Spring 5-23-2021

Tests of Sample-recovery Models of Cued Recall

Jack Harvey Wilson
Syracuse University

Abstract

Sample-recovery models are a predominant class of episodic memory models that seek to explain why sometimes the representation of an experienced event is not retrieved or retrieved incorrectly. In these models, a correct retrieval occurs if the correct target item was sampled among the alternative studied item, then recovered correctly. In cued recall, participants output the representation of a single experienced event, a target, given a presented test stimulus and some defined relationship between the stimulus and the target. This relationship depends on the kind of cued recall and can rely on either studied or pre-experimental relationships. Sample-recovery models of this task share common testable properties related to both sampling and recovery, which we do across two experiments. Experiment 1 tests the property that sampling in sample-recovery models of cued recall is one process: they combine information about test stimulus and its relationship to the target into a single value and sample in a way consistent with the Luce choice rule. We test this assumption by testing whether manipulating the strengths of these relationships generates differential influence on performance in kinds of cued recall where different relationships between test stimulus and response are probed. The pattern of data is inconsistent with one sample process but is consistent with a sampling procedure that separately samples for a cue given the stimulus and a target given a cue. Experiment 2 tests the assumption that recovery performance is independent of other studied items. We allow some cue and target words to be related to some other untested studied words. Targets with a related word on the study list were associated with more correct responses than targets without one. This suggests that recovery in some way uses the memory for the other studied items to help retrieve. We consider how various models of sample-recovery may be adapted to account for these findings, with a particular focus on the Retrieving Effectively from Memory model.

TESTS OF SAMPLE-RECOVERY MODELS OF CUED RECALL

by

Jack Wilson

B.S., College of Charleston, 2011
M.S., Syracuse University, 2015

Dissertation
Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Cognitive Psychology.

Syracuse University
May 2021

Acknowledgements

To Dr. Amy Criss, for never holding back, for, without failure, telling me when I was wrong, and for her incredible support during my time here.

To the Cognitive Area, for being a sounding board for my ideas, for showing me new and better ways to approach problems, and for inspiring me to question settled concepts.

To my fellow advocates in SU GSO and NAGPS, for filling a gap in purpose that science sometimes leaves.

To old friends and family, for keeping me sane these past few years.

Finally, a small bit of advice to any grad student in Syracuse who is reading this, trying to figure out how to survive: gloves, a hat, and warm socks are more important in the cold than a heavy coat.

Table of Contents

Table of Figures

Table of Tables

Chapter 1

Our memory for life experiences is imperfect, as we have all experienced. Although we are capable remembering events that happened to us years ago, we routinely fail to remember desired information about our past, from where our car is parked to the names of individuals we recently met to words presented just moments ago in a psychology experiment. Episodic memory—the memory for temporally dated experienced events (Tulving, 1972)—is the phrase used to describe the act of encoding, storing, and retrieving those kinds of forgettable information. The current dogma among most human episodic memory researchers is that such memory failures are largely failures to either encode or retrieve (Shiffrin, 1970) and that the differences in forgetting are yoked to differences in the process used to retrieve information (Atkinson & Shiffrin, 1968; Shiffrin & Atkinson, 1969). This basic framework has allowed for an understanding of how memory errors can occur in a wide variety of retrieval tasks, given the same memory structure.

Here, we will be focusing on how the *sample-recovery models*, outlined by Atkinson and Shiffrin (1968), account for *cued recall* tasks. Sample-recovery and cued recall will be explored in later paragraphs. To summarize both: in a cued recall test a participant is given a *test stimulus* and asked to retrieve a studied *target word* given both the stimuli and a stated relationship between them. Oftentimes, the *cue word*—the word related to the test stimulus—is the same word as the test stimulus, sometimes they are merely similar. The stimulus, cue, and relationship can differ, resulting in multiple kinds of cued recall tasks. In experimental settings, including the experiments that follow, the cues and targets are usually words and in this document we speak of them as cue words and target words. However, *in vivo* such stimuli need not be words and other experiments have used, for example, pictures as cue or target stimuli in recall tasks (e.g. Shiffrin,

1973). In sample-recovery models, the stimulus, cue, and relationship can be jointly used to sample, or find, the to-be-recalled target which is then recovered, or transformed into a response.

The aim of this paper is threefold. First, to assess whether or not sampling can be partitioned into multiple processes: one consistent with sampling the cue word given the test stimulus, and one consistent with sampling the target word given the cue word. Sample-recovery models tend to treat sampling as a single process, so such a finding would require elaborations to many extant models of memory. If sampling indeed requires multiple processes, we wish to elaborate the inner workings of stimulus-to-cue sampling, cue-to-target sampling, and recovery by considering what kinds of errors may occur at each retrieval phase and how performance at each phase may change as a result of some basic manipulations. Finally, we intend to use this information to suggest updates to these models where appropriate.

Cued Recall

In Vivo

In the lived world, one of the more prominent examples of memory errors is the all-too-common failure to recall the name of a person one has met before. One might recall the time and place they last met and the contents of their last conversation, but not the name of the person. Hopefully, one does recall the name of the person one is conversing with and in so doing give a *correct response*. However, one sometimes forgets, evidenced by calling someone by the wrong name (*intrusions*) or by failing to recall any name at all (*response failures*). This applies equally to those met hours and days ago as much as it does to people one has interacted with mere moments ago in the same social situation, or even in the same conversation.

We call this and similar tasks *cued recall*. In the most general sense, a person attempts to output a single experienced event that possesses a given relationship between itself and a

stimulus. In the example of name-face pairs, this event is a name, but experienced events can include stimuli as varied as paintings, sentences or phrases, or randomly-sampled words presented on a computer screen. For both ease of explanation and consistency with both the arguments by example and the to-be-presented experiments, we will speak of an outputted event as a word[1].

Cued recall can take a number of forms based upon the nature of the stimulus and the relationship between the stimulus and the experienced event. The example we have used thus far—"What is the name of the person across the room?"—is *paired associates cued recall*. Other examples of paired associates cued recall include remembering that you ate a side of chips with your fish the previous night; that you saw Brad walking next to Angelina a week ago; that you left your car keys in the front pocket of your jacket; and so on. The defining characteristic of this kind of cued recall is that the to-be-retrieved word was experienced with the stimulus. In paired associates cued recall, retrieving this word entails successfully remembering that the stimulus was experienced, that the to-be-retrieved word was experienced, and that the two were experienced together.

We can also construct other plausible cued recall tasks within the context of this social situation. One could, later on, ask you the name of that person you were talking to moments ago, "…and, oh jeez, their name has something to do with headphones?" if their name happened to be

[1] We bound cued recall so that the only reasonable correct output to the test stimulus is a single word. Other kinds of recall which fall outside our restricted definition are often called "cued recall." For instance, one might be asked to recall all the women you spoke with at the party, in this case there is one stimulus "women" that prompts many reasonable correct responses—the names of all the women you spoke with at the event. This is often called cued recall, sometimes "category cued recall" (e.g. Mulligan, 2012; Roediger, 1973; Smith, 1971). In fact one can very well argue that all recall is cued recall: a free recall task might prompt you to list out everything you experienced the party, in this case there is no specific word serving as the stimulus, but recall is still prompted and guided by knowing that things at the event should be generated. We bind the scope of cued recall within this paper to those tasks asking for only one response to simplify the issues to a manageable state. However, the general principles learned should still apply to the broader set of tasks.

Dre. This is an *extralist cued recall* task: you must generate the desired word given a stimulus that you did not experience in this setting, but instead has a relationship to the desired word independent of the current setting (the headphones being Beats by Dre, which at the time of this writing were a common sight on the campuses of U.S. universities). The critical component is that the to-be-retrieved word is related to the stimulus for reasons outside the social setting. Other examples might include being asked, "Who was that person whose name rhymes with Gordon?" or, "What was their name again? I think it started with J" (the correct response to both is Jordan). In all cases, the given relationship between the name and the stimulus—a previously learned relationship, a phonological similarity, a word's first letter—was learned outside the social situation in our example.

Finally, one can use more complex relationships to find names: "Oh, we were just talking to someone that looks just like George Clooney, what was his name again?" In this case, the given stimulus, George Clooney's face, was not experienced in this social situation, and the task is to recall the name of someone you just met with a face similar to his. The to-be-retrieved event was not studied with the stimulus or with something related to the stimulus, but instead was experienced at the same time as the something related to the stimulus. This task has components of both extralist and paired associates cued recall. Call the task *hybrid cued recall*.

In Experiments

In the examples above, we used a social setting where the task was to recall a name given a face. In an experimental setting, we often use pairs of words rather than names and faces, but the same principles apply. In a typical cued recall experiment, participants study a list of word pairs. Later on, as one of several test trials, participants are presented with some specific word as a *test stimulus* and are prompted to output a *target* word, given a relationship between the test

stimulus and the target. In these experiments, there also exists a *cue word*, a word related both to the test stimulus and the target word in some way. In some cases, the cue word and the test stimulus are the same word, while in others the cue and target word are the same word[2]. Table 1 offers an example of how such words may map to paired associates cued recall, extralist cued recall, and hybrid cued recall.

Table 1

Mapping of test stimulus, cue word, and target word to cued recall tests, by example

Term	Study word pair APPLE - BASEBALL Word referred to by term in example cued recall task		
	Paired Associates	Extralist	Hybrid
test stimulus	APPLE	BAT	BANANNA
cue word	APPLE	BASEBALL	APPLE
target word	BASEBALL	BASEBALL	BASEBALL

Note: In an experimental setting, the test stimulus is the word presented during the test phase and the target word is the studied word a participant in an experiment should report given the test stimulus.

We wish to have models of memory that are able to jointly account for all three of these cued recall tasks. Extant models that claim to account for cued recall findings tend to account for either paired associates cued recall (e.g. Search of Associative Memory (SAM), Raaijmakers & Shiffrin, 1981) or extralist cued recall (e.g. Processing Implicit and Explicit Relations (PIER),

[2] This set of definitions is something of a departure from how these words have been defined in the context of either paired associates cued recall or extralist cued recall. In both tasks the phrases "cue word" and "test stimulus" are conflated but have different meanings depending upon the task in question. In paired associates cued recall, the cue word is also the word studied alongside the target (e.g. Wilson & Criss, 2017), whereas in extralist cued recall the cue word is unstudied (e.g. Nelson & Goodmon, 2002). If one studies the word pair APPLE-BASEBALL and BASEBALL is the target word, APPLE would be the cue word in paired associates cued recall and BAT would be the cue word in extralist cued recall. APPLE and BAT are different words, so calling both "cues" when speaking of both tasks at once is confusing. The term "context word" does not suffice either. "Context word" refers to, particularly in the older literature, the word studied alongside the target (e.g. Postman, 1975; Roediger & Adelson, 1980). For paired associates cued recall, the context word, cue word, and test stimulus are same word. In extralist cued recall, the context word, if there is one, is either untested or a different target word. In hybrid cued recall, the context word and cue word are the same word, but the test stimulus is a different word.

Nelson et al., 1992 and later versions), with only the Matrix Model offering an implementation for both (Humphreys et al., 1989). Hybrid cued recall is a novel task; no extant model has been extended to it. Rather than consider, model-by-model, how each model jointly accounts for all three tasks, we leverage their common properties as sample-recovery models to assess what predictions such models jointly make about the three cued recall tasks.

Sample-Recovery Models of Memory

A sample-recovery model, in its most general form, retrieves information by sampling—searching either deterministically or probabilistically for some desired information in memory—then recovering that information—preparing it for output and generating a response. Sample-recovery models assume the existence of some kind of long-term memory store containing representations of experienced events, although the assumptions made about the store itself vary from model to model. Sample-recovery models encompass most memory models stemming from Atkinson and Shiffrin (1968) and virtually all modern models of recall.

The concept of sample-recovery stems both from the adoption of visual search principles to memory (e.g. Sternberg, 1966) and the implementation of memory theories inspired by computer science (e.g. Feigenbaum & Simon, 1962). Yntema and Trask (1963), citing Feigenbaum's (1961 as cited by Yntema and Trask) work, proposed that recall could be explained as a sequential search through memory, determining for each individual memory trace whether or not to output. Feigenbaum (1966) and Hintzman (1968) elaborated on this concept, suggesting more general directed searches that are deterministic but not based on input order. Atkinson and Shiffrin (1965), alternatively, suggested that this search may be random and follow the Luce choice rule (Luce, 1959). The concept of memory search was solidified as a central component of retrieval by the Modal Model (Atkinson & Shiffrin, 1968).

From this point, sample-recovery models began to diverge as surrounding assumptions pertaining to the structure of the long-term episodic memory store, statistical regularities of the memory system, and the control processes used to guide retrieval were each tested and refined. One critical distinction was how the memory store was structured, with *discrete trace models* structuring the memory store as a list of traces, and *composite memory models* structuring the memory store as a composite with no unitizable traces. Sample-recovery is employed in both classes of models, even if the mechanism is not immediately clear from the model design. Outside of questions pertaining to catastrophic forgetting, list length, and list strength effects, both classes of models are often considered equivalent and face similar challenges in accounting for cued recall data (Raaijmakers & Shiffrin, 1992).

Discrete trace models assume that the long-term episodic memory store is structured as a set of unitizable, discrete traces that each represent some experienced event. The best known member of this class is probably the Search of Associative Memory model (SAM: Raaijmakers & Shiffrin, 1981). Other members of this class include the Retrieving Effectively from Memory (REM: Shiffrin & Steyvers, 1997) and the Processing Implicit and Explicit Representations model (PIER: Nelson et al., 1992). For these models, how a sample-recovery mechanism works is fairly simple to grok in that traces are "things" that can be selected and recovered.

Other models, known as *composite memory models*, assume that the long-term episodic memory store is a single pool of information and that experienced events perturb this pool in a predictable way. Such models include the Matrix Model (Humphreys et al., 1989) and Theory of Distributed Associative Memory (TODAM; Murdock, 1982). To borrow an analogy from Murdock (1982), such a model's representation of an experienced event can be likened to how the ripples in a pond represent the pond's experience of a rock. Sample-recovery is also used in

this class of models as well, although the verbal connection is not as clear as in the discrete trace model. Composite memory models retrieve by extracting an approximation of an experienced event which is then transformed into something verbalizable. In these models, the process of extraction is a sampling process and the transformation is a recovery process, as will be discussed.

What is to Follow

As sample-recovery models naturally consist of two components, sampling and recovery, we shall consider the sampling and recovery components of cued recall separately in Chapters 2 and 3. In each case, we will outline some of the basic principles of the component; go through implications and critical predictions that arise from those principles; present data that evaluates those predictions; and discuss the general implications of our findings for sample-recovery models.

In Chapter 2, we focus on sampling. Sample-recovery models in cued recall almost universally treat sampling as a single process. This restriction, combined with other universal assumptions about how sampling and memory more generally works, leads to strong predictions both about how the representation of the cue word and of the cue-target association interact and about the relative performance levels of performance in different forms of cued recall. Prior experiments jointly testing extralist cued recall and paired associates cued recall violate these predictions, as do word frequency and set size cued recall experiments. However, these findings are accountable if the representations of the cue and the cue-target association are separately used in two sampling processes. We test the predictions of one- and two-sample processes in a single experiment where cue strength and cue-target associative strength are jointly manipulated at study and measure the effect of these manipulations on paired associates, extralist, and hybrid

cued recall performance. The findings of this experiment are not consistent with the predictions of a one-sample process, but they are consistent with a two-sample process. We discuss implications of these findings for multiple memory models.

In Chapter 3, we focus on recovery. One basic assumption of models that implement recovery is that only the information gained from sampling is used during recovery. In the language of discrete trace models, if the target trace is sampled, the joint probabilities of outputting a word and, if so, of that word being the target word are independent of the other words on the study list. This is in contrast to sampling: the presence of a word similar to the cue at study would be expected to harm paired associates cued recall performance. We test this assumption by, at study, presenting words that are related to the cue word or the target word in a counterbalanced fashion. As expected, the presence of an item related to the cue word at study harms performance. However, contrary to what might be expected from recovery, the presence of a word related to the target at study increases the probability of correctly recalling the target in a paired associates cued recall test. This has different consequences for discrete trace and composite memory models, with the former requiring more extensive adjustments to account for the data.

Finally, in Chapter 4, we delve into several sample-recovery models. We outline in broad strokes how they handle (or might handle) cued recall tasks and propose updates to those models as necessary. Most of this chapter concerns updating the Retrieving Effectively from Memory model to jointly account for paired associates, extralist, and hybrid cued recall.

Chapter 2: Sampling

Sampling is the process by which a to-be-generated memory trace is accessed. There are a number of distinctions that are made subdividing the kinds of sampling procedures: deterministic versus probabilistic, sampling-as-selection versus sampling-as-extraction, and so on. The variety of sampling processes, on a more abstract level, are far more alike than they are different; they share several important and testable properties. Ultimately, the relevant metrics of the sampling process in these models can be expressed through a single Luce choice equation. We call these one-sample processes.

Relevant Sampling Properties Exist in Both Discrete Trace and Composite Memory Models

There are two ways sampling occurs in these models. The first, completed by discrete trace models, is a selection of a trace based upon some amalgamation of the strength[3] of relationship between the test stimulus and the cue memory and the strength of the cue-target association. The second, completed by composite memory models, infers or extracts a representation of the proposed target from the long-term store using a function that approximately reverses the encoding process. In both cases, sampling strongly predicts that increased strength of the cue word or the association should increase the probability of sampling the correct target memory, imposing a monotonically increasing relationship between the strengths and correct response odds.

In discrete trace memory models, experienced events are stored as separate, discrete memory traces. Sampling selects one of these traces as the proposed target; the exact process

[3] Note that strength is a loaded term in the literature. While avoiding as many theoretical pitfalls as possible let us just say that strength refers to more complete and/or higher-quality representations of events. Within the context of Luce choice sampling, the strength of a memory is the size of the number fed into the rule.

differs from model to model. Typically, the selection is completed by random sampling using a Luce-choice rule: the probability of selecting item i is a function of the weighted strength of that item divided by the sum of the weighted strengths of all $k = 1{:}\,N$ items:

$$\Pr(i|j) = \frac{\mathrm{S}(i|j)^{\gamma}}{\sum_{k=1}^{\mathrm{N}} \mathrm{S}(k|j)^{\gamma}} \quad (1)$$

Where $\mathrm{S}(i|j)$ is the net strength of item i given item j and parameter γ weighs the net strengths. If $\gamma = \infty$, then the strongest trace is selected every time and, if $\gamma = 0$, all traces are equally likely to be sampled. We can make similar claims about the extraction of traces from composite memory models because shared variance between target and extracted information is expressible in the same form, see Appendix A.

Different models use different information when Luce-choice sampling. SAM, for instance, relies primarily upon the strength of association between the cue and target words and the strength of association between the target word and the context of study/test. REM relies upon the degree of match between the contents of the test stimulus and the representations of the cue word and context. PIER relies on semantic strengths gathered from the free association task. The Matrix Model uses a combination of strength (the magnitudes of the studied vectors) and cosine similarity, again see Appendix A. As we will discuss later, this makes some models better adapted to account for some cued recall tasks than others. Critically, all this information is used in the same way, so we can make some claims about the entire class.

Properties of Sampling

Within the confines of episodic memory models, there are four sampling properties common to all sample-recovery models that are relevant to our research questions. These properties are: The probability of sampling the correct item is monotonic to the strength of the item; sampling is competitive; sampling amalgamates the different information attended to and

available about a sample-able trace into a single value; and sampling odds are improved when more relevant information is provided at test.

The probability of sampling a trace is monotonic to the strength of that trace, as we demonstrate here. The relationship between two variables is monotonically increasing if increasing the value of one never leads to a decrease in the value of the other. We have established that sampling is expressible as a Luce choice rule, so we can demonstrate that this first property is true using the Luce choice rule. In this framing, we can show that selectively increasing $S(i|j)$ in the Luce choice equation will never decrease the probability that item i will be sampled given j. Consider the Luce choice function $f(a) = \frac{a^\gamma}{a^\gamma + \sum b_i{}^\gamma}$, where $f(a)$ is the probability of sampling a, a is the strength of the correct trace, and b_i gives the strength of the incorrect traces i. These variables are only ever expressed as values greater than or equal to 0. The first partial derivative of the Luce choice equation outlined above, with respect to a, is:

$$f'(a) = \frac{\gamma a^{\gamma-1} \sum b_i{}^\gamma}{(a^\gamma + \sum b_i{}^\gamma)^2} \quad (2)$$

Positive real-valued numbers a and b_i are added, divided, multiplied, and raised to a positive real-valued power γ. These functions in combination can only produce positive numbers, therefore the probability of sampling a monotonically increases with the strength of a.

Sampling is a competitive process. In other words, increasing the probability of selecting a given item must decrease the probability of selecting any other item. In the Luce choice rule, increasing the odds of sampling trace i decreases the odds of sampling one of the other k traces and this can only be so. In composite memory models, a stronger competing cue will make the extracted information look more like the target word the competing cue was studied with and therefore (assuming the competing words were studied with unrelated targets) will make that

extracted information look less like the correct target. Thus, the sample-recovery model class predicts that, in paired associates cued recall, strengthening the stored representations of the tested words should increase the correct response rate and lower the intrusion rate, before considering other elaborations or effects of recovery.

Sampling processes amalgamate what information is used during sampling, which makes it difficult to dissociate information sources. This is a consequence of episodic memory models taking the form of *compound cueing models* (Medin, 1975)[4] for reasons that go beyond the confines of sample-recovery models of recall. Take SAM as an example of a compound cueing model. In SAM (Raaijmakers & Shiffrin, 1981), the sampling strength of a target trace in paired associates cued recall is a product of its binding to the context of study-test S_{cj} and its binding to the test stimulus / cue word S_{ij}: $S(j|c,i) = S_{cj}S_{ij}$. These values are computed for each sampleable trace and the sampling probability is a Luce choice function of those values. SAM does and other sample-recovery models may allow strategic weighting of the information used during sampling, for example by attending to only context during some stages of the free recall search and both context and item-to-item associations during others.

