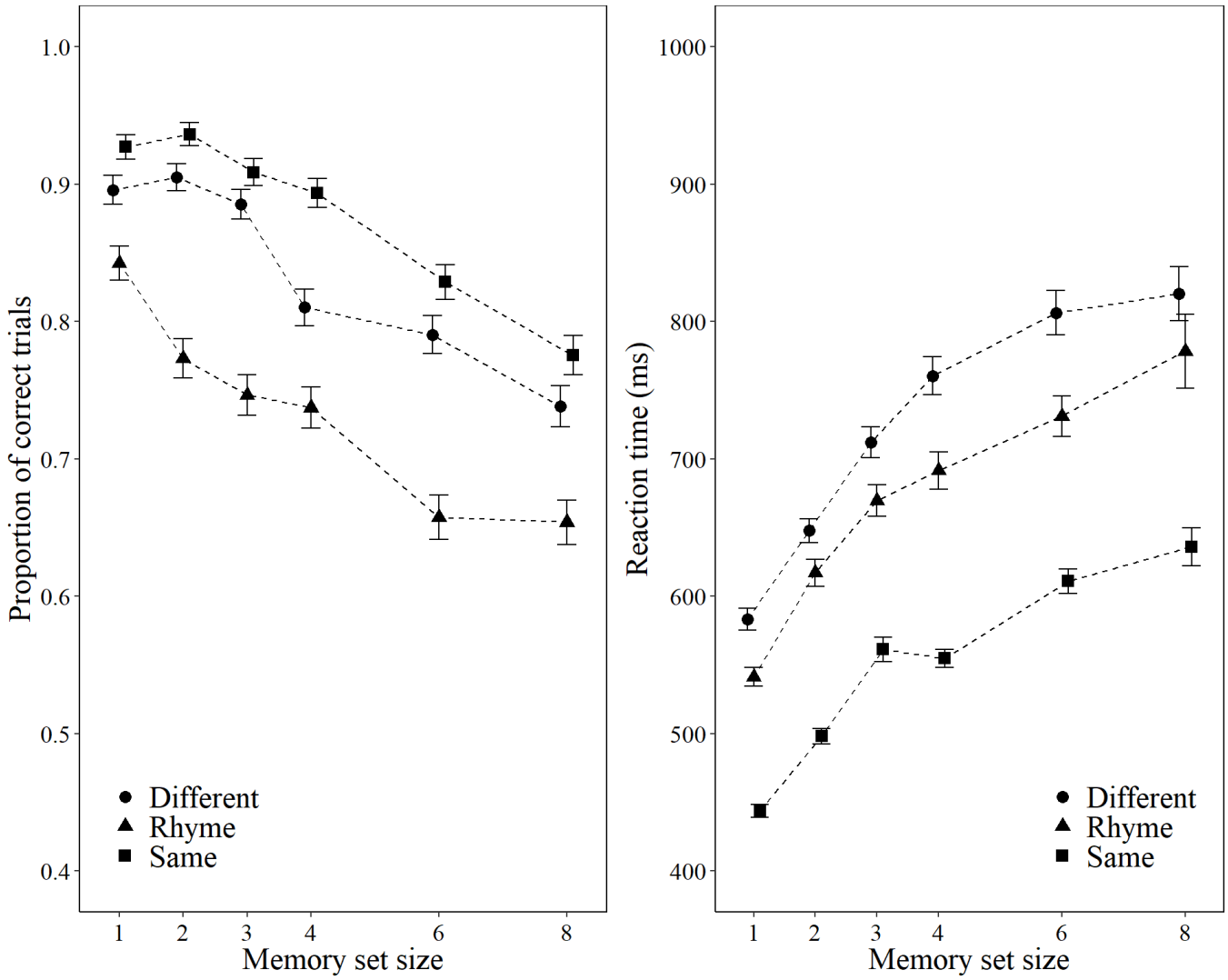

Figure 3. Mean accuracy (proportion of correct trials; left panel) and mean reaction time for correct responses (in milliseconds; right panel) as a function of probe type and memory set size for Experiment 2 (rhymes classified as same or similar). Error bars show standard error of the mean.