This final property has two consequences of note. The first is that the compound cueing process amalgamates the information sources into a single number. The Luce choice equation sees the product, not the factors. Doubling the strength of one factor will have the same consequence no matter which factor was doubled. The second consequence is that adding more discriminatory information to the sampling process will always make sampling better. Essentially, these models state that the apparent similarity of an item that matches across two

[4] Medin and others, e.g. (Nosofsky, 1984) call them "context models." To disambiguate the concept they describe from context as a form of incidental or temporal or situational information that retrieved context memory models (among others) rely upon, we borrow the term "compound cue" that Ratcliff and McKoon (1988) use to describe the principle.

points of comparison is greater than the apparent combined similarity of two items that match across just one (and no further comparisons are made in either case). The more information that is used to sample, the more likely it is that the correct trace will be sampled. For example, adding representations of cue-target associations to the sampling process would tend increase the odds of sampling the correct target trace. Omitting that information due either to lack of explicit utility in the task or lack of knowledge would tend to harm performance. At the very least, increasing the amount of useful information available should never make matters worse, nor should decreasing the amount of useful information available make matters better.

That strengthening the representations of events through additional study improves retention and increases odds of later retrieval is one of the oldest findings in the memory literature (Ebbinghaus, 1895). Accounting for that point is a necessary feature of any memory model. Likewise, presenting more and more accurate information at test about the correct target (such as using the cue word as the test stimulus rather than a related word) in cued recall should increase its likelihood of recall. Sampling, as currently implemented in cued recall models, naturally allows for these things to happen. However, in much the same way that recognition testing cannot separate familiarity and recollective-based information (Dunn, 2008) because it is all transformed into a single dimension during retrieval, amalgamating cue and association during sampling leads to the prediction that effects from manipulating either form of information cannot be isolated. Likewise, the fact that all additional information should potentially help memory performance (and certainly not hinder it) leads to the prediction that cued recall tasks which utilize more information should generally perform better.

We can therefore say that the sampling process is consistent with the effects of manipulating cue and associative strength if such manipulations produce data with the following

properties: strengthening the correct cue and/or cue-target binding increases the probability of correct responding, this in turn decreases the probability of incorrect responding; that this should also be true of tasks that cue with more information versus less; and that manipulations of cue strength and associative strength cannot be fully disentangled. Violations of these properties should be construed as evidence against sampling, in its current form, as described by strengths amalgamated by compound cueing and fed into a single Luce choice equation.

Sampling in Cued Recall Tasks

A review of the word frequency and set-size effects in cued recall, as well as a consideration the relative performance of paired associates versus extralist cued recall, highlights issues for a sampling process as outlined above. Word frequency measures how often a word is expected to appear throughout one's lifetime, while pre-experimental set size is a measure of how many words a given word is semantically related to. We will discuss in-depth the word frequency and pre-experimental set size findings for cued recall and other tasks in the discussion. Suffice it to say, paired associates cued recall performance is better for high frequency words and words with large pre-experimental set sizes, while extralist cued recall favors low frequency words and words with small pre-experimental set sizes.

The evidence available suggests no reliable ordinal relationship between extralist cued recall and paired associates cued recall correct response rates. Multiple studies show greater extralist cued recall correct response rates than paired associates cued recall, contrary to prediction; the lack of consensus suggests multiple factors at play. Postman (1975), for example, intermixing tests of paired associates cued recall versus extralist cued recall for word pairs versus single items, found greater correct response rates for extralist cued recall than paired associates

cued recall for unrelated words or word pairs[5]. For some instances where paired associates cued recall gave more correct responses than extralist cued recall, see (Nelson et al., 1993).

This pattern also poses a problem for sample-recovery as a single Luce choice rule. Sampling represented as a single Luce choice rule predicts that paired associates cued recall and the hybrid task will yield more accurate responses than extralist cued recall. As per the properties of the Luce choice rule and of compound cueing, more information and higher-quality information increase chances of correct sampling and decrease chances of incorrect sampling. In paired associates cued recall, the test stimulus is related to the cue in that they match, while in extralist cued recall they are different words. Paired associates cued recall should therefore have better information about the cue word given the test stimulus than extralist cued recall should. At the same time, successful paired associates cued recall should rely on the information associating the cue and target words, whereas successful extralist cued recall need not.

This problem exists because a sampling process that amalgamates all information into a single Luce choice rule does not account for task difficulty. Although the single sampling process can use more information to complete paired associates cued recall, it does not require that extra information to accurately do so. Hybrid cued recall, for example, should be harder to complete than extralist cued recall because successful hybrid cued recall requires access to information (the cue-target association) that extralist cued recall does not need. Returning to the social example, recalling the name of the man that looked like George Clooney or the name of the person standing in front of you requires more information to do successfully than just recollecting their face, or alternatively recalling the name of that person after being told their name rhymes with "Gordon." However, sampling as presented does not allow for these kinds of

[5] As per the mean stringent scores for W-W + C-C for paired associates cued recall versus W-S for both C-C and N-C for extralist cued recall.

task difficulty adjustments because all the necessary information is fed into a single sampling rule.

To be clear: this is not a strawman. The models we have mentioned or discussed (SAM, REM, PIER, Matrix Model, TODAM) and others all implement cued recall with a single sampling process and share the critical properties.

Is Sampling More Than One Process?

The failure of the sample-recovery models outlined here is not a problem with sample-recovery as a general theory, necessarily, nor is it necessarily a failure of the theories that use them. The failure presented may instead simply be a lack of elaboration of the control processes the models use to complete cued recall. No such elaboration has been necessary up until this point, in part due to a dearth of relevant data. The lived example of recalling a name given a face and different kinds of name-face relationships obviates the point. Paired associates and hybrid cued recall may simply require sampling the correct cue word trace given the test stimulus **and** sampling the correct target word trace given the cue word trace. Extralist cued recall might bypass this second step because the representation of the cue and of the target are the same representation.

We outline this *two-sample* framework below as an alternative to the *one-sample* framework that is used by extant sample-recovery models. For simplicity, let us assume that errors in the retrieval system are not rectifiable: once something goes wrong, it results in either an incorrect response or a failure to respond. Of course, mathematically speaking, if retrieval is a random process then rectification could happen by random chance. However, given the possibility space involved, the probability of rectification by random chance is exceedingly small. A participant in a typical long-term episodic memory experiment may study a list of 16 or

more word pairs and is expected to select the representation of one word from those of 32 or more studied words as well as representations of other words from before study, from any intermediate distractor task, and from test. On top of that, a college student knows, by some estimates, approximately 17,000 words (D'Anna et al., 1991; Goulden et al., 1990). On top of that, the odds of rectifying by random chance are the odds of making an error that happens to result in the correct response (the probability that two wrongs will make a right). The memory system samples correctly at a rate greater than chance. Therefore, the probability of two wrongs making a right is even smaller than expressed by the outcome space. If rectification happens at a rate worthy of consideration, then it must be because of additional processes in sample-recovery.

We outline the two-sample framework predictions on the level of correct responses because the primary concerns about the one-sample framework rest on correct response rate predictions. The framework divides the process of sampling into two separate processes: *stimulus-to-cue sampling* and *cue-to-target sampling*. Stimulus-to-cue sampling represents the process by which the cue trace is sampled given the test stimulus, and cue-to-target sampling represents the process by which the target trace is sampled given the cue trace. Sampling correctly in both is necessary to select the target trace. Let x_{right} be the strength of the cue trace, and y_{right} be the strength of the association between the cue and target. Let $\{x_{wrong}\}$ be the set of other item strengths—the set of incorrect alternatives to the cue trace—and $\{y_{wrong}\}$ be the set of incorrect alternatives to the target trace. These are all strength values, in that they represent some factor that determines the probability of being sampled in a Luce choice function. Let $p = \mathrm{L}(a, \{b\})$ be the Luce choice probability of selecting item a over the set of incorrect alternatives $\{b\}$. The model functionalizes the probability of selecting the correct target in each task:

$$\Pr(correct|ELCR) \sim s\mathrm{L}(x_{right}, \{x_{wrong}\}) \quad (3)$$

$$\Pr(correct|PACR) \sim \mathrm{L}\left(x_{right}, \{x_{wrong}\}\right)\mathrm{L}(y_{right}, \{y_{wrong}\}) \quad (4)$$

$$\Pr(correct|HYBR) \sim s\mathrm{L}\left(x_{right}, \{x_{wrong}\}\right)\mathrm{L}(y_{right}, \{y_{wrong}\}) \quad (5)$$

We multiply the probabilities in extralist cued recall (ELCR) and hybrid cued recall (HYBR) by $s < 1$, representing the concept that finding the correct cue trace should be more likely in paired associates cued recall (PACR) than the other two tasks.

A number of predictions arise from this framework. The first is that the hybrid cued recall task (in the limit) will have the lowest correct response rate of the three tasks. This is because both of the Luce choice probabilities and the parameter s are less than 1. Therefore, successful hybrid cued recall will be less likely than successful extralist cued recall by a factor of $\mathrm{L}(y_{right}, \{y_{wrong}\})$ and less likely than successful paired associates cued recall by a factor of s. The model is agnostic as to whether paired associates cued recall should have more or fewer correct responses than extralist cued recall because *s* may be larger or smaller than $\mathrm{L}(y_{right}, \{y_{wrong}\})$ depending what stimuli are used and how the experiments are designed.

The framework also strongly predicts that manipulations of cue strength and cue-target association strength will generate differential influence on extralist cued recall versus the other two tasks. Simply, extralist cued recall is unaffected by cue-target association strengths in this framework. Therefore, strengthening the cue-target association is predicted to improve performance in paired associates and hybrid cued recall, but not in extralist cued recall. Further, the same manipulations will not differentially influence paired associates and hybrid cued recall.

This gives us four hypotheses to be tested:

1. Increasing the strength (e.g. study time) of the cue word or the target word should increase correct response rates generally.

2. Increasing the strength of the cue-target association should increase correct response rates in paired associates cued recall and hybrid cued recall and have no effect on the extralist cued recall correct response rate.
3. Hybrid cued recall should have the lowest correct response rate.
4. Joint manipulation of cue strength and strength of the cue-target association should reveal no differential influence when considering paired associates and hybrid cued recall correct response rates as dependent variables, but differential influence when extralist cued recall correct response rates are jointly considered with those from either task.

Towards Identifying a Two-sample Process

The first critical task of this manuscript is to determine whether sampling can be decomposed into two separate samplings: one related to the relationship between test stimulus and the cue (stimulus-to-cue sampling), and one related to the relationship between the cue and target (cue-to-target sampling).

We start by performing a conceptual replication of Hockley and Cristi (1996) to determine whether the half-seesaw effect observed in item and associative recognition is also observed when comparing extralist cued recall to paired associates or hybrid cued recall. The half-seesaw effect is the observation that manipulating the study task such that participants are instructed to study the pair members separately rather than together (in two sentences or one sentence, respectively) selectively improves performance for associative recognition performance without harming item recognition performance. The presence of that effect in cued recall would indicate that the strength of the cue-target association does not influence extralist cued recall, but alone this is not sufficient to favor the two-sample framework.

To evaluate the two-sample framework, we employ multiple forms of analysis, each of which does well at assessing some of the predictions made by the framework, but not others. Namely, in addition to the analysis of variance which will among other things help determine the presence or absence of the half-seesaw effect, we employ state-trace analysis to more robustly consider whether the patterns of data can be described by one versus multiple latent processes, and multinomial modeling to drill down on qualitative aspects of the two-sample framework such as how error-prone certain forms of sampling are and what kind of errors result from what kinds of sampling.

ANOVAs

The ANOVA is used to analyze the presence or absence of the half-seesaw effect, whether cue and target strengthening universally improves cued recall performance, and the relative overall performance of the three cued recall tasks. We run all ANOVAs in JASP (JASP Team, 2020) using both frequentist and Bayesian methods with default parameters ($r_{fixed} = 0.5$, $r_{random} = 1.0$, $r_{covariates} = 0.354$).

State-trace Analysis

We use state-trace analysis (specifically, conjoint monotonic regression) to determine whether manipulations of cue strength and association strength load on two latent sampling processes underlying cued recall. State-trace analysis offers a robust test of interactions and in turn allows us to properly infer whether multiple latent monotonic processes are necessary to account for the relationship between dependent and independent variables. State-trace analysis states that two or more independent variables differentially influence two dependent variables if the relationship between the dependent variables under manipulation is not monotonic. We

recommend reading Dunn and Kalish (2018) for the details and general logic of state-trace analysis.

The two-sample framework predicts that jointly manipulating cue and cue-target association strength will cause differential influence on extralist cued recall response rates versus either paired associates or hybrid cued recall correct response rates (Figure 1). Because paired associates and hybrid cued recall use both stimulus-to-cue and cue-to-target sampling, the framework predicts no differential influence of the joint manipulation on paired associates and hybrid cued recall correct response rates. In contrast, a single sampling process that utilises the same information to make cued recall decisions does not predict differential influence of cue and cue-target association strength on the different cued recall correct response rates.

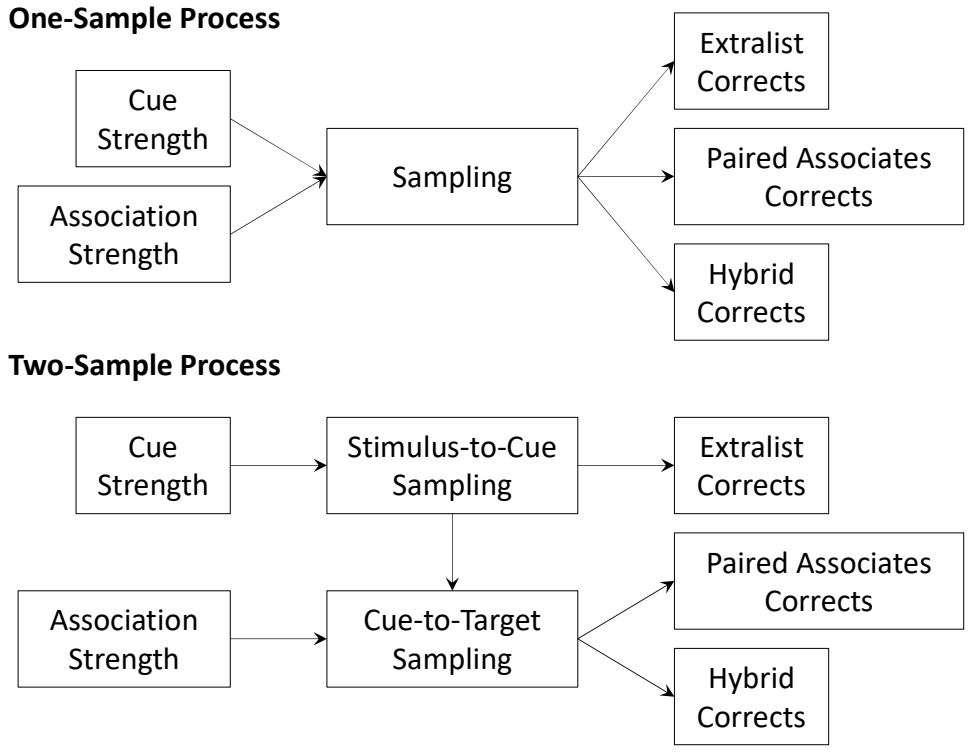

Figure 1. Diagrams of a one-sample process versus a two-sample process.

Bayesian Multinomial Modeling

If the two-sample framework holds, we can then use a Bayesian multinomial model to get a more complete picture of what our experimental manipulations are doing. We use a Bayesian multinomial model to assess the level of performance and the effect of manipulations on three processes in whole: stimulus-to-cue sampling, cue-to-target sampling, and recovery. We construct a model that treats cued recall as the completion of those three phases in sequence. A correct response requires successful completion of all three phases. We constrain the model in several ways. We impose selective influence of cue manipulations, cue-target association manipulations, and target manipulations on the three phases respectively. We equate stimulus-to-cue sampling in extralist cued recall and the hybrid task because in both tasks the test stimulus is merely similar to a studied item. We equate cue-to-target sampling in paired associates and hybrid cued recall. We equate recovery across all three tasks as we believe the critical difference in the tasks is on the sampling side. Finally, we fix the probability of successfully completing cue-to-target sampling in extralist cued recall to one: once you have found the cue trace you have found the target trace because, in this framework and task, cue and target are one in the same.

Experiment 1

Methods

Participants

$N = 189$ native-English-speaker undergraduate students at Syracuse University participated in this experiment for course credit.

Materials

The words used in this experiment were selected from the University of South Florida (USF) free association norms (Nelson et al., 2004). In all, we constructed two lists: first, a list of related word pairs, then a list of words unrelated to any of the word pairs.

The USF norms offer two key measures of the semantic relatedness of two words: forward strength (the probability of responding with word A when word B is the test stimulus) and backward strength (the probability of responding with word B when word A is the test stimulus). A pair of related words was included in the list of pairs if both members had a forward and backwards strength to each other between 0.14 and 0.30, were between 4 and 12 letters long, and had logHAL and SUBTL entries in the English Lexicon Project (Balota et al., 2007). This yielded a list of 132 pairs, or 264 words, after eliminating the pair STATE-FLORIDA (the USF norms were collected in Florida and our experiment was conducted in Syracuse, New York).

We then built a list of unrelated words from the remaining, unused words in the USF norms. We excluded any word that had a forward or backwards strength greater than 0 with any of the 264 words from that first list. Then, for each of the 264 words, we selected the two words in the pool that best matched along the dimensions of logHAL word frequency and logSUBTL context diversity using `knnsearch` in MATLAB. Some words selected more than once by `knnsearch`, so this yielded a list of 370 unique words unrelated to any of the 264.

Design and Procedure

This first experiment utilized a 3 (encoding strength: none strong vs. cue strong vs. target strong) x 2 (study task: study together vs. study separately) x 3 (test task: paired associates vs. extralist vs. hybrid cued recall) within-subjects design, with strength of cue and of target

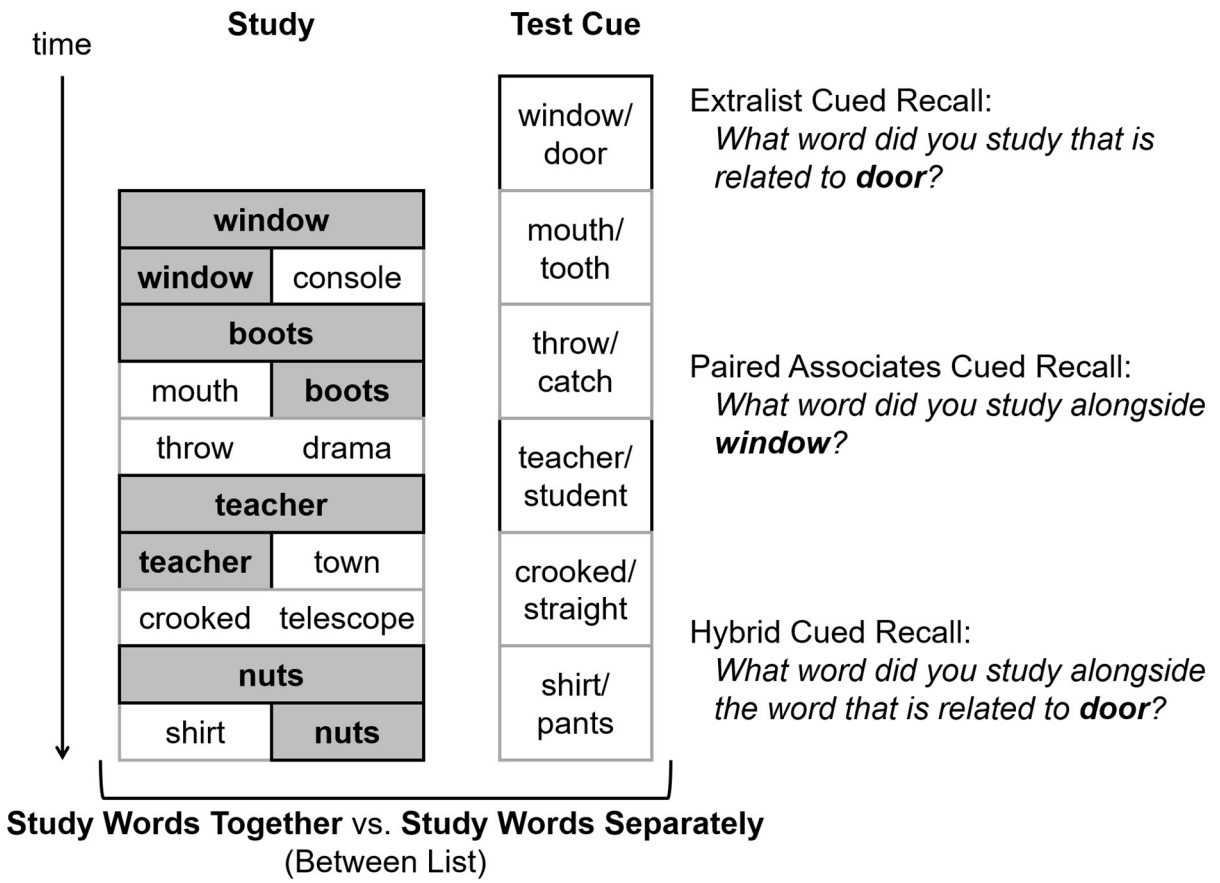

Figure 2: Conditions of Experiment 1. List length is not represented. Study time is not to scale. Left-right position of strong words was randomized. The role of a pair member as cue or target was post-cued.

manipulated within list via study time and both study and test tasks manipulated between lists (Figure 2). Participants in this experiment studied and were tested on 6 lists of 18 word pairs, each study list was followed by a distractor task and test of recall. Encoding strength of each item was manipulated via study time. On 6 of the 18 pairs, both members of the pair were presented together for 3.5 s, with one member presented on the left half of the screen and the other on the right half. For 12 of the pairs, one member of the pair (half cue, half target) was presented for 3.5 s on its own in the center of the screen, followed by the presentation of the pair for 3.5 s. Each 3.5 s presentation was separated by a 0.2 s blank screen. The order of the word pairs on the study list and whether the cue or target word was presented on the left half or right

half of the screen was randomized. On half of the lists, participants were instructed to study the two words together by placing them into a single sentence. On the other half, participants were instructed to study the two words separately, by placing them into separate sentences. Participants were not told to enter any responses during the study phase, merely place the words in separate sentences or the same sentence in their head. Study list order was randomized.

Two of the six lists were tested with paired associates cued recall, two were tested with extralist cued recall, and two were tested hybrid cued recall. For instance, if a participant studied WINDOW and CONSOLE together in a pair and was later cued to retrieve CONSOLE, hybrid cued recall provided the word DOOR as a test stimulus with instructions to recall the word studied alongside something similar to the stimulus (which, in this case, is WINDOW). One may compare this to paired associates cued recall, where WINDOW would be the test cue, and extralist cued recall, where XBOX could be the test cue because it is related to CONSOLE. The order of the test tasks was randomized.

Results

We consider each of the listed predictions in turn. Note that for these primary analyses we only consider correct response rates. We present an exploratory analysis of intrusions and response failures later. See Figure 3 for means of each response type in each condition. Appendix B includes a secondary analysis of output interference by task which not directly relevant to the question at hand.

Analysis of Experimental Manipulations

Beginning with a 2 x (study separately vs. together) x 3 (none strong vs. cue/target strong vs. untested item strong) Bayesian repeated measures ANOVA of the extralist cued recall correct responses, we fond that study task had no effect on correct response rates, $F(2, 188) < 1$,

$BF_{inclusion}$ = 0.050. The item encoding strength manipulation did affect performance, $F(2, 376) = 11.9$, $p < .001$, $BF_{inclusion} = 120$. Additional cue/target study increased correct response rates versus no additional study, Bonferroni-corrected post hoc $t = 3.46$, $p < .001$, $BF_{10} = 3170$, and versus study of the untested item, Bonferroni-corrected post hoc $t = 3.61$, $p = .001$ $BF_{10} = 20.6$. Correct response rates when the untested item received additional study were the same as those for when no item was given additional study, post hoc $t = 1.38$, $p = .501$, $BF_{10} = 0.13$. Item strength and study task did not interact, $F(2, 376) < 1$, $BF_{inclusion} = 0.007$.

We now turn to the paired associates and hybrid cued recall data. In the 2 (paired associates vs hybrid cued recall) x 2 x (study separately vs together) x 3 (none strong vs. cue strong vs. target strong) Bayesian repeated measures ANOVA, we found that studying pairs together led to greater correct response rates than studying them separately, $F(1, 188) = 30.4$, $p < .001$, $BF_{inclusion} = 4.6*10^{14}$. The ANOVA had trouble parsing the item effects in the gestalt, finding a significant main effect, $F(2, 376) = 9.04$, $p < .001$, but finding little evidence for or against an effect, $BF_{inclusion} = 0.61$. The Bayesian component of the ANOVA discounts precision obtained by sampling large numbers of participants but with few observations per participant, as we had in this experiment. Breaking down by effect of cue and target strength, we found evidence that the strong conditions were each associated with more correct responses than the condition where no words received additional study (using the Bonferroni correction, cues: post hoc $t = 4.16$, $p < .001$, $BF_{10} = 120$; targets: post hoc $t = 3.25$, $p = .004$, $BF_{10} = 4.5$).

Finally, we want to know the overall level of performance by test task. For this, we combined the three tests in a 3 (extralist vs. paired associates vs. hybrid cued recall) x 2 (study separately vs. together) x 3 (none strong vs. cue or cue/target strong vs. target or untested item strong) repeated measures ANOVA and considered the main effect of task. We found that

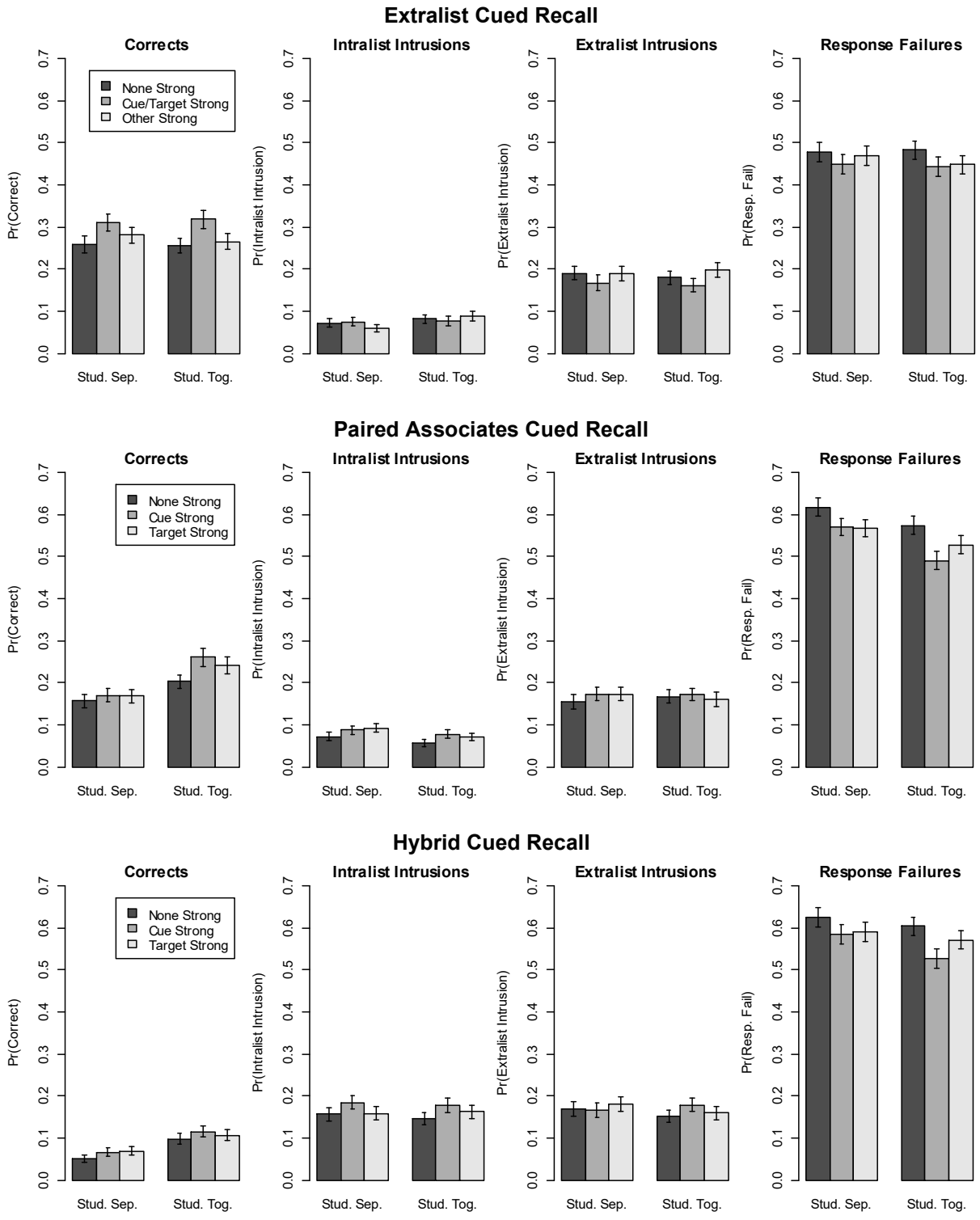

Figure 3. Bar graph of means in the task. Error bars: ± 1 SEM (SD/sqrt(N) method)

extralist cued recall had the most correct responses overall, paired associates cued recall in the middle, and hybrid cued recall the fewest, $F(2, 376) = 117$, $p < .001$, $BF_{inclusion} = 3.2*10^{15}$. Bonferroni-corrected post-hoc *t* tests showed that hybrid cued recall had the fewest correct responses (vs paired associates: post hoc $t = 17.7$, $p < .001$, $BF_{10} = 1.8*10^{58}$, $d = 1.29$; vs extralist: post hoc $t = 23.7$, $p < .001$, $BF_{10} = 1.2*10^{97}$, $d = 1.72$). For point of comparison, the difference between extralist and paired associates cued recall correct response rates (post hoc $t = 9.06$, $p < .001$, $BF_{10} = 2.8*10^{9}$, $d = 0.66$) was a large effect with decisive evidence, but not nearly as large or decisive as the comparisons to hybrid cued recall.

Summary. Overall, we found that the study task improved correct response rates in the paired associates and hybrid cued recall tasks and had no effect on correct response rates in the extralist cued recall task, consistent with hypothesis 2. We also found the expected effects of cue and target strength in all three tasks, consistent with hypothesis 1, but in paired associates and hybrid cued recall these effects were small. Finally, we found strong evidence that hybrid cued recall yielded the lowest correct response rates of the three test tasks, consistent with hypothesis 3.

Differential Influence of Cue Strength and Study Task

Here, we present tests of the prediction that joint manipulations of cue strength and cue-target association strength, manipulated via study task, should produce joint monotonicity of paired associates cued recall and the hybrid cued recall correct response rates, but differential or selective influence when comparing extralist cued recall correct response rates against those from either task.

We tested these predictions with state-trace analysis. Our frequentist method was conjoint monotonic regression; we used the binomial analysis functions of the STACMR-R package

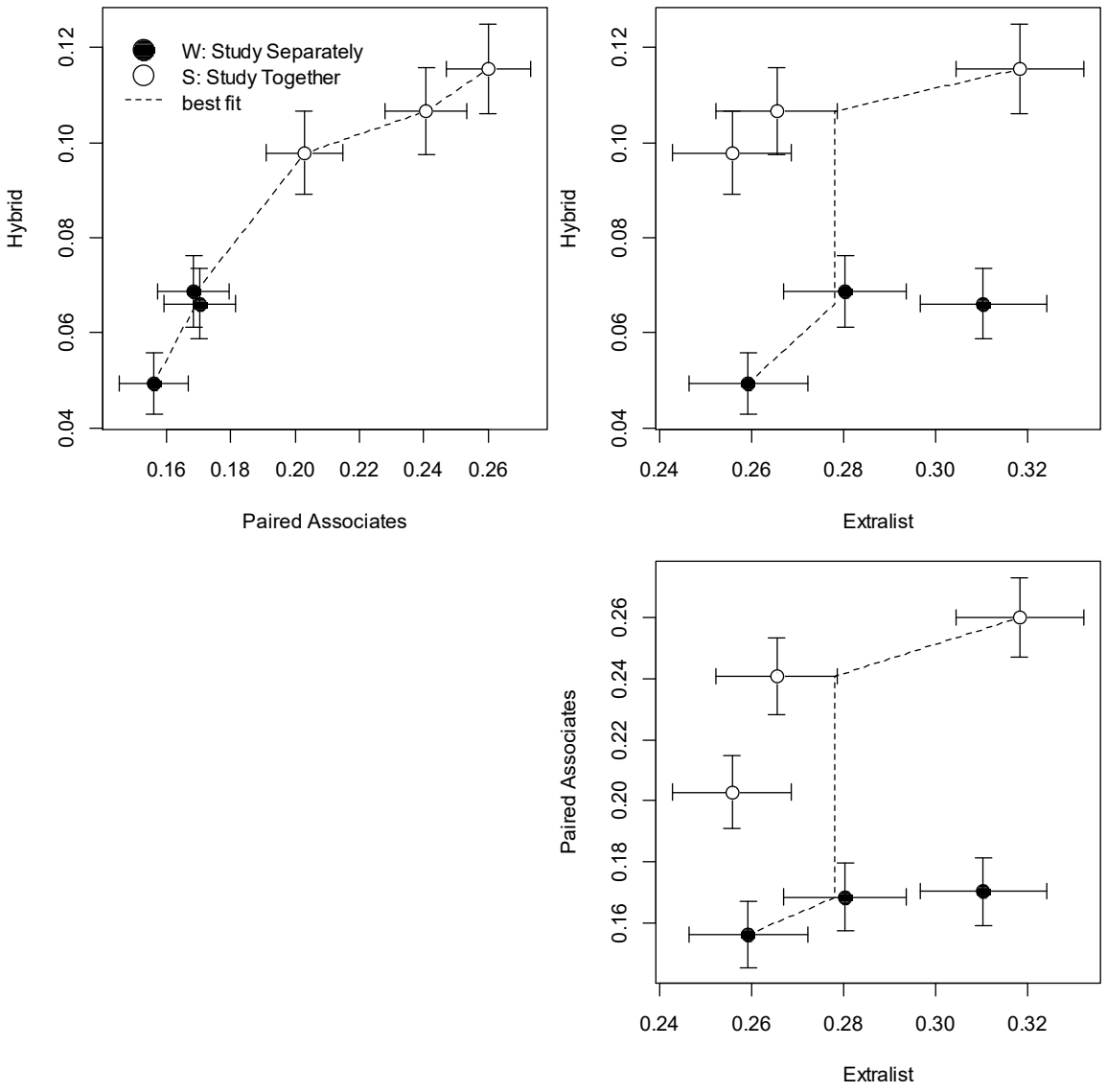

Figure 4. State-trace plot with study task and cue/target strength as independent variables and correct response rates in each task as dependent variables. Points and error bars give means +/- 1 SEM under binomial distribution assumptions. Dotted line is the best fitting monotonic regression using the STACMR-R package.

(Dunn & Kalish, 2018) with `nsample` = 10^4. Coupled monotonic regression tests a null hypothesis that two measures are jointly monotonic by bootstrap-resampling the data and comparing, for each resample, the goodness-of-fit of a best-fit monotonic function to one that need not be monotonic. The null hypothesis is rejected if the monotonic function fits best in less than 5% of the samples ($\alpha = .05$).

In all, we failed to reject the null hypothesis of joint monotonicity for paired associates vs hybrid cued recall correct response rates, $p = .84$. Comparing extralist cued recall to either task, we rejected the null hypothesis of joint monotonicity (vs. paired associates: $p = .01$; vs. hybrid: $p = .01$). See Figure 4.

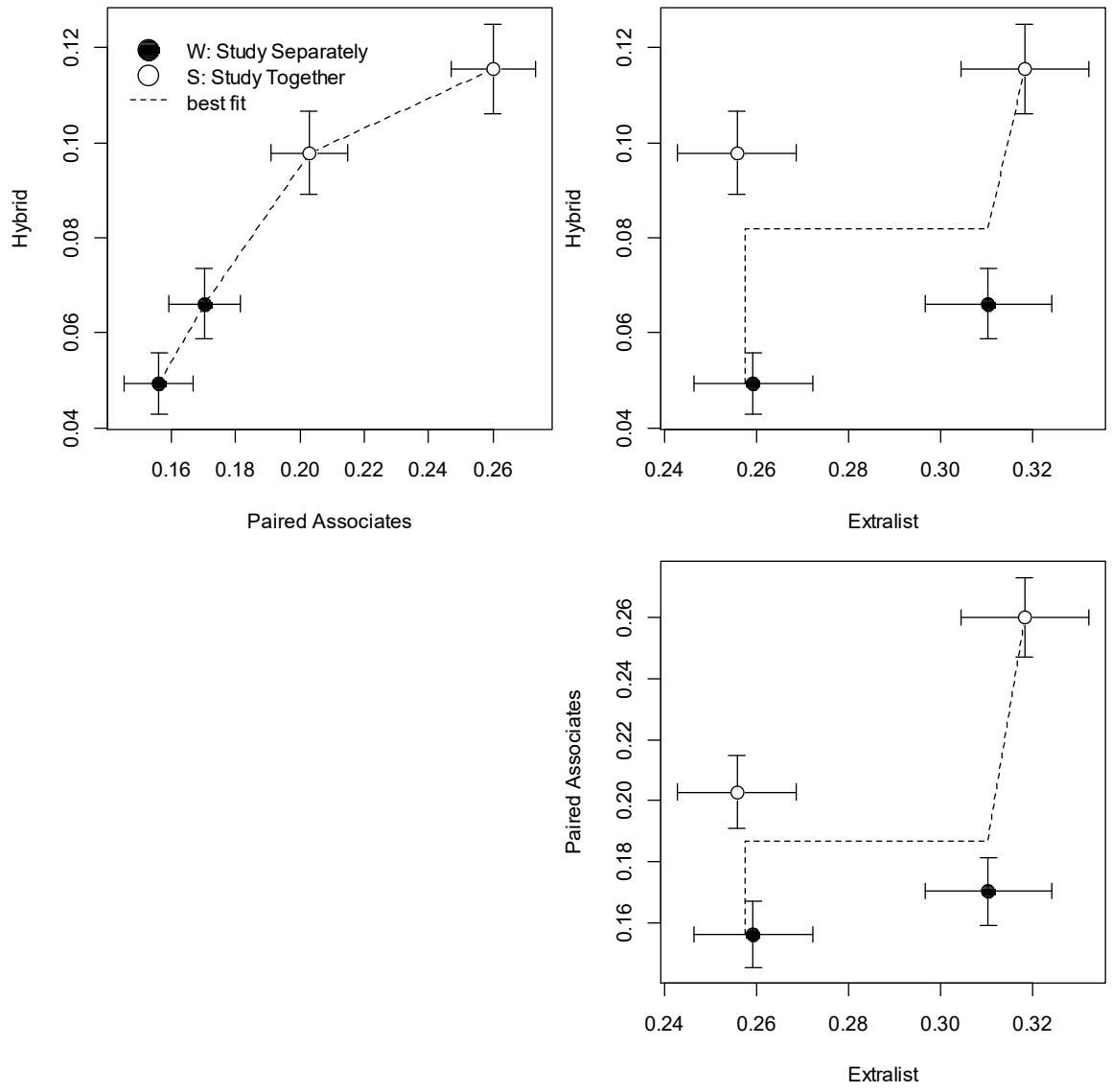

Figure 5. State-trace plot with study task and cue strength as independent variables and correct response rates in each task as dependent variables. Points and error bars give means +/- 1 SEM under binomial distribution assumptions. The dotted line is the best fitting monotonic regression using the STACMR-R package.

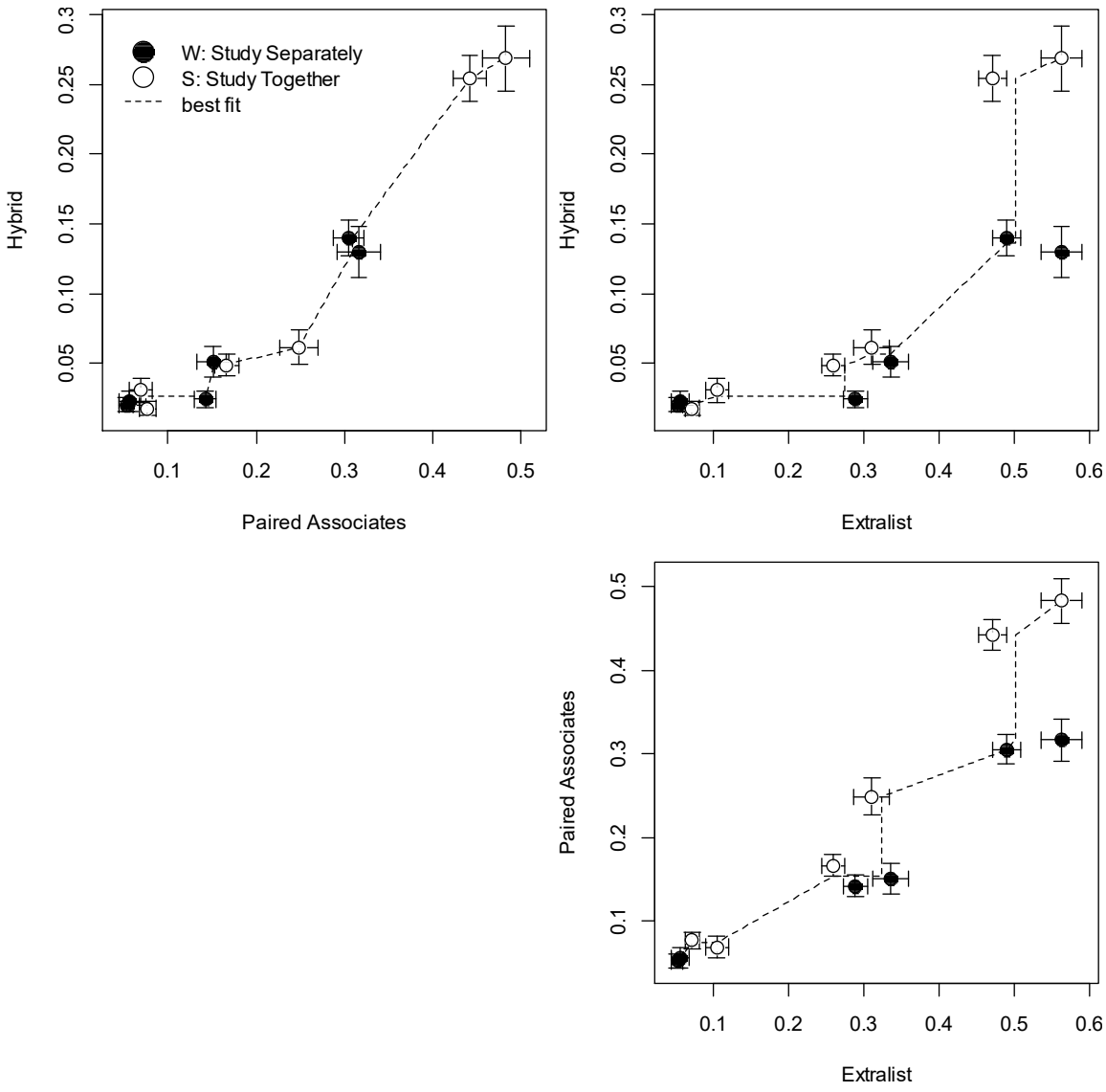

Figure 6. State-trace plot with study task and cue strength as independent variables and correct response rates in each task as dependent variables, sweeping across tercile. Points and error bars give means +/- 1 SEM under binomial distribution assumptions. The dotted line is the best fitting monotonic regression using the STACMR-R package.