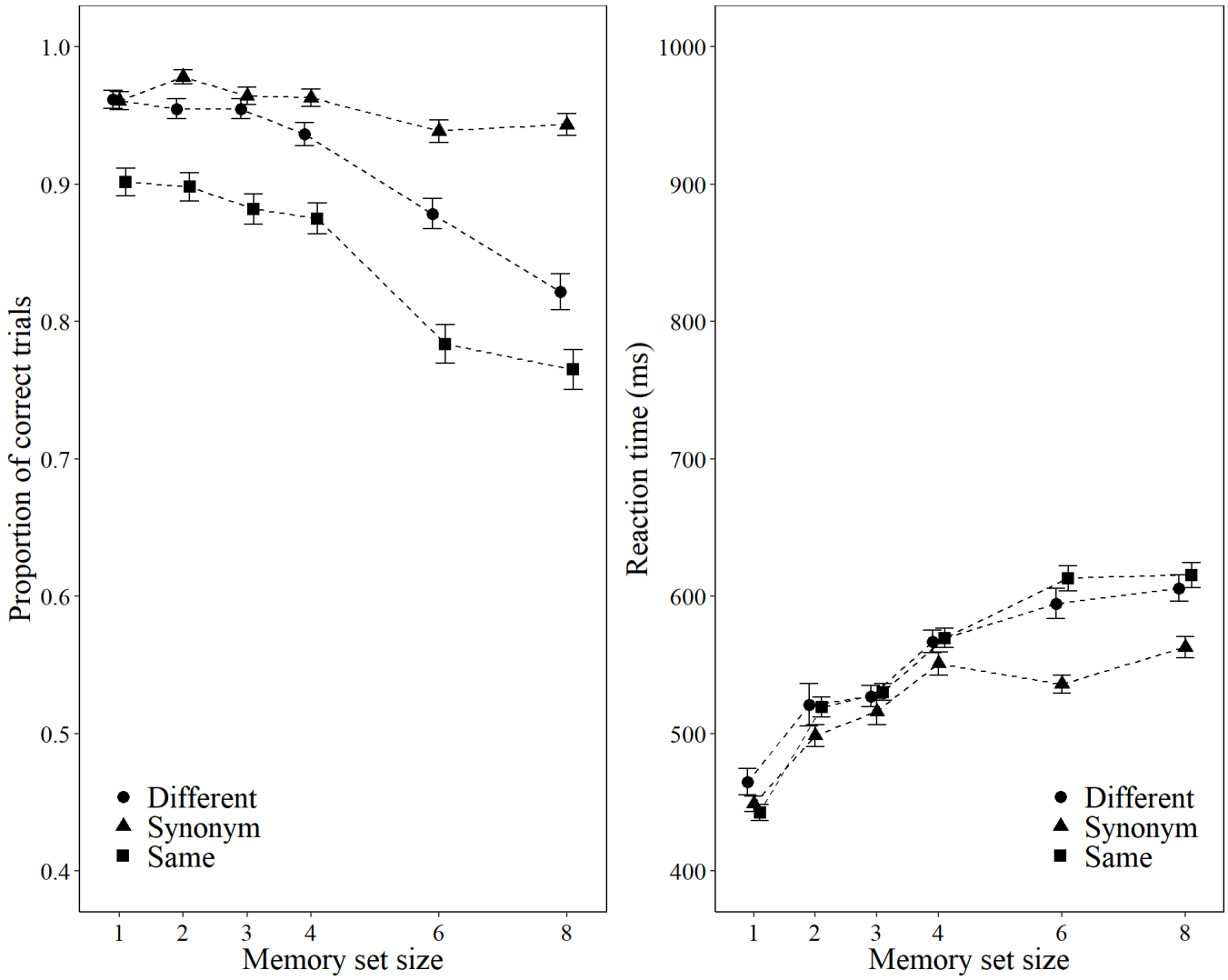

Figure 4. Mean accuracy (proportion of correct trials; left panel) and mean reaction time for correct responses (in milliseconds; right panel) as a function of probe type and memory set size for Experiment 3 (synonyms classified as different). Error bars show standard error of the mean.