The prior analysis considered both cue and target manipulations, meaning that differential influence of target strength and some other manipulation could have caused the non-monotonicity, rather than differential influence of cue strength and study task. To ascertain whether differential influence could be attributed to just the manipulations of cue strength and study task, we reran these analyses without the condition where targets receive extra study time

(Figure 5). Comparing paired associates cued recall to hybrid cued recall, we failed to reject the null hypothesis of joint monotonicity, $p = 1.00$. When we compared extralist cued recall to either task, we rejected the null hypothesis of joint monotonicity (vs. paired associates cued recall: $p < .05$ vs. hybrid cued recall: $p = .01$).

Finally, to confirm that failure to reject the null hypothesis of joint monotonicity of paired associates and hybrid cued recall was not driven by the degree of separation imposed by the study task, we performed a tercile split over the average correct response rate of a participant across the whole experiment[6] (Figure 6). The conjoint monotonic regression model used here accounted for the partial order imposed by the tercile split. We found the same patterns as before and failed to reject the null hypothesis of joint monotonicity between paired associates and hybrid cued recall, $p = .73$, while rejecting the null for extralist cued recall versus paired associates cued recall, $p = .02$ and versus hybrid cued recall, $p = .02$.

Summary. In all, these findings are consistent with hypothesis 4 and both predictions from the two-sample framework. Differential influence of cue strength and study task was observed for extralist cued recall versus either of the other two tasks. Paired associates and hybrid cued recall were jointly monotonic under these same manipulations. Altogether, this means that paired associates and hybrid cued recall correct responses are produced via the same retrieval process, but a different process is (or combination of processes are) employed during extralist cued recall. This is consistent with the notion that paired associates and hybrid cued

[6] A failure to reject H_0 in state-trace analysis can happen when all values in one condition are larger than all values in another condition, irrespective of the presence or absence of differential influence. We therefore wish to create conditions where the "traces" of a manipulation under two conditions could potentially overlap, such that the smallest point in a condition with larger numbers may be smaller than, or to the bottom-left of, the largest point in a condition with smaller numbers. Conducting a tercile split, here, fits the bill.

recall use both stimulus-to-cue and cue-to-target sampling, while extralist cued recall uses only stimulus-to-cue sampling.

Bayesian Multinomial Model Analysis

Method. This analysis served to measure what sorts of errors result from the two sampling phases and recovery.

The model consisted of three stages (Figure 7): stimulus-to-cue sampling, cue-to-target sampling, and recovery. A correct response occurred if and only if there was success at all three stages. If failure occurred at any given stage, then the model existed with either an intralist intrusion, an extralist intrusion, or a response failure. Allowing m, c, and r to be the probability of success in stimulus-to-cue sampling, cue-to-target sampling, and recovery respectively, we can write:

$$\Pr(\text{correct}) = mcr \tag{6}$$

The intralist intrusion and extralist intrusion probabilities are:

$$\Pr(\text{intralist intrusion}) = (1-m)m' + m(1-c)c' + mc(1-r)r' \tag{7}$$

$$\Pr(\text{extralist intrusion}) = (1-m)m'' + m(1-c)c'' + mc(1-r)r'' \tag{8}$$

Where m', c', and r' are the probability of exiting the process with an intralist intrusion at the respective phase, given the correct path is left. Parameters m'', c'' and r'' do likewise for extralist intrusions. Response failure probabilities can be written in the same fashion and take up the remaining probability space.

The probability space for a phase was fit independently for each manipulation of that phase. In all, we fit four versions of stimulus-to-cue sampling reflecting the impact of both cue study time and the test task: one strong and one weak version for paired associates cued recall and one strong and weak each for extralist and hybrid cued recall, strength depending on cue

study time. We fit one strong and weak version of cue-to-target sampling used only by paired associates and hybrid cued recall, strength depending on the study task. Finally, we fit one strong and weak version of recovery used by all three tests, strength depending on target study time. Each parameter was used based on the relevant manipulation. For example, the "strong" version of the recovery parameter was used only when the target word receives additional study. The model assumed that cue-to-target sampling was automatically successful in extralist cued recall. Finally, because the cue and target word in extralist cued recall are the same word, we assumed that additional study of the cue word was associated with the strong version of both the stimulus-to-cue sampling parameter and recovery parameter. See Table 2 for a mapping of parameter to task.

Table 2

Versions of each correct processing parameter for each test task and study condition

		Parameter Version in Study Condition					
		Study Separately			Study Together		
Cued Recall Task	Parameter	none strong	cue strong	target/other strong	none strong	cue strong	target/other strong
Extralist	$m_{similar}$	weak	strong	weak	weak	strong	weak
	c	1	1	1	1	1	1
	r	weak	strong	strong	weak	strong	strong
Paired Associates	m_{same}	weak	strong	weak	weak	strong	weak
	c	weak	weak	weak	strong	strong	strong
	r	weak	weak	strong	weak	weak	strong
Hybrid	$m_{similar}$	weak	strong	weak	weak	strong	weak
	c	weak	weak	weak	strong	strong	strong
	r	weak	weak	strong	weak	weak	strong

Note. For extralist cued recall, $c = 1$; the phase is bypassed. Parameters for incorrect processing correspond in the same way.

We fit this model using jags in R and find the posterior distribution of the probability spaces.

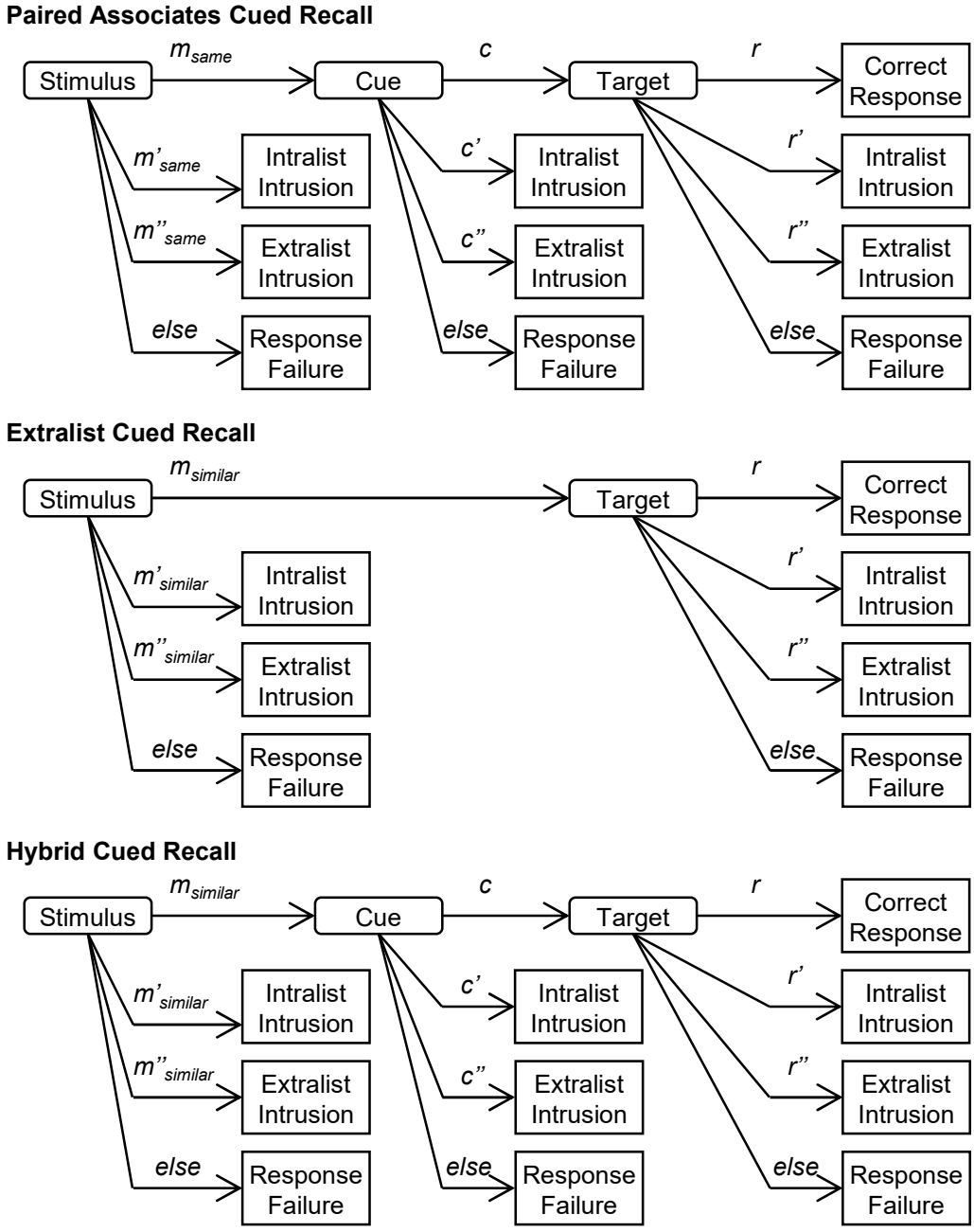

Figure 7. Schematic of the Bayesian multinomial model. $x + x' + x'' + else = 1.0$ for all parameter types.

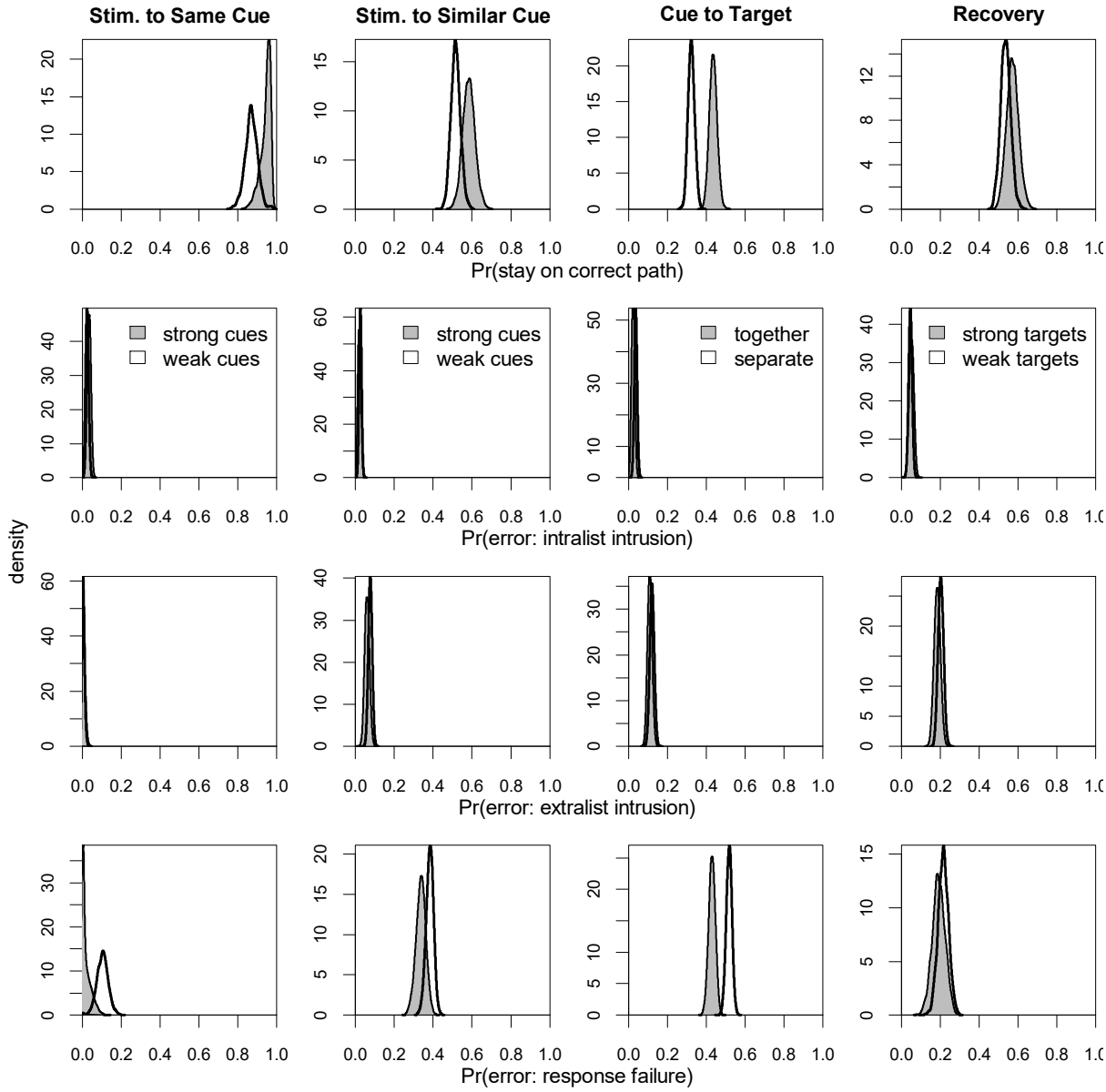

Figure 8. Distributions of the probability parameters from the multinomial model. Top row: correct processing. Second row: exit with an intralist intrusion. Third row: exit with an extralist intrusion. Bottom row: exit with a response failure.

Results.

Correct Processing at Each Phase. Stimulus-to-cue sampling for paired associates cued recall was highly accurate, C95%= [.88, .98] for strong cues C95% = [.80, .93] for weak cues. When the stimulus and cue represent similar words, as in extralist and hybrid cued recall, the rate

of success was much lower, C95% = [.52, .64] for strong cues and C95% = [.47, .56] for weak cues. Cue-to-target sampling was flagged by the model as the hardest phase with the lowest chance of success, C95% = [.40, .47] when words were studied together, C95% = [.29, .36] when studied separately. Recovery was, qualitatively, about as successful as stimulus-to-cue sampling in extralist and hybrid cued recall, C95% = [.51, .63] for strong targets and C95% = [.49, .59] for weak targets.

Table 3

95% Credibility Intervals for exit probabilities from the multinomial model

		C95% for Probability of Exiting at Phase With a(n):		
Phase	Study Condition	Intralist Intrusion	Extralist Intrusion	Response Failure
Stimulus-to-cue				
similar	weak cues	[.02, .03]	[.06, .09]	[.34, .42]
	strong cues	[.01, .03]	[.04, .08]	[.29, .38]
same	weak cues	[.01, .04]	[.00, .02]	[.04, .16]
	strong cues	[.02, .05]	[.00, .02]	[.00, .08]
Cue-to-target				
	study separately	[.03, .05]	[.10, .14]	[.49, .54]
	study together	[.01, .03]	[.09, .13]	[.40, .46]
Recovery				
	weak targets	[.03, .06]	[.18, .23]	[.16, .27]
	strong targets	[.03, .07]	[.16, .21]	[.13, .25]

Errors at Each Phase. We report here the qualitative findings, see Table 3 for 95% credibility intervals. First, perhaps reflecting the fact that the most common outputs in this experiment were response failures, the majority of errors during the sampling phases were response failures. Improvements in performance due to cue strength or study task manipulation tended to result in fewer response failures. Recovery differs according to the model: the most

common error during this phase was an extralist intrusion, followed by a response failure. Although response failures also happened due to recovery errors, they happened proportionally less often in this phase in comparison to the sampling phases.

Summary. In all, cue-to-target sampling was generally noisier than stimulus-to-cue sampling. The effect of manipulation on correct processing at each phase mirrored the findings of the ANOVA, which was expected given that manipulations of cue, association, and target selectively influenced stimulus-to-cue, cue-to-target, and recovery probabilities respectively. Generally, sampling phases were more likely to result in response failures. The recovery phase was more likely than other phases to exit with an extralist intrusion.

Discussion of Experiment 1

The outcomes of this experiment are consistent with the predictions of the two-sample framework. We observe differential influence of cue strength (cue strong vs. none strong) and study task on correct response rates in extralist cued recall versus the other two tasks, but not for paired associates cued recall versus hybrid cued recall correct response rates. While increased cue study time generally increased performance, study task only affected performance in paired associates and hybrid cued recall—not extralist cued recall.

While the two-sample framework can account for this data, the one-sample framework, where the relationship between stimulus and cue and between cue and target are amalgamated, runs into problems. Put simply, a single sampling process as described in Chapter 1 cannot jointly predict the pattern differential influence observed here and the relative performances of extralist, paired associates, and hybrid cued recall.

Dissociations of Extralist and Paired Associates Cued Recall

Paired associates cued recall performance is greater for high frequency words and words with large pre-experimental set sizes, while extralist cued recall performance is greater for low frequency words and words with small pre-experimental set sizes.

Word Frequency. Word frequency is, simply, a metric for how often a person is expected to experience a word over a lifetime. It is measured by (more or less) counting words from text corpora (e.g. Brysbaert & New, 2009; Kucera & Francis, 1967; Lund & Burgess, 1996). Effects of word frequency on memory performance in various memory tasks have long been observed: words spoken in noisy settings were easier to identify when of higher frequency (Howes, 1957) is an early example. Word frequency effects are observed throughout much of the episodic memory literature. High frequency words are associated with greater correct response rates in free recall, category cued recall, and serial recall. Conversely, participants are more successful at distinguishing studied from unstudied words in recognition tasks if those words are of lower frequency.

A typical account for this word frequency dissociation between free/category/serial recall and recognition tasks is that low frequency words are more distinct, while high frequency words are more available or accessible. To wit, availability/accessibility drives recall while distinctiveness drives item recognition, thus a dissociation along word frequency. SAM, for instance, states that representations of higher frequency words are more tightly bound to those of other studied words and the context of study-test. High frequency words therefore have greater retrievability and also experience less differentiation than low frequency words (Gillund & Shiffrin, 1984; Shiffrin et al., 1990). Adaptive Character of Thought – Rational (ACT-R) argues that high frequency words are more likely to be recalled because they are generally more likely

to be needed in any given circumstance than a low frequency word. This leads to worse performance in recognition memory for the same reasons. On the other end, low frequency words are so rarely needed that they are less likely to be endorsed as studied when they have not been studied, and more likely to be endorsed as studied if they were studied (Schooler & Anderson, 1997).

As in other recall tasks, high frequency words are associated with more correct recalls in paired associates cued recall. This pattern generally holds for manipulations of both cue and target frequency but is less robust for cue words (Criss, Aue, et al., 2011). In extralist cued recall, unlike other recall tasks, low frequency words receive the advantage: low frequency targets are more likely to be correctly recalled than high-frequency targets, with a very small or no effect observed for the word frequency of the test stimuli (Nelson & McEvoy, 2000).

Pre-experimental Set Size. The pre-experimental set size of a word is, simply, the number of other words that are semantically related to said word. This can be measured through analysis of text corpora, or (and this method is more valid for cued recall tasks) by behavioral outcomes of the free association task. In a free association task, a word's set size is measured both by the number of unique responses to the word when presented as a test stimulus and the number of test stimuli that elicit the word as a response. As with word frequency, there are task dissociations related to this metric. Words with few pre-experimental relationships hold an advantage when these relationships are used during retrieval, while words with many pre-experimental relationships hold an advantage when the critical relationships used at test are formed in the context of the experiment.

In extralist cued recall, smaller set sizes for test stimulus and for target are associated with greater correct response rates, a consistent finding across many studies (Nelson & Zhang,

2000) that is robust to other word property manipulations (Nelson et al., 1992). Set size derived from the free association task offers the most robust effect on extralist cued recall performance, nonetheless the effect is also observed when the set size is category size (e.g. McEvoy & Holley, 1990) or the number of extralist rhymes (e.g. McEvoy & Nelson, 1990; Nelson, Bajo, et al., 1987; Nelson & Friedrich, 1980) . It shares this property with word fragment cued recall for the set size of the word fragment (e.g. Nelson, Bajo, et al., 1987; Nelson & McEvoy, 1979, 1984), perceptual identification (Nelson et al., 1984), and, to a lesser extent, item recognition (Nelson, Cañas, et al., 1987).

In paired associates cued recall, however, the opposite finding occurs. Provided the cue and target words are unrelated, larger set sizes promote an increased correct response rate (Nelson et al., 1990). If the cue and target are related, a small set size advantage appears (Nelson et al., 1992) but it is noticeably smaller than in extralist cued recall. This suggests two competing effects: a small set size advantage in cases where the test stimulus is semantically related to the target word, and a large set size advantage when the target must be found by utilizing the cue-target association. Paired associates cued recall shares this advantage with target words in free recall (e.g. Nelson, McEvoy, & Janczura, 1992, as cited by Nelson et al., 1992).

In other words, it appears that accessing the pre-experimental relationships between items advantages words with few such relationships, whereas accessing relationships defined within the context of the experiment advantages words with many pre-experimental relationships. This closely parallels the word frequency findings, where tasks that use the representation of the item advantage low-frequency words experienced less often and tasks that use the learned associations between items advantage high-frequency words. The two factors are somewhat correlated but

both word frequency and set size effects appear after statistically controlling for the other (Nelson & McEvoy, 2000).

Implications. Both word frequency and pre-experimental set size cause opposite effects in extralist and paired associates cued recall. In paired associates cued recall, high-frequency words and words with large pre-experimental set sizes are advantaged, while in extralist cued recall, low-frequency words and words with small pre-experimental set sizes are advantaged.

As previously mentioned, this dissociation causes problems for a one-sample process. However, it can be accounted for under the two-sample framework: stimulus-to-cue sampling is better when the cues are low frequency and/or have small set sizes, while cue-to-target sampling is better when the cues and targets are of high frequency and/or have large set sizes. Extralist cued recall relies on the relationship between the test stimulus and the cue, but not on the knowledge that two words were studied together. Paired associates cued recall, instead, relies on both the stimulus-cue relationship and the cue-target association. High-frequency words and words with large set sizes have an easier time forming associations with other words in the experimental context than low-frequency words or words with smaller set sizes. Low frequency words and words with small set sizes, on the other hand, are easier to find given a test stimulus.

Multinomial Modeling of Recall

In the multinomial model applied to this experiment, strengthening cues, targets, and cue-target associations were each associated with increases in their related correct processing probabilities. Overall, manipulations of the cue-target association, by adjusting the study task, had the greatest impact on the model parameters. In paired associates and hybrid cued recall, the model reports that most of the errors take place during the cue-to-target sampling phase. In contrast, the model reports that stimulus-to-cue sampling, when the stimulus and cue are the

same word, is quite accurate. Stimulus-to-cue sampling in extralist and hybrid cued recall is less accurate, reflecting the fact that in these two test tasks, the cue and test stimulus are related but nonetheless different words.

The use of multinomial models to analyze recall data has an extensive literature, including as process dissociation models (e.g. Jacoby, 1998), but has not been applied in the way offered here to analyze the process of retrieval. Process dissociation models suggest that response probabilities are governed by dominant and non-dominant processes, with the outputs from the dominant process receiving priority over the non-dominant. Multinomial models have also been used to attempt to distinguish between storage and retrieval effects. Reifer and Rouder (1992) considered free- and cued recall-tested memory for common versus bizarre sentences with a multinomial model for storage and retrieval of said sentences.

Most relevant might be Ross and Bower's (1981) analysis of the properties of associations in episodic memory. They compared two multinomial models (along with a third "fragment" model, which fit the data poorly and we therefore skip over): horizontal models and schema models. The models were designed for a task where participants studied sets of noun tetrads/quintets with a common schema and were later cued to recall the set given one, two, or three members as test stimuli. Experiment 1 used tetrads and strength of association was manipulated by informing or not informing participants of the schema. Experiment 2 used quintets and all participants were aware of the schema. Experiment 3 returned to tetrads and used a sequential cueing procedure.

The horizontal model quantifies cued recall with independent all-or-none storage probabilities for the individual tetrad members (probability p) and the possible unidirectional links between them (probability θ). Successful recall of a whole tetrad/quintet requires encoding

of all 4/5 items and sufficient unidirectional links to path from the test stimuli to all target members. The schema model instead assumes the presence of a schema—a higher-order representation of the tetrad—which has probabilistic all-or-none unidirectional links to (probability r) and from (probability a) the tetrad members. Successful cued recall requires links from the cue words to the schema and from the schema to the targets.

These two models share some important similarities to the multinomial framework proposed here, but neither offers useful evidence in either direction for or against more than one sampling process. The more successful schema model suggests that the probability of paired associates cued recall (for a pair of two items) is the product of two probabilities: that of a link from cue to schema and that of a link from schema to target: $\Pr(\text{recall}) = ar$. This can distinguish between cue and target strength but conflates schema-to-target and successful recovery, in effect. The horizontal model is the more similar of the two to our two-sample framework: outputting a target requires successful encoding of the cue, the target, and the link from cue to target—three probabilities. However, by experimental design, the cue and target are studied for the same amount of time hence encoding probabilities are equal, $\Pr(\text{recall}) = p^2\theta$, and only the strength of the association was varied. Therefore, neither model suggests more than two processes, whereas our full retrieval model suggests three: two for sampling and one for recovery.

Implications for Sample-Recovery Models

The two-sample framework includes assumptions that have interesting implications for sample-recovery models. Foremost amongst these, of course, is that there are two stages of sampling in paired associates cud recall. The implications of this point for models are clear.

Many theories of memory divide roughly along the lines of using knowledge from the study list (e.g. context) versus using pre-existing knowledge (e.g. semantic relations) to aid in retrieval. On the one hand, models such as SAM and the Temporal Context Model (TCM; Howard & Kahana, 2002) argue that finding the target requires learning strong associations between representations of individual words during study. In SAM, these are direct links between the cue and target word, in TCM this is increased knowledge that the two words were experienced at the same time. On the other hand, models such as REM and PIER argue that a strong semantic relationship or a high degree of similarity between the test stimulus and the cue word drives performance. In REM (Diller et al., 2001), the more similar to the test stimulus the cue trace is, the more likely the trace is to be sampled. In PIER, the best predictors of extralist cued recall performance are the strengths of the forward and backward semantic associations between test stimulus and target word (Nelson et al., 2013).

That we can separate stimulus-to-cue sampling from cue-to-target sampling suggests a way forward in jointly accounting for both the association-driven and semantic-driven approaches of the models. Stimulus-to-cue sampling should rely on the semantic relationships and degree of similarity between the test stimulus and the contents of the long-term episodic memory store. Cue-to-target sampling relies on the relationship between cue and target learned during the study list.

The models that have seen more success in paired associates cued recall as well as free recall, namely SAM and the TCM class, focus on the newly learned associations including between cue and target. The Bayesian multinomial model analysis suggests that this is because cue-to-target sampling is where most errors occur in the task. These models' focus on accounting for the factors that determine what trace is selected given another has therefore

proven fruitful when accounting for most recall tasks. Likewise, that stimulus-to-cue sampling is rather successful in paired associates cued recall has allowed the models to abstract this phase without much loss in ability to account for phenomena in paired associates cued recall.

Models focusing on similarity and semantic associations—REM and PIER—have been most successful in tasks such as item recognition and extralist cued recall because the primary source of noise in those tasks comes from the relationship between the test stimulus and the contents of the long-term episodic memory store. It is more difficult to find the representation of the cue word if the test stimulus is merely related, and it can be difficult to decide whether or not a test stimulus has been studied or not when there are many studied items, each possessing some degree of relation to the test stimulus. Overall, this suggests that adopting the two-sampling framework may be a way forward to integrate these various theories of memory.

Chapter 3: Recovery

Recovery is the process by which the information extracted or selected during sampling is transformed into a response. In a broad view, this means transforming an episodic representation of an event into the lexical-semantic space, then transforming that lexical-semantic representation into either a written or gestured or verbal form. From that perspective, recovery is a complicated process which recruits memory, language, and mechanical skills and thus transcends the episodic memory literature. However, sample-recovery models have relatively simple assumptions about what governs the success of recovery. Recovery, in our computational models, uses the information gained from sampling and the lexicon to decide whether and what to output. Representations of the target word with more and more accurate information are more likely to result in the target word being output.

In computational models, the algorithm by which this occurs varies considerably. In SAM, the strength of an item is represented by stronger associations to other items. If a representation of a word is sampled, then either that word will be output or recovery will fail—if the representation of "house" is sampled, then the model will only ever output "house" or nothing at all. The probability of output is a function of the associations that item has to the test stimulus and text context. Thus, only information contained within the to-be-recovered item will be used in the output decision. Most instantiations of REM (Lehman & Malmberg, 2013; Shiffrin & Steyvers, 1998) make similar assumptions. The sampled item is either recovered and the word it represents is output, or recovery fails; the probability of recovering is a function of the proportion of information stored in the sampled item. However, the variant by Diller, Nobel, and Shiffrin (2001) allows for recovery to sometimes generate a word that is not the sampled item; the probability depends on the evidence that the trace is a representation of the target.

More elaborate models of recovery impose various multinomial processing trees on the problem. The prototypical example of this was implemented by Schweickert (1993). In this model, a to-be-recovered trace is either intact or degraded. If degraded, there is a chance that the trace can be reconstructed. Reconstruction may be based on lexical or phonological properties. In either case, the underlying probabilities of the tree are probabilities about the status of the sampled information or the capacity to transform that information into something recoverable.

Yet more elaborate approaches essentially use the sampled information as a baseline to "sample" the lexicon: given the information obtained during sampling, what lexical entry is most similar? This is the implementation used by TODAM (Murdock, 1982). In that implementation, the dot-product of the extracted information and the possible responses is compared. Whatever lexical entry has the largest dot-product is output if the dot-product is larger than some set criterion. Although few other memory models have implemented recovery in this way, it would be relatively straightforward to implement this approach in the other models we have discussed.

The most elaborate approaches attempt to model the process of transforming an episodic representation into a lexical entry; these models generally impose a recursive algorithm to achieve this task. Multiple algorithms have been proposed, including the Brain-State-in-a-Box model (J. A. Anderson et al., 1977; proposed for the Matrix Model, Humphreys et al., 1989) and Hopfield networks (Hopfield, 1982; proposed for TODAM, Murdock, 1982). These algorithms take as input the information extracted during sampling and each lexical entry, which serve as attractors. The algorithm transforms the extracted information over time until it converges on one of the lexical entries. Whatever entry the algorithm converged upon is the response given. Again, however, the other contents encoded during study do not influence this outcome, outside of its influence on what information was extracted during sampling.

Verbal Theories of Recovery

The metamemory literature has suggested multiple recovery mechanisms which have only been implemented in memory models to varying degrees. These generally center around the concept that participants are in one way or another able to evaluate the accuracy of their responses and withhold information depending on that evaluation and task demands. In cued recall, the task is to respond with a word; the relevant metamemory decision would be whether or not to respond with the recovered word. We thus focus on considering what information is thought to be used in making these decisions and how it may apply to computational memory models.

Broadly, the metamemory literature suggests that a response is more likely to be given if the to-be-generated word is more relevant to the question at hand and if it is more correct. Naturally, this includes direct assessments of recovery accuracy (Koriat & Goldsmith, 1996), which is already implemented in TODAM (Murdock, 1982). However, other sources of information are also thought to contribute. Some argue that traces that are vivid or distinct, or which have stronger source representations are less likely to be suppressed (e.g. Schacter et al., 1999). Implementations of this in memory models might use the to-be-generated item's binding to the context of study-test as a proxy for its accuracy. This is in line with SAM's formulation of recovery, in which binding to context is used both to determine what to sample and whether to recover. Alternatively, this verbal account may be construed as testing the to-be-generated item's familiarity. Generally, checks of familiarity are implemented in the context of the recognition task: a global match of a test stimulus to the contents of the long-term store is performed and the probability of calling said item is familiar is weaned to the degree of global match.

Some suggest that information surrounding the circumstances of recovery may influence the decision as to whether or not to recover, on the basis that these circumstances are predictive of response accuracy. For instance, if recall takes longer than expected this may be used as a basis to suppress the response (T. O. Nelson & Narens, 1980). Alternatively, participants may be more likely to suppress a response if they generally feel they are bad at the task (Perfect, 2004). These factors do not map cleanly onto the kinds of information included in our models; we do not consider this factor further.

Some theories suggest that recollections of other items can be used as a basis to suppress information (Brainerd et al., 2003). For instance, if a proposed response is similar to another recalled item on a list of unrelated items, that response may be suppressed. Yet others suggest that recall of an item may be suppressed if other, similar items, have already been recalled (M. C. Anderson et al., 1994). A similar mechanism exists within SAM, in that a to-be-recovered item is not recovered if the memory system sees that it had already been output during test. However current evidence suggests that this is based on information local to the to-be-recovered information and thus similar responses would not influence this form of suppression (Wilson et al., 2020).

Finally, responses may be suppressed if they are not relevant. Participants pay attention to task instructions and suppress responses that do not follow instructions. Assessments of relevance necessarily require semantic knowledge alongside knowing that the word was studied. In a category cued free recall task, one must know what four-legged animals are in order to respond with only the studied four-legged animals and not the other studied words. Extant models already consider this point. It is, for instance, a considered component of category-cued

free recall in SAM: if the to-be-recovered item is not found to be a member of the cued-for category then the response is suppressed (Gillund & Shiffrin, 1981).

A general theme we see here is that, with the exception of the familiarity check, the only information learned during the experiment that is used during recovery is the information related to the to-be-recovered word itself: time to recover; relevance; distinctiveness; and so on. Anderson's (1994) approach technically isn't even a recovery-based mechanism; the core of his argument is that information is forgotten. Other possible sources of information, like a participant's judgement of how good they are at a task, are formed prior to learning the study list. Thus, the verbal accounts of recovery are consistent with the models in that other items on the study list should not impact recovery. This general point will be tested in Experiment 2.

Experiment 2

The broad goal of this experiment is to assess recovery. In Chapter 2, we found evidence that sampling is composed of stimulus-to-cue and cue-to-target sampling. In the context of paired associates cued recall, the memory system first conducts stimulus-to-cue sampling and identifies the trace it believes to be the representation of the test stimulus; this should be the cue trace. The system then conducts cue-to-target sampling and attempts to determine what trace represents the item that was studied alongside the cue; this should be the target trace. Recovery then transforms that trace into a response.

The extant models of recall make different predictions about how the confusability of the cue and the target impact cued recall performance, stemming from the differences between stimulus-to-cue sampling and recovery. In stimulus-to-cue sampling, the traces in the study list compete to be sampled. Therefore, the representation of a cue word that confusable with another word on the study list is less likely to be sampled, resulting in worse paired associates cued recall

performance. In recovery, only the sampled trace is compared to the lexical-semantic store to determine what and whether to give a response; the other studied items play no role. Therefore, outside of effects of encoding, the presence or absence of a similar target is predicted to not influence paired associates cued recall performance.

A stimulus-to-cue sampling process identifies the trace that best matches the test stimulus. In this, two factors are critical in determining whether the correct trace is sampled, namely the distinctiveness or *confusability* of the cue trace and its memory strength. This is exemplified by the two models of cued recall that effectively use stimulus-to-cue sampling: REM and PIER. In REM, additional study of an item leads to more information about the item being encoded in its representative trace. As more information about the cue is encoded, the stronger the representation of the cue becomes and the more likely it is to be sampled upon presentation of the test stimulus. However, if another trace is similar, by virtue of sharing many features with the cue trace, then that other trace may be confused for the cue and sampled instead. In PIER, this is modeled at a higher level. Representations of cues that are studied longer are more likely to be sampled. If the cue has a strong semantic association to another studied item, as measured by the free association task, then the trace representing the other word is more likely to be sampled upon presentation of the test stimulus. Both models make two similar predictions. First, that increased study of the cue should increase the accuracy of stimulus-to-cue sampling. Second, that the presence of an item like the cue on the study list should decrease performance because they are confusable. Experiment 1 offered a test of the former. The latter prediction concerning similarity will be tested here.

In stimulus-to-cue and cue-to-target sampling phases, the traces compete to be sampled and so what is sampled depends upon the entire representation of, or a *global match* to, the study

list. In contrast, recovery is assumed to be based on a *local match* of the sampled information. In other words, although the global match affects what information enters the recovery process, it is not a factor in the recovery process *per se*. The non-reliance on a global match during recovery leads to the prediction that manipulating the content of the rest of the list will have no influence on the outcome of the recovery process. Our critical manipulation in Experiment 2 will therefore be the confusability of the cue and target words.

Methods

Participants

$N = 78$ undergraduate students at Syracuse University, all native English speakers, completed this experiment.

Materials

Words were drawn from the same pool of words used for Experiment 1.

Design & Procedure

All participants studied and were tested on six lists of 28 word pairs. For each list, 16 of the 28 pairs were later tested with paired associates cued recall. The remaining 12 served as "lures" and each lure was related to one tested pair. The nature of this relationship was determined by whether and what items in the lure pair were confusable with the tested pair. A cue in a tested pair could be *confusable* or *distinct* from a lure. Here, two confusable words are semantically related, in other words they have forward and backwards strengths between .14 and .30 as measured in the USF free association norms. For distinct words, there is no semantically related word in the lure. Likewise, targets could be confusable or distinct from a lure. Each of the 16 tested pairs contained a confusable cue, a confusable target, both, or neither. In this way

the experiment is a 2 (confusable cue vs. distinct cue) x 2 (confusable target vs. distinct target) x 2 (short lags vs long lags) within-by-between design (Figure 9).

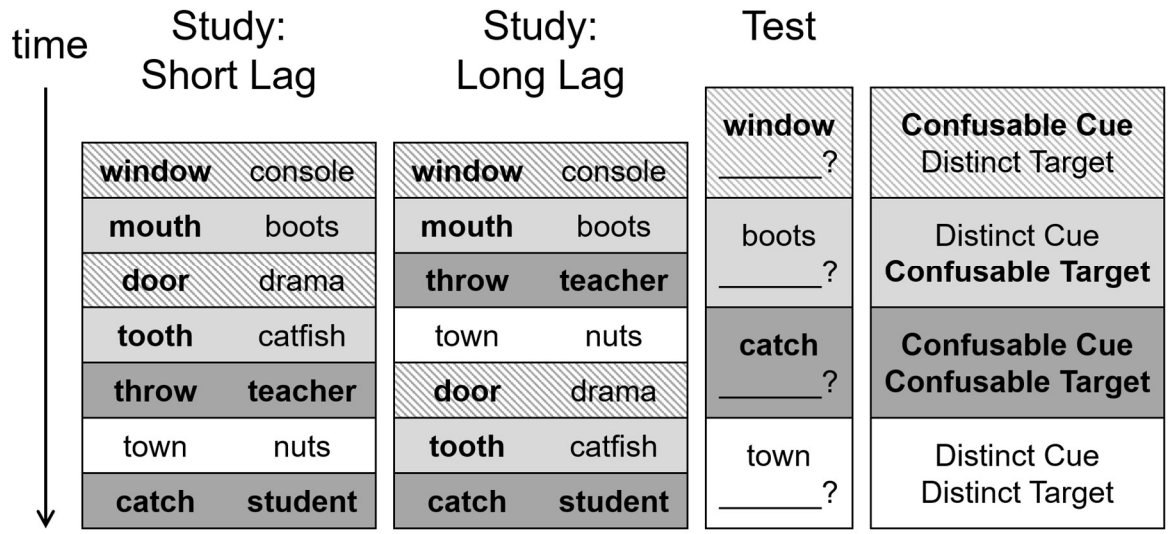

Figure 9. Condition schematic for Experiment 2. All pairs were presented in the same font and background color. Bolding and shading here are used respectively to highlight confusable words and the different within-list conditions. Study lists include 4 tested pairs per condition. Participants were uninformed of what words were cues or targets.

For one group of participants ($N = 40$), the lure pairs and the tested pairs were arranged such that they would be studied on opposite halves of the study list, called the long lag condition. For the other group of participants ($N = 38$), the lure and tested pair were studied on the same half of the list, called the short lag condition[7]. Study order was otherwise randomized. On each study list, each pair was presented for 2.5 s each with a 0.5 s interstimulus interval. Participants were instructed to study the pairs by placing both words of the presented pair into the same sentence. Each study list was immediately followed by a 30 s running addition distractor task followed by a test of paired associates cued recall. Participants were not informed of the study or test structure; they were informed that would need to study the words for a later test of memory,

[7] Note that the participants in short lags group were tested before those in the long lags group. Because this manipulation ultimately does not impact the pattern of data in a meaningful way and because it is secondary to the point, we treat the two groups as conditions in a single experiment rather than two variations of the same. We report the effect of group only when there is one.

and that they would not be tested on a study list again. Which study pair member appeared on the left vs right and the order in which pairs were tested were randomized.

Results

We analyze the data in a 2 (distinct vs confusable cue) x 2 (distinct vs confusable target) x 2 (short vs long lags) repeated measures ANOVA, with Bayes factors computed in JASP with standard parameters (r_{fixed} = 0.5, r_{random} = 1.0, $r_{covariates}$ = 0.354) using matched model effects as evidence. We consider correct responses, intrusions, and response failures (Figure 10). For a breakdown of what kinds of intrusions participants make, e.g. are their incorrect responses from the lure pair and if so was it the cue's lure or target's lure, see Figure 11.

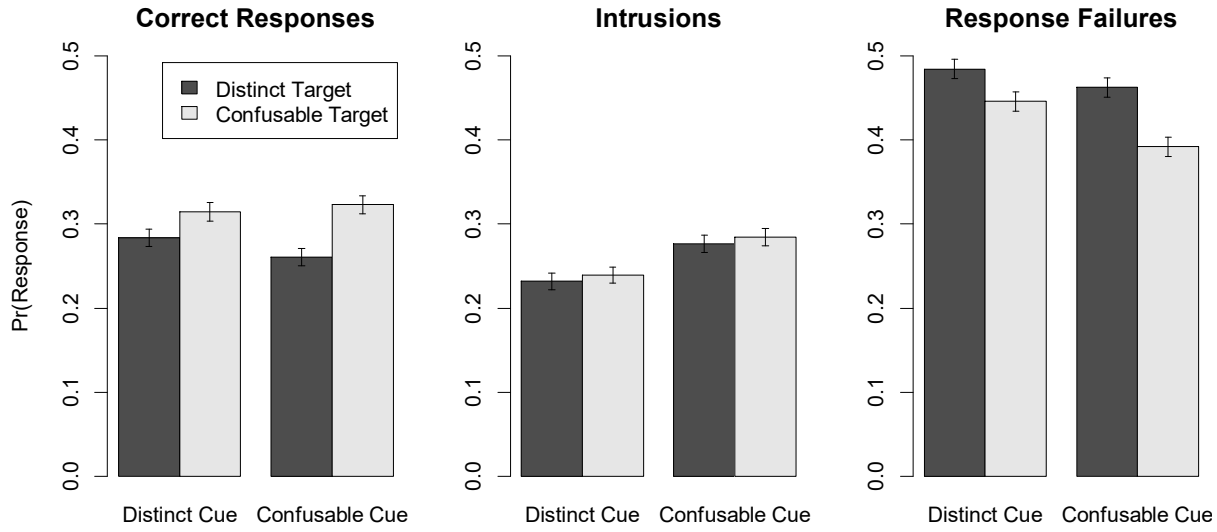

Figure 10. Response rates by condition. Means ± 1 SEM (SD/sqrt(N)).

More correct responses happen when the target word is confusable than not, $F(1, 76) = 20.03$, $p < .001$, $BF_{Inclusion} = 3000$. Whether the cue is distinct has no impact on correct response rates, $F(1, 76) < 1$, $BF_{Inclusion} = 0.16$. No interactions were observed (cue-by-target: $BF_{Inclusion} = 0.55$).