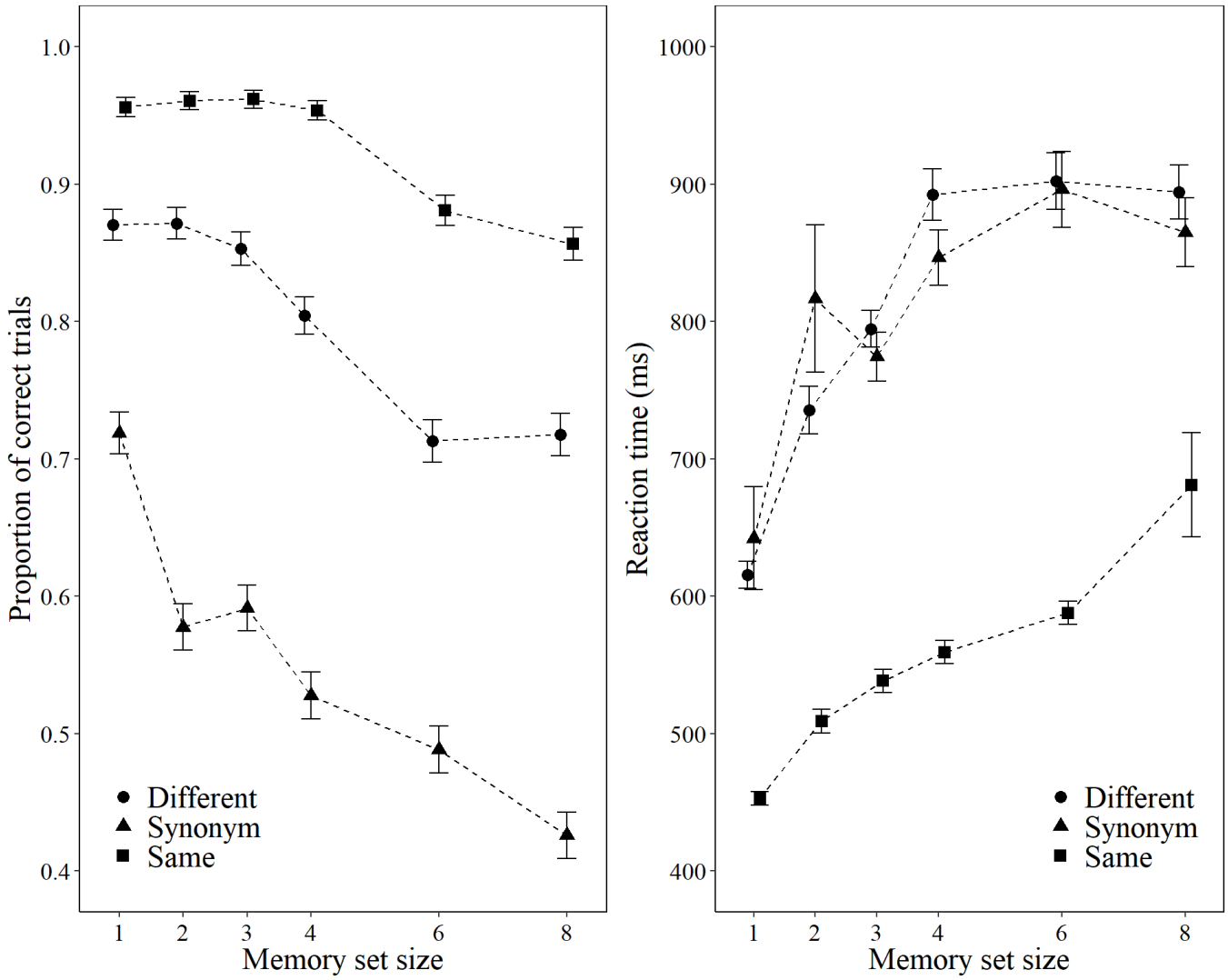

Figure 5. Mean accuracy (proportion of correct trials; left panel) and mean reaction time for correct responses (in milliseconds; right panel) as a function of probe type and memory set size for Experiment 4 (synonyms classified as same or similar). Error bars show standard error of the mean.