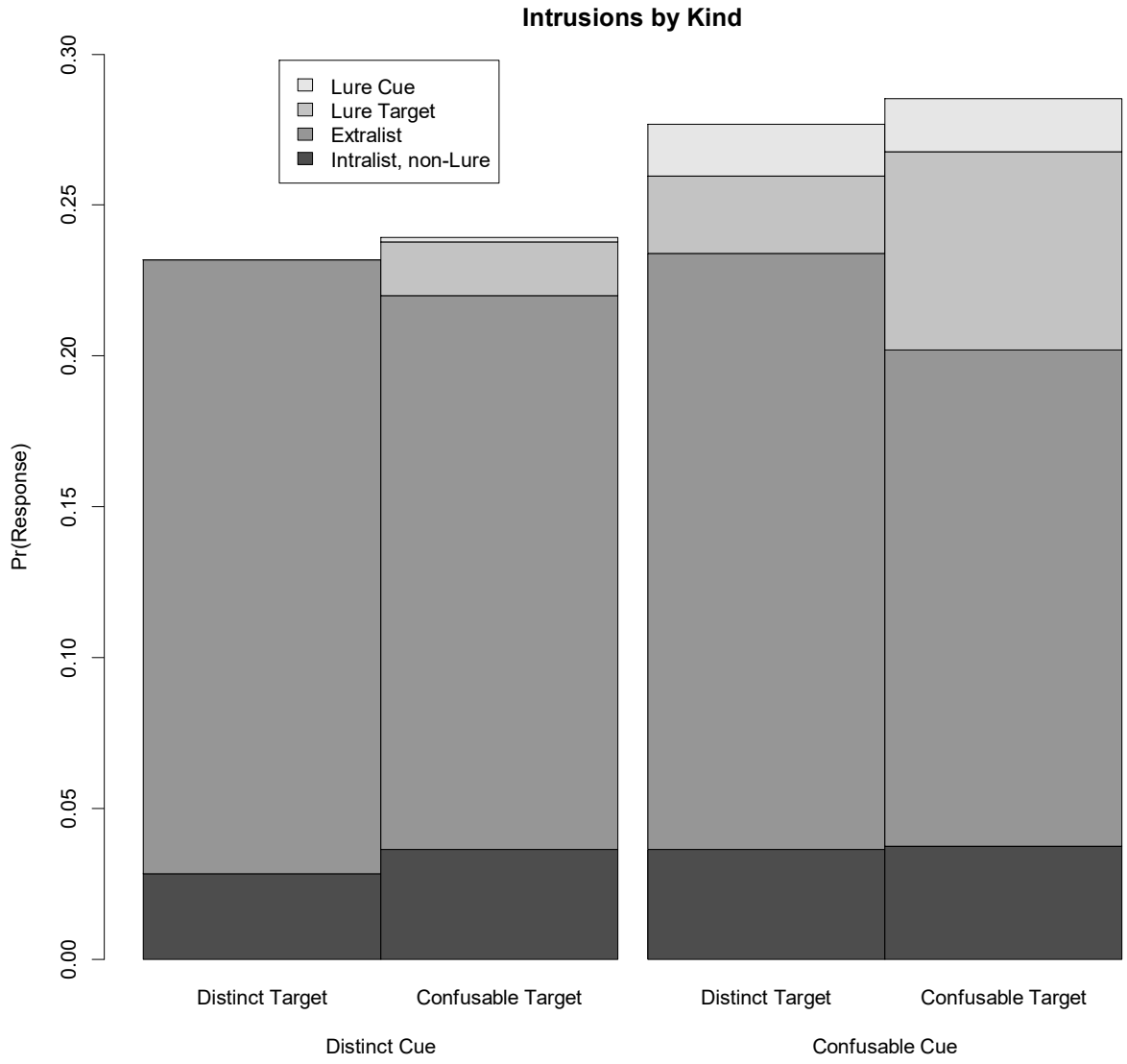

Figure 11. Intrusions rates broken down by kind of intrusion. Lure Cue: intrude with the cue word from the similar pair. Lure Target: intrude with the target from the similar pair. Extralist: responses outside the study list. Intralist, non-Lure: intruding responses from studied pairs other than the lure.

The overall level of intrusions is higher when the cue word is confusable than not, $F(1, 76) = 24.03$, $p < .001$, $BF_{Inclusion} = 4.4*10^4$. Target confusability had no impact on overall intrusion rates, $F(1, 76) < 1$, $BF_{Inclusion} = 0.18$, nor was a cue-by-target interaction observed, $F(1, 76) < 1$ $BF_{Inclusion} = 0.17$. A statistically significant interaction of lag and target confusability was observed, $F(1, 76) = 5.526$, $p = .021$, but there is little evidence for or against the effect either

way, $BF_{Inclusion}$ = 2.0. The absence of main effects of target or lag on intrusions suggests that this interaction is spurious.

Both distinct cues and distinct targets are associated with more response failures (cues: $F(1, 76) = 16.14$, $p < .001$, $BF_{Inclusion} = 130$; targets: $F(1, 76) = 34.87$, $p < .001$, $BF_{Inclusion} = 2.4*10^5$). No other factor significantly affects this measure (cue-by-target: $BF_{Inclusion} = 1.4$).

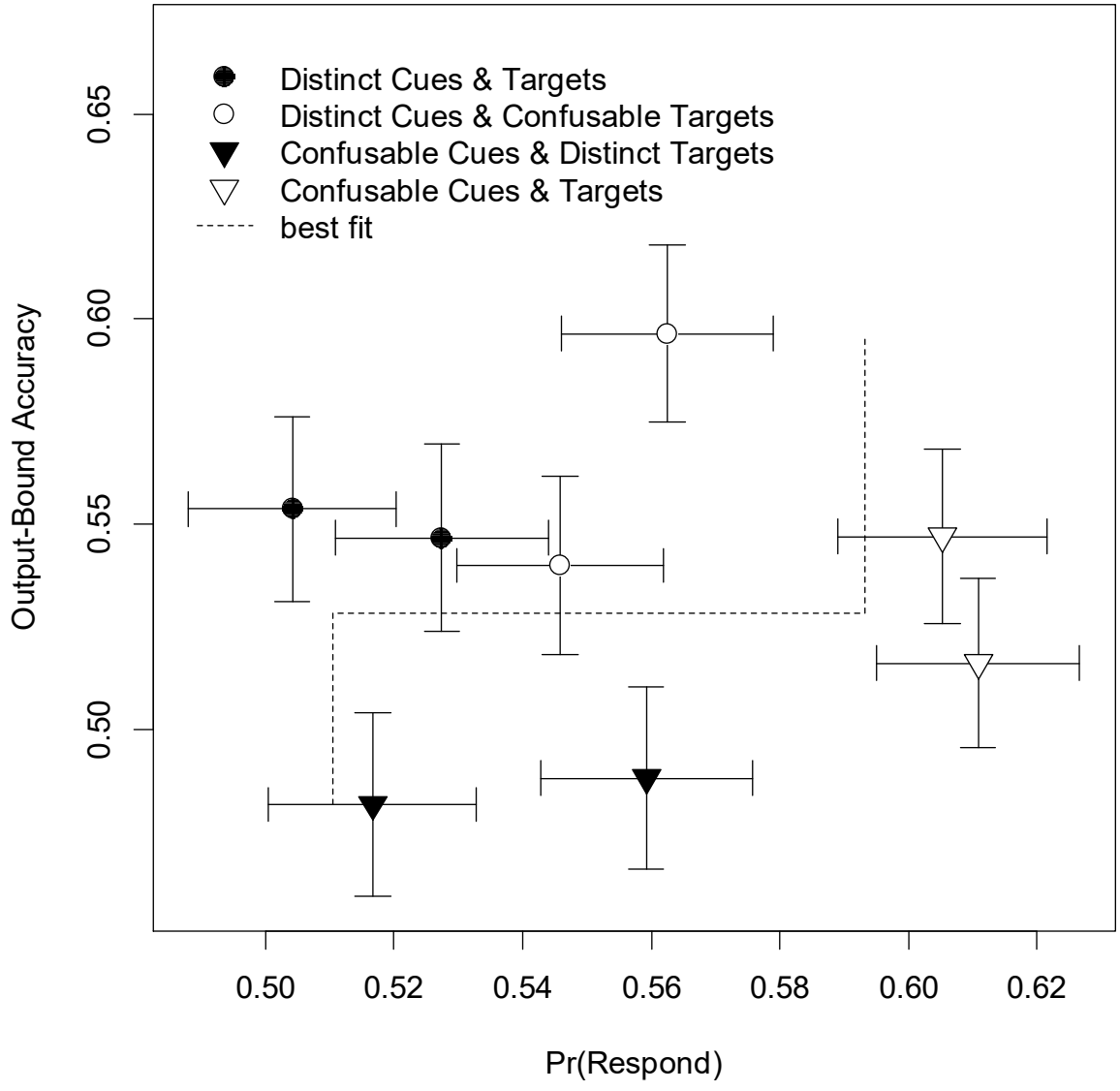

Figure 12. Output-bound accuracy (proportion of responses that are correct) by proportion of trials where a response (a correct response or an intrusion) is given. Black: distinct targets. White: confusable targets. Circles: distinct cues. Triangles: confusable cues.

State-Trace Analysis: Pr(Response) vs Output-Bound Accuracy

The two-sample framework predicts that stimulus-to-cue sampling and recovery treat confusable items differently. If so, then the broad pattern of responses should yield differential influence of cue and target confusability on the response profile. As an assessment of this, we performed a state-trace analysis with conjoint monotonic regression. Our dependent measures were the probability of giving a response (the sum of the correct and intrusion rates) and the conditional probability that the response was a correct response (Figure 12). These two measures are statistically (if not theoretically) independent, in that the rate of responding does not limit the kind of responses given. We reject the null hypothesis of joint monotonicity, p = .02. The confusability manipulation is leading to different effects for cues than for targets overall, consistent with the ANOVA. Generally, confusable targets lead to more responses while confusable cues lead to both more responses and less accurate responses.

Discussion

In all, the effect of confusability depends upon whether it is the cue word or the target word that is confusable. This is indicated both by the differing effects of cue and target confusability and, more robustly, by the differential influence of confusability on the probability of responding and the conditional probability that the response given is correct. Confusable cues yield greater numbers of intrusions, while confusable targets yield more correct responses and more intrusions.

The finding that the presence of a word similar to the target on a study list increases the correct response rate is critical. A local match recovery process has significant trouble accounting for this finding. Once the target trace has been sampled, a local match process would use only the information in the target trace to recover, not information from traces that were not

sampled. Therefore, under a local match, the presence or absence of a confusable target in the lure pair should not affect the correct response rate. It seems, then, to account for this pattern, the recovery process must utilize in some fashion some form of global match.

Of course, one might also wonder whether the presence of a similar item may have impacted how the study list is encoded, which would necessarily affect the outcome of both stimulus-to-cue sampling and recovery. Experiment 2 is similar in some ways to the proactive interference and facilitation paradigm for paired associates cued recall (e.g. Aue et al., 2017). In that paradigm, the cue or target in a to-be-tested pair is presented again alongside a new item. This results in an increase in both correct responses and intrusions for the pair. A number of hypotheses have been offered to explain this phenomenon; all of them involve this repetition influencing how the pair is encoded. If you consider similarity as a continuous measure, with "completely dissimilar" and "the same word" being on opposite ends of the continuum, then it is possible that presentation of an item similar to the cue or target may yield similar outcomes with similar underlying processes.

We can distinguish between effects of encoding and effects of retrieval by noting that encoding effects should be the same for cues and targets. If confusability affects encoding somehow, it should impact how an item or pair is encoded in the same way regardless of whether that item is ultimately a cue or target word at test. Therefore, encoding explanations for this data necessarily make specific predictions about both cue and target confusability manipulations. These can be distinguished from the stimulus-to-cue sampling and recovery predictions.

To that end, a proactive interference and facilitation account would expect similar effects of cue and target confusability on paired associates cued recall performance. Repeating either the cue word or the target word, in the proactive interference and facilitation paradigm, increases

correct response rates. Two encoding-based explanations for this have arisen: either both the cue and target traces are strengthened when either is repeated, or such a repetition creates a "memory triad" where the repeated item links the representations of the two other items it was studied with together with the representation of itself. A reasonable extension of the final account to this experiment would be that the two confusable words are linked to each other and, in so doing, link the other two items they were individually studied with in some memory tetrad.

Stronger item encoding can be ruled out in this experiment because, while the presence of a confusable target is associated with more correct responses, the presence of a confusable cue is not. The presence of a confusable cue should strengthen both the cue and target traces, resulting in a greater chance of sampling the target trace and recovering it and a corresponding increase in correct responses. We do not observe this pattern.

We can also rule out triad/tetrad formation. The triad/tetrad hypothesis predicts the formation of associations, rather than increasing item strengths. The triad/tetrad account therefore does not claim that the strength of the target increases with confusability. It therefore offers no explanation for how, under local match recovery, the presence of an item confusable with the target word on the list could increase the probability of responding correctly.

Having ruled out the encoding-based explanation for the data (see also Appendix C for experiments that do the same), we are left with the explanation that the recovery process in some fashion uses information from the rest of the study list to help produce a response.

Global Match Recovery

Here, we focus on the implications for discrete trace models. We discuss the composite memory models separately. In composite memory models, some information from the similar

item can bleed into the extracted information. This means that the recovery process in these models already uses, in a way, information from other studied words.

The outcome of this experiment suggests that a global match recovery process should use information in such a way that the presence of a confusable item on the study list improves correct response rates. Having a global match process do so is not a trivial matter. In global matching, as increased confusability increases the chances of an error, so too does it decrease the chance of a correct response. Sampling for a correct item is less likely to be successful if there is an incorrect alternative that is similar to the correct item. Some consideration of how a global match recovery process might predict greater correct responding from greater confusability is therefore necessary.

One solution would be to check the familiarity of the proposed response before responding. Suppose a recovery system that attempts to find the word that best matches the to-be-recovered trace then performs a familiarity check of the word as if it were a test stimulus in a single-item recognition test. The probability of selecting the right word is driven by the relative similarity of the to-be-recovered trace to the correct word (versus other words) and by the global match of the selected word to the representation of the study list. The global match familiarity check increases the odds of a correct response because the presence of a similar word on the study list, in addition to the target word, makes the target word more familiar.

Alternatively, the recovery system might use the representation of the study list to assist in transforming the episodic representation of the target into a lexical-semantic representation. Suppose that whatever recursive algorithm used to complete this can decide that there is not enough information to successfully complete the transformation or, alternatively, that the information is too noisy to do so. The addition of a word related to the target in the study list

gives additional information about the target. For instance, if KITTEN is the target word and PUPPY is a related studied word, then there is additional information in the study list about cute, small, furry baby household pets, and this can be used to aid in transforming the trace into a lexical-semantic representation.

The key difference between these two possibilities is in how information is used to determine what to output. Both can use the net familiarity to determine whether to output. The former uses net familiarity to decide what in a competitive process: confusable targets are less likely to be sampled but more likely to be familiar. The latter mixes this competition with cooperation between related items.

On Generate-Recognize. To argue that the memory system considers the global match familiarity of a proposed response before responding is, in essence, to argue for a kind of generate-recognize model. Such models have seen some controversy over the decades. Here, we briefly outline the key components of the model class, the criticisms it has faced, and critique the assumptions made by both the model class and its tests in the context of modern memory theory.

Generate-recognize models are cross-task theories of both recognition and recall which assume, simply, that recall occurs by generating proposed responses that are then fed into a recognition process. Successful recall entails successful completion of both processes, as such the probability of recall is the product of the probabilities of generating the word and of recognizing the word (J. R. Anderson & Bower, 1972; Bahrick, 1970). As per Bahrick's model:

$$\Pr(\text{recall}) = \Pr(\text{generate})\Pr(\text{recognize}) + \Pr(\text{guess}) \quad (9)$$

Such models make two critical predictions. First, recall probabilities will almost always be less than recognition probabilities. This entails two special cases: recalled words must always

be recognizable, and unrecognizable words are never recallable. Second, there is a functional dependence between recognition and recall (Tulving & Wiseman, 1975).

This class of models underwent several tests during the 1970s and 1980s that, at the time, led to the rejection of the model class. Tulving and others found that recall of studied items was often higher than recognition of studied items (Tulving, 1974; Tulving & Thomson, 1973; Watkins & Tulving, 1975; Wiseman & Tulving, 1975), which would seem to directly contradict the first prediction. Tulving and Wiseman (1975) summarized that the proportion of words both recalled and recognized is a relatively clean function of the recognition probability, with the former generally being larger than the later. Gardiner (1988) tested the dependence prediction by, for each participant, testing their memory for the same study list with single-item recognition, free recall, and cued recall in turn. Recognition probabilities were statistically independent of both cued and free recall probabilities. Thus, generate-recognize, so generated by the field, was recognized by the same as an incorrect account of recall processes.

This conclusion may be worth reconsidering. The argument against generate-recognize rests on certain implicit assumptions about the retrieval strategy used by the memory system during recall and on the impacts of testing on the long-term store. To identify two: 1) generate-recognize and its tests assume a single attempt at generate-recognize, 2) the state of the long-term store does not significantly change over the course of repeated testing.

To differing extents, neither of these assumptions are widely held by extant memory models. Modern models of free recall allow attempts to retrieve words many times. More critically, empirical tests of the model class often rely on an implicit assumption that the state of the long-term memory store is unaffected by testing. Most of tests of the model prompt participants to use different, repeated tests of the same study list, using within-subject

comparisons of the test types to draw conclusions. For instance, Tulving and Thompson (1973) found cued recall performance exceeded performance on a preceding recognition test, which was construed as evidence against generate-recognize. This conclusion is based on the implicit assumption that the state of the long-term memory store is the same in both tests and therefore differences in performance are due solely to differences in retrieval process. It is widely accepted that, to the contrary, learning does occur at test and that testing memory has measurable impacts on future tests (e.g. Malmberg et al., 2014; Raaijmakers & Shffrin, 1981; Roediger & Karpicke, 2006).

Composite Memory Models

Composite memory models assume that individual experienced events are stored together in a combined representation. These models (e.g. Matrix Model, TCM, Context Maintenance and Retrieval 2 (CMR2); Lohnas et al., 2015) are thus in contrast to discrete trace models (e.g. SAM, REM), which have unitizable representations of events. Composite memory models should be capable of accounting for this data without much alteration. In composite memory models, the value of the extracted information is a sum of the studied item inputs weighted by the degrees of match between the test stimulus and the individual cues. Study a word pair and probe with one word; the extracted information will be the target representation plus some noise representing the incorrect targets. If a word similar to the target was also studied, then some of that noise will be replaced with information that is like the correct target. As a result, the extracted information will look more like the correct target than otherwise and also resemble the similar item. Any decline in accuracy from the target and similar word being confusable may then be offset by the better match increasing the odds of responding, depending upon the specific assumptions a given composite memory model makes about recovery. For instance, CMR2

assumes that a post-recovery check of the retrieved context to make sure the recovered word was on the study list. Provided that the probability of passing that check increases by an amount greater than that lost when mapping context to item due to target confusability, then CMR2 should be able to account for this finding without alteration. Otherwise, CMR2 may benefit from a more sensitive post-recovery check.

Summary

The data shows differential influence of cue and target confusability on paired associates cued recall performance. For discrete trace models, this is inconsistent with a local match recovery process. In other words, recovery should use the representations of other studied words to help retrieve. Critically, this must be done in such a way that confusability increases odds of correct recovery. In composite memory models the to-be-recovered information is a mixture of the target word and the other studied words. Therefore, extant composite memory model recovery processes should account for this finding provided that the increased familiarity of the information outweighs the increased confusability.

Chapter 4: General Discussion

In two experiments, we have tested several properties of sample-recovery as a broad theory of cued recall. In this model class, the test stimulus is used in some fashion to find a to-be-recalled target, which is then recovered, or transformed into a response. Our data highlight the need for certain elaborations in models that use a simple sample-recovery procedure.

In Experiment 1, we observed that the pattern of data is inconsistent with a sample-recovery model with a single sampling process, but is consistent with the two-sample framework. Paired associates and hybrid cued recall correct response rates are jointly monotonic when cue study time and cue-target associative strength are jointly manipulated. However, under those same manipulations, extralist cued recall correct response rates are jointly monotonic with those from neither task. This suggests that paired associates and hybrid cued recall are completed by chained sampling for the cue trace, then the target trace, while extralist cued recall skips the cue-to-target sampling phase. No extant models use a two-sample process.

In Experiment 2, we further observed that recovery, not just sampling, involves a global match to the study list. Words that are similar to other words on the study list are more recoverable than otherwise. In discrete trace models such as SAM and REM, updates will be necessary to allow for the recovery process to have access to more studied information than what was sampled, including, possibly, a global match to the study list. In composite memory models, we can restrict the space of recovery outcomes to those where the increase in correct response rate from net similarity to the target outweigh any decline in correct response rate stemming from increased confusability.

In this chapter, we offer modifications to several sample-recovery models that should allow for them to handle these findings. We focus particularly on the REM model, as this model

has seen the most success in in accounting for memory phenomena across a wide variety of test tasks. For the other models, and to varying degrees, we offer modifications more as tentative suggestions. These suggestions prompt implementation questions that warrant further inquiry.

On Current Sample-Recovery Models

On Implementing Two Sampling Phases in Models

These results suggest that we implement, in the cued recall retrieval process of all models, multiple sampling phases prior to recovery. The same general process works in all cases. For paired associates and hybrid cued recall, attempt to identify the cue trace given the test stimulus, then attempt to identify the target trace given the cue trace. In extralist cued recall, identify the to-be-retrieved cue/target trace given the test stimulus. Generally, stimulus-to-cue sampling relies on the encoded strength of the cue trace and its degree of similarity to the test stimulus and, otherwise, shares properties with recognition memory. Models with fleshed out accounts of cue to target sampling also tend to have well developed accounts of free recall, and to that extent often rely on causal contiguity.

Viewed in this light, extralist cued recall theoretically resembles recognition memory more than free recall. Paired associates and hybrid cued recall exist on the boundary between recognition and free recall. The two tasks share properties with recognition because the stimulus is used to find the representation of the cue word. They share properties with free recall because the cue trace is used to find the target trace; traces are used to find traces.

On Implementing Global Match Recovery

Discrete trace models and composite memory models need differing levels and kinds of modification to account for the data from Experiment 2. Discrete trace models appear to need something resembling a global match to reproduce the pattern of data for confusable targets in

Experiment 2. Composite memory models have such a mechanism built in. As such, the modifications required to reproduce the finding are less structural and more parametric.

We do not delve much deeper into this topic. Instead, we offer brief thoughts on a variety of models and a proof of concept in the REM model that a mechanism which ties global familiarity to the probability of recovering an item could reproduce the pattern of data observed here.

Retrieving Effectively from Memory (REM)

To jointly account for extralist and paired associates cued recall we updated several aspects of the model. The highlights: We have included association information, which now serves as the only means by which the model can infer what item was studied with what. To make that point clear, items studied together are not concatenated. Instead, they are encoded separately and concatenated with associative information. The associative information represents an item's co-occurrence with some other item. If two items were presented at the same time, the associative information in the two traces will be similar. We have implemented two sampling algorithms. Stimulus-to-cue sampling finds the best matching item trace to the test stimulus. Cue-to-target sampling uses the associative information concatenated to each item to determine which items were studied together. This yields a joint representation of the two forms of cued recall that also accounts for the task-based dissociations in word frequency.

Core Elements of REM

We first outline the core elements of the model; these are left unaltered from their original form as outlined by Shiffrin and Steyvers (1997, 1998). In REM, stimuli (and/or their lexical-semantic representations) and episodic representations of encoded events are both represented as lists of $w = 20$ features. Lexical-semantic representations of stimuli are complete

and errorless, while an episodic representation of a stimulus is noisy (in that some features do not match the stimulus) and incomplete (in that there are some features with no information at all). Stimuli are generated from a geometric distribution, typically with parameter $g_{environment} = .4$, such that the smallest and most common feature value is 1. Higher g values tend to generate higher frequency—more common—stimuli, while lower g values tend to generate lower frequency stimuli.

An episodic representation of a stimulus is both noisy and incomplete. A stored representation will on average have some fraction *u* of the features encoded, the remaining features will take empty values. Of those features that are encoded most, usually $c = .7$ on average, will be directly copied from the stimulus. If the feature is not directly copied it is instead randomly sampled from the geometric distribution with g_{system} = .4. This means that randomly sampled features will sometimes match the stimulus features by chance.

The centerpiece of REM is, arguably, how the similarity between stored representations and stimuli is computed. For any stimulus and representation, the similarity is a marginally informed likelihood ratio that the representation was encoded from the stimulus, versus the possibility that it was not. The value of the comparison between stimulus and representation *j* and *k* is written as λ_{jk}:

$$\lambda_{jk} = (1-c)^{n_q} \prod_{\forall i} \left(\frac{c + (1-c) g_{system} (1-g_{system})^{i-1}}{g_{system} (1-g_{system})^{i-1}} \right)^{n_{im}} \tag{10}$$

Where c and g_{system} are the same parameters used during encoding operations, n_q is the number of encoded, mismatching feature values in j and k, and n_{im} is the number of encoded, matching features with the value i. The value of λ approaches zero as more features mismatch and the value of λ approaches infinity as more features match and as those features become larger and

less likely to have been randomly sampled. Strongly encoded words and words with many uncommon features yield higher λ when compared to their episodic representations, versus weakly encoded words and words with more common features.

Cued Recall in REM

Upon presentation of a word pair, two traces are added to the long-term store. Each trace contains the representations of one item, the context of the experiment, and the association between that item and the other item it was studied with. This is a departure from prior implementations of the model. In prior implementations, words studied at the same time are associated by concatenation: they are encoded together in the same trace; once an item is identified the location of its paired associate is known. Consequentially, REM predicted that the probability of successfully determining what item was studied alongside the sampled item was $\Pr = 1.0$. Any REM model where concatenation occurs in this manner must therefore "play dumb" to make an error in determining what words were studied together. To emphasize that determining what was studied with what relies on the quality of the associative information, we do away with item-to-item concatenation.

Without the concatenation mechanism, the model must instead rely on the content of each trace—the associative features—to determine what was studied with what. Upon presentation of a word pair, in addition to the two stimuli, REM sees an association stimulus of length $w = 20$ representing their co-occurrence. The encoding of the word pair generates two traces representing the association stimulus, one per word. Each trace includes an episodic representation of the word and an episodic representation of the association. The model encodes the association separately for each trace, thus the association vectors concatenated to two co-occurring items in the long-term store are similar, but not the same. Strength of the

association—manipulated by the study task—is determined by the degree of encoding u_{assoc} (we distinguish this from the encoding strength of the item: u_{item}).

Paired associates and hybrid cued recall retrieval in this new implementation of REM use three stages. The first stage uses the contents of the test stimulus to find the best matching item; the model continues only if the likelihood ratio of the sampled item is greater than a threshold. If the cued recall task requires knowledge of the word pair, activate the associative features in the trace and compare it to the associative features in the other stored traces; take the best match. Proceed if that trace is greater than some threshold. Next, probabilistically decide whether to attempt recovery, where the probability is related the fraction of features encoded in the sampled trace. If so, activate the best matching lexical entry; that entry is the response given by the model. If the model fails to pass either threshold check or decides that not enough information is encoded in the to-be-recovered trace, a response failure is recorded.

Typically, REM uses the Luce Choice rule when determining which item to sample. This process probabilistically determines the sampled item; the odds of sampling an item are equal to the weighted evidence for the item divided by the weighted evidence for all the options:

$$LC(b|a) = \frac{\lambda_{ab}{}^{\gamma}}{\sum_{d=1}^{N} \lambda_{ad}{}^{\gamma}} \tag{11}$$

For the purposes of this model, we sample by using the MAX rule, then check to see that the sampled trace is sufficiently similar to the stimulus by comparing λ to some criterion ε. The only difference between stimulus-to-cue sampling and cue-to-target sampling is what information is used. In stimulus-to-cue sampling, the contents of the test stimulus are compared to the item and context information in each trace. The probability of using item i among possible items l, given a test stimulus h, is:

$$\Pr(\text{sample } i \text{ given } h) = \Pr(\lambda_{hi} = \max(\lambda_{hj}))\Pr(\lambda_{hi} > \varepsilon_1) \quad (12)$$

Where ε_1 is the relevant threshold for this comparison.

For cue-to-target sampling, the associative information from the sampled trace is treated like a stimulus and compared to the associative information in the other traces. The same equation is used for selecting target k among possible items l, given sampled trace i, using a second threshold ε_2: $\Pr(\text{sample } j \text{ given } i) = \Pr(\lambda_{ij} = \max(\lambda_{il}))\Pr(\lambda_{ij} > \varepsilon_2)$. We use a second threshold parameter because the distributional properties of the association-to-association comparison differ.

Note that extralist cued recall skips cue-to-target sampling because, once stimulus-to-cue sampling has been completed, the to-be-recovered trace has already been found.

To recover, the model decides whether to output based upon the proportion of features encoded, then decides what to recover by comparing the to-be-recovered trace to the set of stimuli that were studied and selects the entry the trace most likely represents, using the MAX rule. The process by which this model selects the lexical entry to recover is novel for REM models. For recovery, the probability of outputting word m, rather than other possible words n, given trace k, is:

$$\Pr(\text{output } m \text{ given } k) = {p_k}^{\tau}\Pr(\lambda_{km} = \max(\lambda_{kn})) \quad (13)$$

where p_k is the proportion of encoded features in trace k, $\tau \in [0, 1]$ weighs that proportion, λ_{km} is the likelihood ratio of to-be-recovered trace k matching word m, and $\max(\lambda_{kn})$ is the largest likelihood ratio among all comparisons of the target trace to the lexical-semantic store. In this, we extend the standard recovery formula, $\Pr(\text{recover } m) = {p_k}^{\tau}$, to allow the model to allow for different kinds of errors in determining what word a trace represents by adopting an analogous mechanism to that used in early versions of TODAM. This recovery process allows for semantic

intrusions; lexical entries related to the target words are more likely to be incorrectly selected than those unrelated to the target.

Previous cued recall REM models used a less detailed mechanism. Diller, Nobel, and Shiffrin (2001) offers the most detailed mechanism in REM thus far. They implemented recovery by, first, deciding whether or not to recover using $\Pr(\text{recover}) = {m_k}^{\tau}$ where m_k is the proportion of encoded features that match the correct target, then stating that the probability of correct recovery is a function of the item's likelihood ratio similarity to the correct target word: $\Pr(\text{recover correctly}) = 1 - e^{-\psi {\lambda_{km}}^{\gamma}}$, where γ is the same weighting parameter used during Luce choice sampling, λ_{km} is the likelihood ratio that to-be-recovered trace k is a representation of correct response m, and $\psi = 0.86$ is a "best fit" scaling of the probability curve. If an incorrect target is sampled it is simply assumed that the response, if given, will be an intrusion. This function thus states that stronger target items are more likely to lead to recovery of the correct target.

Our mechanism improves on this in two critical ways. First, it offers a more detailed account as to why weaker representations of target words is associated with more intrusions. Weaker representations look less like the target word and more like other words. Second, it simulates what item is being intruded with and thus allows for more detailed predictions of performance as a function of the semantic relationships between stimuli. This second mechanism is critical if the model is to distinguish between extralist and intralist intrusions.

Simulations and Predictions

We fixed several parameters across simulations of Experiment 1 and 2. Similarity between stimuli in REM is parameterized by s. Similar stimuli are generated using the following algorithm: generate a stimulus, for each feature copy that value to the similar stimulus with

probability s, then randomly generate the remaining features. This parameter governs the similarity between test stimuli and cue words in these simulations. Both experiments use the same stimulus set pulled from the USF free association norms (Nelson et al., 2004) such that related or confusable words have forward and backwards strengths to each other between 0.14 and 0.30. Similarity between words is a critical driver of the performance in extralist and hybrid cued recall relative to paired associates cued recall, the critical manipulation of Experiment 2, treated as constant across subjects, and formed outside of and independently of the experiments. It therefore seemed important to take extra effort to derive an appropriate value for the model, rather than choose a one by hand. To derive what similarity value corresponds to this range of forward and backward strengths, we simulated the free association task (see Appendix D) and computed the average similarity value for word pairs that met the criteria. As per the algorithm to generate similar items, the probability $\Pr(m_i)$ that two given stimuli match along a dimension with value v_i is $\Pr(m_i) = s + (1-s)G(v_i)$ and the probability $\Pr(q_i)$ that they mismatch (given a predetermined stimulus has feature value v_i in the dimension) is $\Pr(q_i) = (1-s)(1-G(v_i))$, where $G(v_i) = g(1-g)^{v_i-1}$. By utilizing these equations, we get a similarity parameter of $s = .55$.

We set the paired associates cued recall stimulus-to-cue sampling threshold to a default value of $\varepsilon = 1.0$ and fix the extralist/hybrid stimulus-to-cue sampling threshold to $\varepsilon = 0.5$. A lower threshold for extralist and hybrid cued recall is reasonable because the test stimulus does not match a studied item but, instead, is merely similar. One should therefore expect less evidence that the stimulus and the correct representation are the same (a lower λ) in extralist/hybrid cued recall than in paired associates cued recall.

Table 4

Parameters of the REM model

		Parameter Value			
		REM-a		REM-b	
Parameter	Description	Expt. 1	Expt. 2	Expt. 1	Expt. 2
$w_{context}$	# of context features	6 [c]			
w_{item}	# of item features	20 [a]			
$w_{association}$	# of association features	20 [a]			
g	geometric distribution parameter	.4 [a]			
c	probability of copying stimulus value during encoding	.7 [a]			
$u_{context}$	encoding strength, context	.30	0.40	.55	.65
$u_{item,weak}$	encoding strength, weak items	.30	0.40	.55	.65
$u_{item,strong}$	encoding strength, strong items	.40	NA	.65	NA
$u_{association,weak}$	encoding strength, study separately	.45	NA	.52	NA
$u_{association,strong}$	encoding strength, study together	.65	0.75	.65	.75
s	similarity of 2 related items	0.55 [b]			
$\varepsilon_{stim2cue}$:	cue-to-stimulus sampling threshold				
PACR	stimulus and cue word match	1.0 [a]			
ELCR&HYBR	stimulus and cue word are similar	0.5			
$\varepsilon_{cue2target}$	association sampling threshold	5.0		8.0	
τ	recovery odds parameter	0.5 [a]			
$\tau_{confusable}$	… when the target is confusable in Experiment 2	N/A	0	N/A	0

Notes: a: default value. b: derived value. c: Chen et al, yet to be published. Otherwise: hand fits. The model makes one attempt to sample and recover, so *Kmax* = 1.

Finally, we set the recovery weighting parameter $\tau = 0.5$. This is the same as prior models, but because our recovery function is different in substantial ways from prior versions of recovery in REM and we should have no theory-driven prior beliefs about where this should be.

The value from the old literature just happens to fit the recovery outcomes found using Bayesian multinomial model from Experiment 1.

Experiment 1

For Experiment 1, we implement the model as we have described it using the study list structure outlined in Experiment 1. We separately simulate lists of 18 pairs where associations are strong or weak and test them using paired associates cued recall, extralist cued recall, and hybrid cued recall. During the two sampling phases, what could be sampled or recovered was limited to the items/associations within that study-test block. During extralist and hybrid cued recall, the 18 related items served as test stimuli. We included context in stimulus-to-cue sampling. As per unpublished work by Chen, Cox, Wilson, and Criss, we set the number of context features to 6 (vs 20 features per item) and kept the encoding strength for contexts and items equal. The context stimulus remained constant throughout a study-test block.

For Experiment 1, we set the encoding strengths for weak and strong items to $u = .30$ and $u = .40$ respectively. Study task is designed and modeled as a manipulation of associative strength. We set the encoding strength for weak (study separately) and strong (study together) associations to $u = .52$ and $u = .65$ respectively. Context encoding strength was set to $u = .30$. Context encoding strength was constant across conditions (Malmberg & Shiffrin, 2005).

As seen in Figure 13(A), the model captures the pattern of means rather well. The model is also capturing the presence or absence of differential influence where it was observed in the data (Figure 14). By design, the model naturally captures joint monotonicity of paired associates and hybrid cued recall because the same sampling processes are used to recall. It also captures

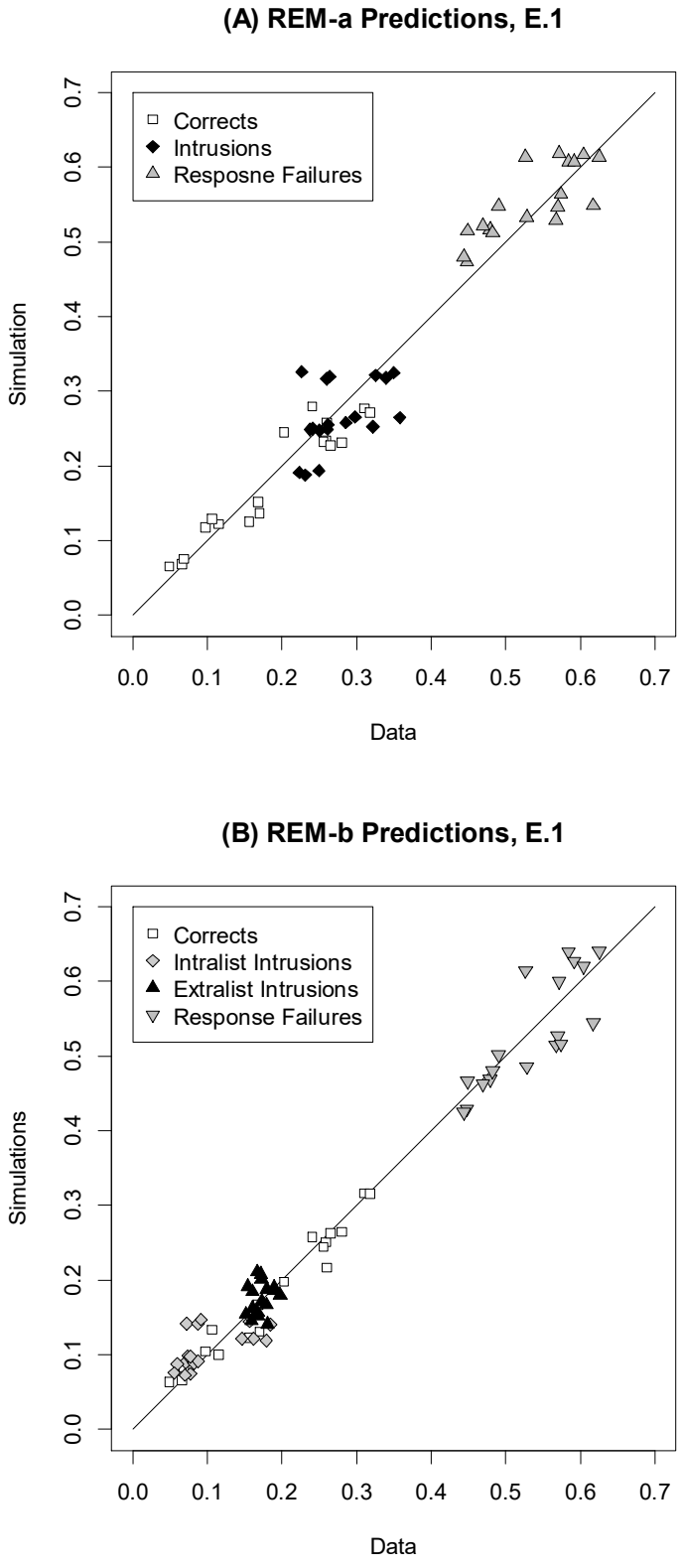

Figure 13. Model fits to Experiment 1 for (A) REM-a and (B) REM-b.

the lack of joint monotonicity between either of the two tasks and extralist cued recall because extralist cued recall does not sample from cue to target.

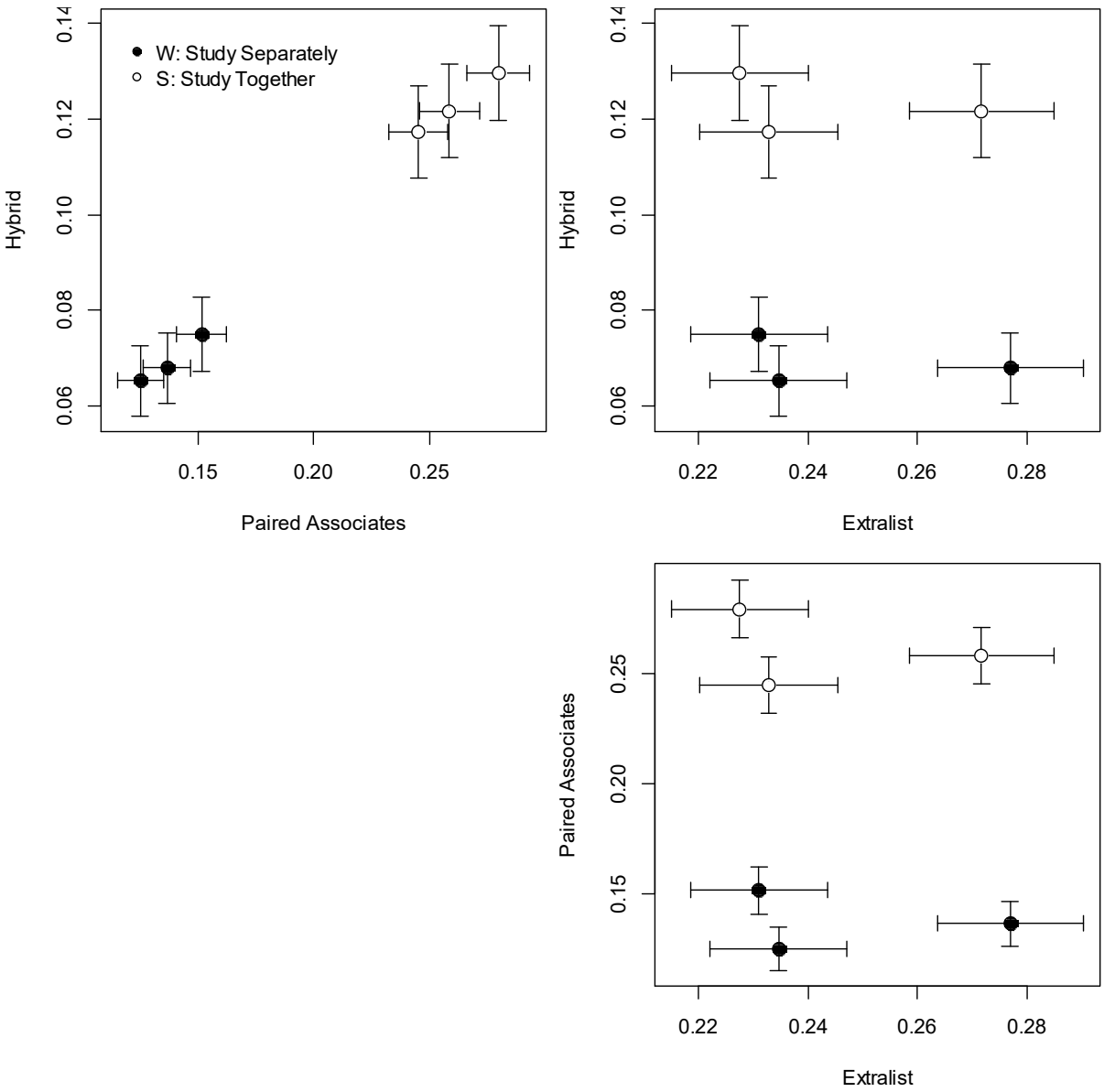

Figure 14. REM-a simulated correct response probabilities for Experiment 1, in the format of state-trace plots.