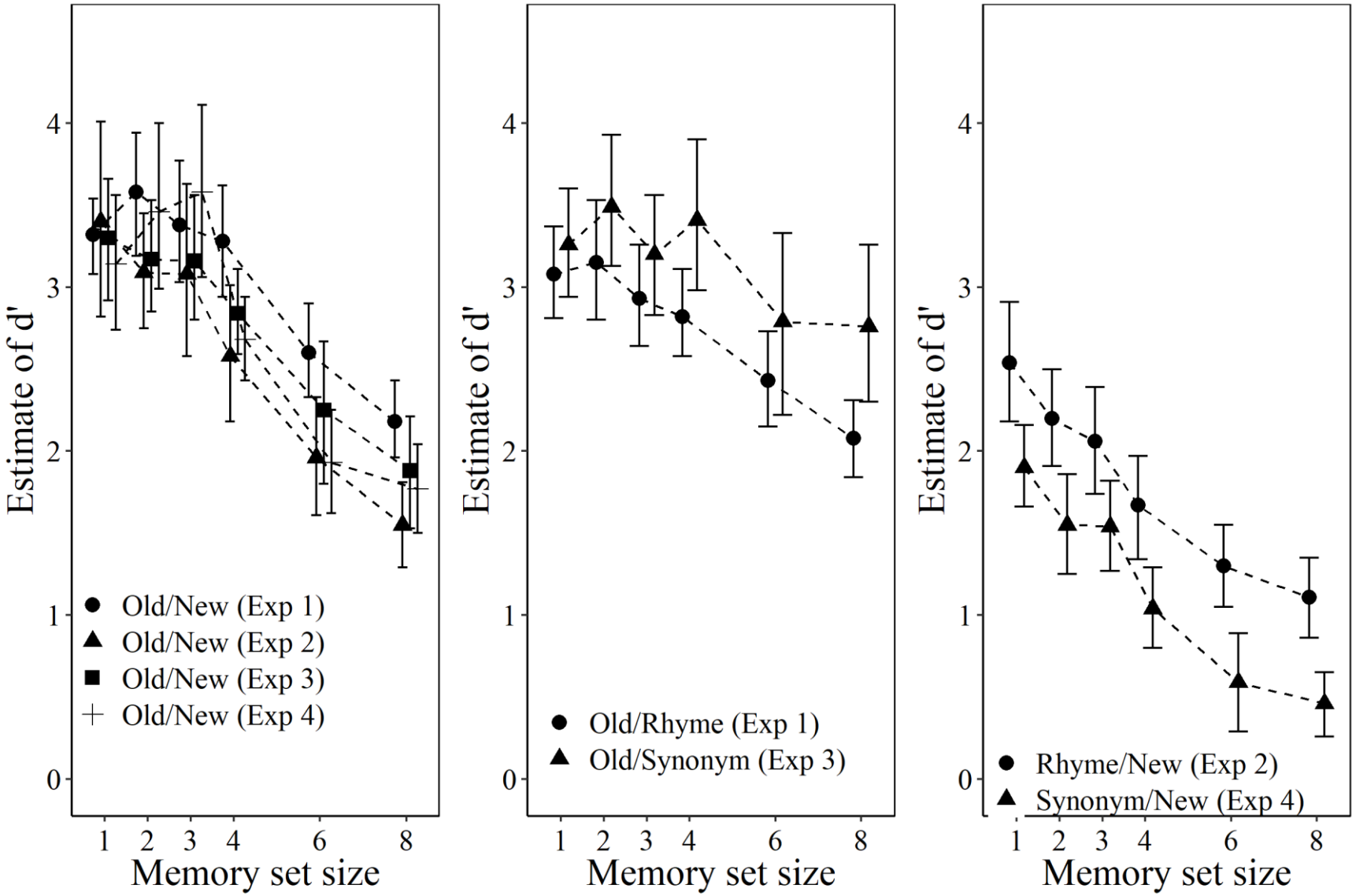
Estimate of d'
Memory set size
Old/New (Exp 1)
Old/New (Exp 2)
Old/New (Exp 3)
Old/New (Exp 4)
Old/Rhyme (Exp 1)
Old/Synonym (Exp 3)
Rhyme/New (Exp 2)
Synonym/New (Exp 4)

Figure 6. Mean estimates of *d'* across set sizes. Left panel: Discrimination of Old items from New items, separated by experiment. Middle panel: Discrimination of Old items from Similar distractors (Rhymes in Experiment 1, Synonyms in Experiment 3). Right panel: Discrimination of Similar items from New items in the accept-similar conditions of Experiments 2 and 4. Error bars represent the 95% highest posterior density interval (HDI). Estimates whose 95% HDIs do not overlap are considered reliably different from each other.