Accounting for Extralist Intrusions – REM-b. In Experiments 1 and 2, more than half of the intrusions observed were extralist intrusions. To account for this pattern, we modified the recovery algorithm to include all possible lexical entries, rather than just those that were present

at study or test. The model, during recovery, now takes the best matching lexical entry from among 17,000 stimuli, an estimate of how many words an undergraduate student knows (D'Anna et al., 1991; Goulden et al., 1990). Naturally, a randomly generated set of 17,000 words would not be expected to possess the same semantic structure seen in real life, however certain properties can and do arise simply from modeling free association in REM as a Luce choice sample of the likelihood ratio comparisons between randomly generated stimuli (Appendix D).

The additional noise during recovery necessitates parametric adjustments to allow the model to match observed performance. We therefore set the encoding strength for weak items to $u = .55$, strong items to $u = .65$, and weak associations to $u = .52$ and the cue-to-target sampling threshold to $\varepsilon = 8$. To disambiguate this version of the model from that used before, call the prior instantiation REM-a and this new instantiation REM-b. Figure 13(B) plots observed and simulated performance across correct responses, intralist intrusions, extralist intrusions, and response failures. As one can see, REM-b fits the data quite well with these adjustments.

Experiment 2

Here, we use a simple modification to REM-a and REM-b as a proof of concept that a process where confusable items are more readily recovered can recreate the pattern of data in Experiment 2 (Figure 15). For to-be-recovered words with a similar item on the studied list, we set $\tau = 0$. In other words, if another studied word is similar, recovery will occur. Paired associates cued recall performance in this experiment is generally better than what was observed in Experiment 1 and so we set the item and context encoding parameters $u = .40$ for REM-a and $u = .65$ for REM-b and association encoding parameter $u = .75$ in both variations. As one might expect, adjusting τ in this way increases correct response rates and intrusion rates for

similar targets, leaving the rate of intruding with an unrelated pair constant. This is proof-in-concept that a global match for similar pairs is utilized.

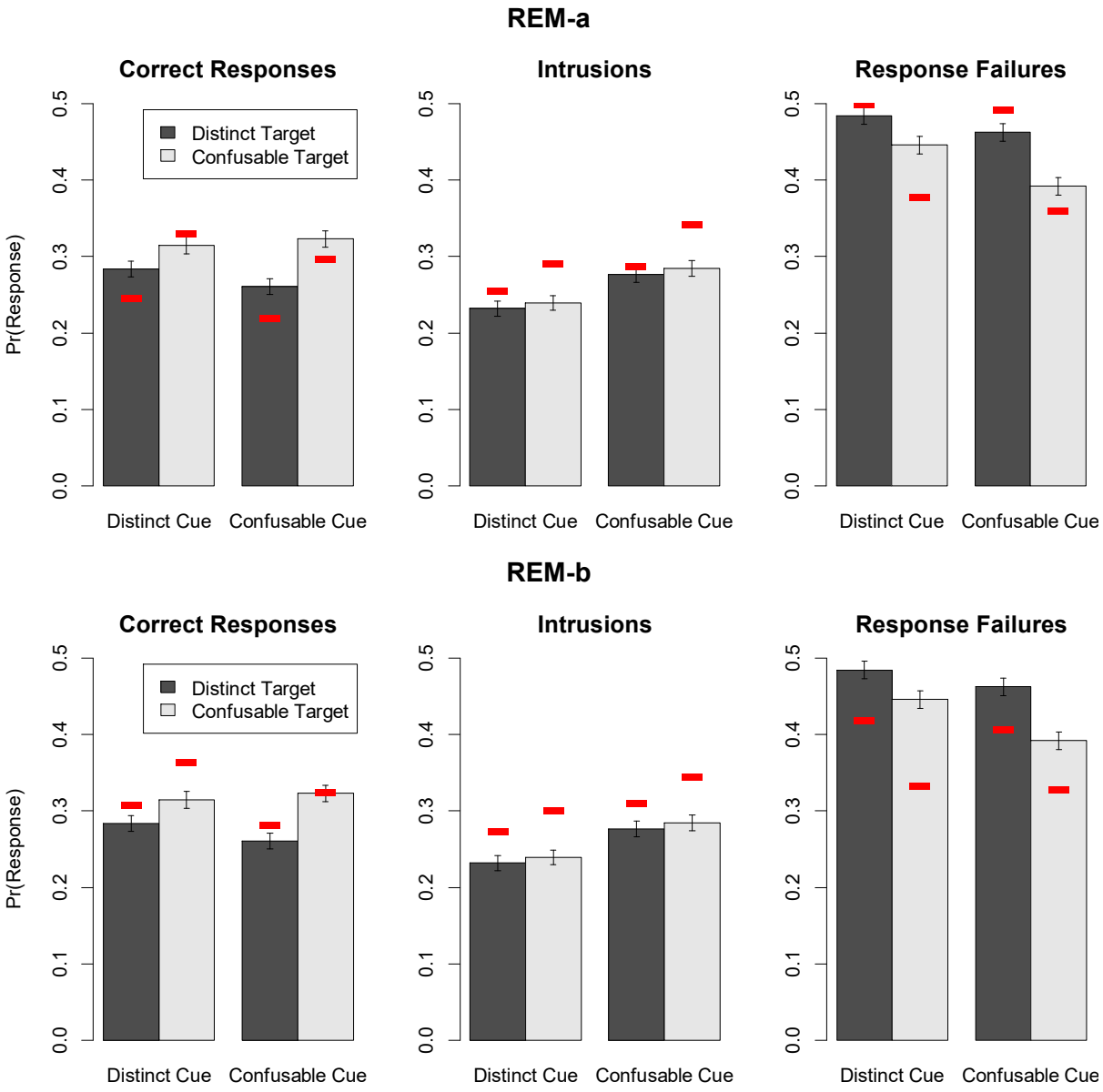

Figure 15. Correct responses, intrusions and response failures from (bar plots) Experiment 2 and (red lines) model predictions from REM.

Attempts to implement this as a process have not yet managed to recreate this pattern, however. Suppose, in addition to the current recovery process, the model performs a final recognition check to see if the proposed target was on the study list. This is done in REM by

averaging the likelihood ratio comparisons between the proposed target stimulus and the traces encoded at study and comparing them to a criterion. This average, Φ, is larger on average for confusable targets and confusable targets are therefore more likely to be recognized. However, among the set of target words that could be recognized, the probability of being correctly sampled is greater for distinct words than it is for confusable words. This outcome is paradoxical because, by experimental design, the probability of correctly sampling the target stimulus is the same for both confusable and distinct targets. Distinct and confusable words both have a related word in the lexicon (and that related word happens to on the study list for confusable words) and this is modeled in REM-b. The net effect is that these two trends cancel each other out, leading to equal chances of correct recovery for distinct and confusable targets.

Word Frequency Effects

In paired associates cued recall, high frequency words are associated with greater correct response rates. In extralist cued recall, low frequency words are associated with greater correct response rates. As previously explained, models of cued recall that employ a single sampling process cannot account for this task dissociation.

Because we have now split sampling into stimulus-to-cue sampling and cue-to-target sampling, a REM account for the low-frequency advantage in extralist cued recall and high-frequency advantage in paired associates cued recall becomes clear. The low-frequency advantage for extralist cued recall occurs for the same reason as the low-frequency advantage in recognition: uncommon words are less confusable than common ones by virtue of having fewer common features. The advantages for high-frequency words occurs for the same reason as in SAM: they form stronger cue-to-target associations.

As a demonstration, we run REM-a for extralist and paired associates cued recall of high-frequency versus low-frequency pairs. For these purposes we mirror three experiments across two studies. For extralist cued recall, we simulate Experiments 1 and 2 from Nelson and McEvoy (2000). For paired associates cued recall, we simulate Experiment 2 from Criss, Aue, and Smith (2011) which, like Nelson and McEvoy, manipulates word frequency fully within-list.

In Nelson and McEvoy's (2000) first experiment, the similarity between test stimulus and target was crossed with the word frequency of the test stimulus, within-list. Participants studied 24 target words and were tested with extralist cued recall. Six targets were tested in each condition: high frequency and similar; high frequency and dissimilar; low frequency and similar; low frequency and dissimilar. In their second experiment, the word frequency of the test stimulus was crossed with the word frequency of the target word within-list and in the same way as before; list lengths were also 24 words and 24 test trials. Stimulus-target similarity in this second experiment was close to that of the similar condition from the first experiment.

To model these two experiments, we manipulate the word frequency and similarity parameters. In REM, word frequency is parameterized by $g_{enviornment}$, we set the frequency parameter for low frequency words to $g = .35$ and high frequency words to $g = .45$. We set the similarity parameter to $s = .6$ for similar cues and $s = .45$ for dissimilar cues. We set the item encoding parameter to $u = .65$ and keep the other REM-a parameter values as they are in Experiments 1 and 2.

In Experiment 2 of Criss, et al. (2011), participants were studied lists of 20 word pairs for 3 s each and tasked to study the two words by placing both members of each pair in one sentence or image. Each study list was followed immediately by a 60 s distractor task and then a paired associates cued recall test. The word frequency of the cue and target were counterbalanced such

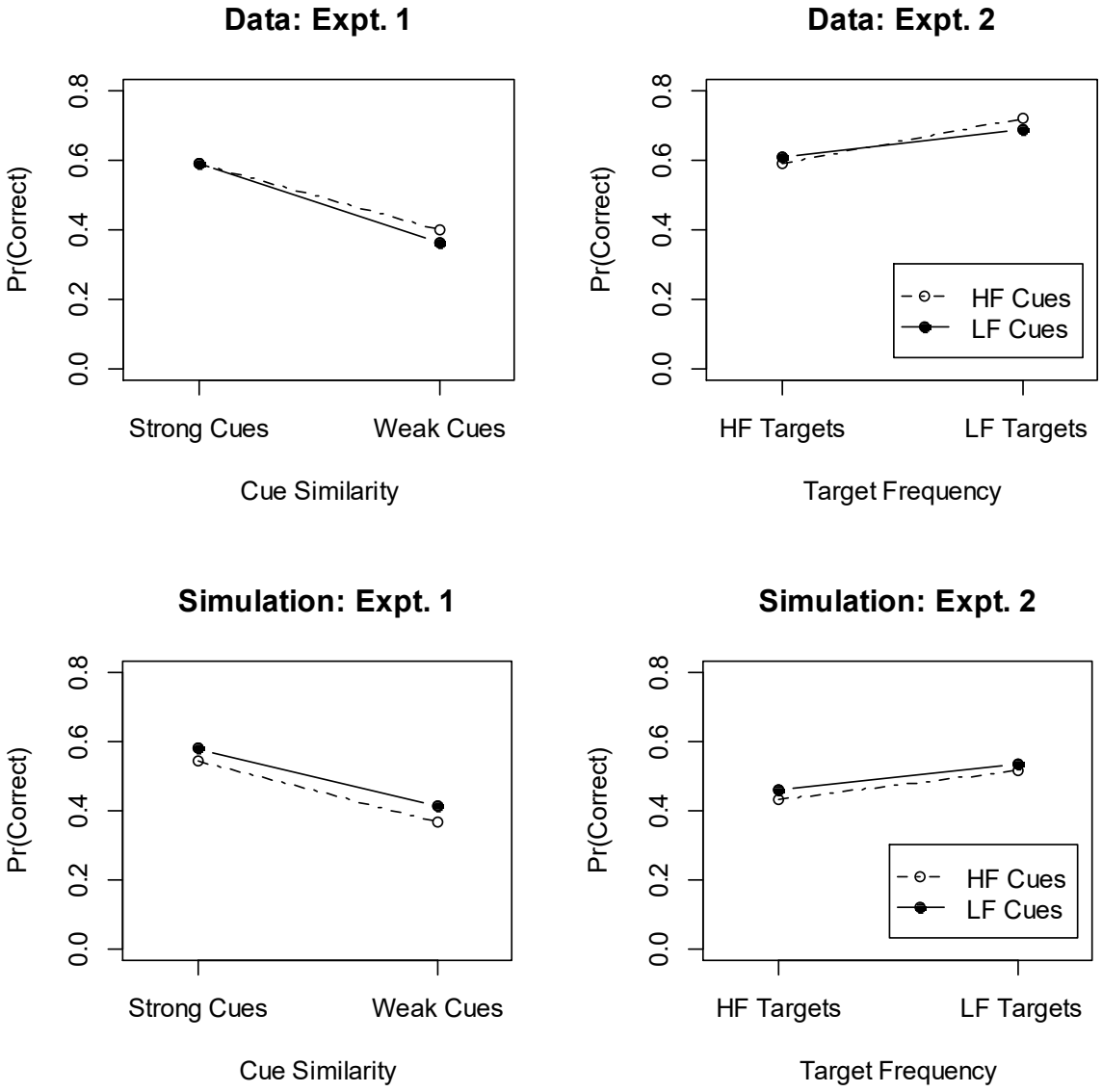

Figure 16. Simulations of Nelson and McEvoy (2000). Top: Data from their experiments. Bottom: simulations from REM-a.

that 5 pairs were both low frequency, 5 pairs had a high frequency cue and lower frequency target, 5 pairs had a low frequency cue and a high frequency target, and 5 pairs had high frequency cues and targets.

To simulate this experiment in REM-a we use the same word frequency adjustments to $g_{environment}$ as before. As per the SAM theory, we also allow the strength of encoding for cue-

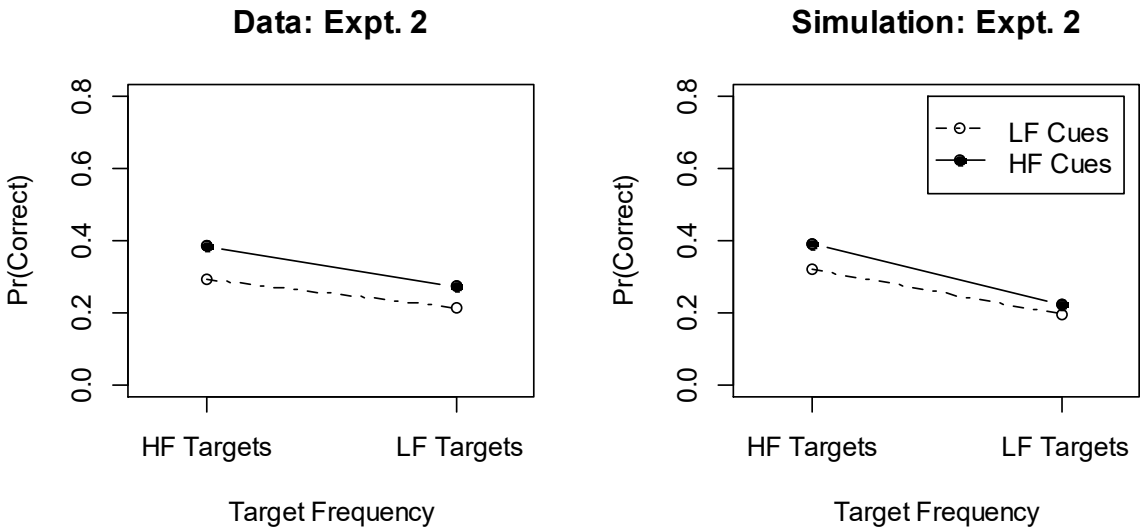

Figure 17. Simulations of Criss, Aue, and Smith (2011) Experiment 2. Left: Data, averaging across the study task condition. Right: Simulations from REM-a.

target associations to adjust based upon the word frequency of the pair. Item encoding strength $u = .50$ was the same for all conditions. We set by hand fitting the association encoding parameter to $u = .90$ for high frequency pairs, $u = .60$ for pairs of mixed frequency, and $u = .45$ for low frequency pairs. We keep the other parameters as they are for the REM-a simulations of Experiments 1 and 2.

Perhaps unsurprisingly, these parameterizations allow for the paired associates – extralist word frequency dissociation. In extralist cued recall, low frequency targets are less confusable than high frequency targets, resulting in a greater probability of sampling the correct item. We are thus able to reproduce the results of Nelson and McEvoy (2000), see Figure 16. In paired associates cued recall (Figure 17), word frequency makes low-frequency cues less confusable as well, but the resultant increase in stimulus-to-cue sampling accuracy is not as large as the high-frequency benefit of stronger cue-target associations. Stimulus-to-cue sampling accuracy is, already, much greater in paired associates cued recall than in extralist cued recall and so the

benefit derived from low frequency pairs is smaller for paired associates cued recall as than for extralist cued recall.

Set Size Effects

To recap from earlier in this document, (pre-experimental) set size refers to the number of words semantically associated with the word in question. In terms of the free association norms, this is some combination of the number of unique words generated upon presentation of the word as a test stimulus and the number of test stimuli that generate that word as a response. In extralist cued recall, words with smaller set sizes are associated with a greater correct response rates (Nelson & Zhang, 2000), while in paired associates cued recall words with greater set sizes are associated with greater correct response rates (Nelson et al., 1990).

The updated REM framework offers a clear account of the phenomena. We do not offer simulations; such simulations would require a modification to REM-b to also include episodic representations of prior-list experiences. Words with smaller set sizes have a fewer number of words with which it is confusable. Stimulus-to-cue sampling is improved for these words because the relative paucity of related words makes it is less likely that stimulus-to-cue sampling will select an item related to the cue. At the same time, in the same manner as high frequency words, words with higher set sizes form stronger episodic associations with unrelated words. This results in greater performance for high set size words during cue-to-target sampling. In extralist cued recall, low set size words are advantaged because only stimulus-to-cue sampling is used to find the to-be-recalled item. In paired associates cued recall, cue-to-target sampling is less accurate than stimulus-to-cue sampling, and so the high set size advantage in cue-target associations overcomes its disadvantage during stimulus-to-cue sampling, resulting in a modest advantage for high-set-size words.

Future Work: Extension to Free Association

Having developed a model of extralist cued recall in REM, the model can be further extended to account for free association data. Most of the examinations of extralist cued recall consider the probability of recalling a target from an extralist cue in relation to the probability of responding with that target given the cue in the free association task (see, e.g.: the collected works of Douglas L. Nelson). We therefore know a great deal about the relationship between these two tasks. Furthermore, models of this relationship already exist (Nelson et al., 1992; Steyvers et al., 2005) and should offer some guidance on how to proceed in development. Preliminary work on extending REM to free association has been promising (see Appendix D).

A number of challenges will need to be resolved. The first is philosophical. REM does not map specific words to lexical representations. Instead, it treats the lexical representations as a random draw. In the preliminary model, it is a random draw from the geometric distribution, but even if the lexicon was simulated with a generative process, lexical representations would still be a random sample from some distribution. The free association task considered specifically what words are responded with given the presentation of other words. Absent a mapping of individual words to a specified vector of features, REM could only ever model the general properties observed for, and relationships between, words of a given kind. Take word frequency: REM can predict or simulate the relationship between test stimulus and response word frequency, or the relationship between that property and response metrics like set size or forward/backward strengths, but would not be able to fit the network of associations for specific words as they exist. That said, REM's approach can demonstrate generalizability, which we consider to be an advantage overall.

Another challenge is posed by the observation that response probabilities in free association violate the triangle inequality. Given three words A, B, and C, the triangle inequality states that the psychological distance between A and C should never exceed the summed psychological distances between A and B and between B and C. This constraint is routinely violated, leading to the development of models that allow for non-symmetric similarities between items (e.g. Griffiths et al., 2007). Models like REM can handle this violation if they include a dimensional attention mechanism such as that used by the Generalized Context Model (Jones et al., 2018; Nosofsky, 1984). REM has no such mechanism. The adoption of one would require a principled account as to what features are attended to more or less in a given circumstance.

Finally, a REM model of free association would need to account for the impacts of episodic priming on free association responses. Study of words prior to free association can impact free association response probabilities in a number of ways. Prior study of probes contextually influences response probabilities. If one studies the sentence "The boat rested along the bank." participants would be more likely to respond with RIVER given BANK as a test stimulus, but if they instead studied "Sue went to the bank yesterday." they would be more likely to respond, instead, with MONEY (Zeelenberg et al., 2003). Study of a mediating word can induce responses indirectly related to the test stimulus and lower response times. For instance study of ANIMAL could prime participants to respond with VEGETABLE when probed with DEER, because ANIMAL is associated with both words (Nelson & Goodmon, 2002). Priming effects also seem to depend on the symmetry and directionality of the semantic relationship. Responses with a strong backward association to the test stimulus can be primed by prior study of the response. Not so for responses with only a strong forward association to the test stimulus.

As a model of episodic memory and a proposed model of semantic memory and retrieval as well, it would be incumbent of a REM model of free association to account for these and other related findings.

Summary

The adaptation of REM presented here extends the model to extralist cued recall and the novel hybrid cued recall task by including associative information that is used during cue-to-target sampling, in contrast to stimulus-to-cue sampling which relies on item information. The adaptation accounts for word frequency effects in paired associates and extralist cued recall and offers a clear verbal explanation for the pre-experimental set size dissociations in those tasks.

Considerations of Other Memory Models

Search of Associative Memory (SAM)

In prior implementations of SAM, the cue trace is activated automatically in paired associates cued recall simulations. In other words, sampling in SAM's current formulation implements cue-to-target sampling and not stimulus-to-cue sampling. Adapting SAM thus requires including a stimulus-to-cue sampling mechanism in the model and allowing this mechanism to take similarity to the stimulus as an input. Ultimately, what one would desire is a mechanism that incorporates both the basic principles of the Processing Implicit and Explicit Relations (PIER) family of models (Nelson et al., 1992, 1998, 2007, 2013) and a differentiation mechanism for items. Further empirical work is needed to determine the best way to do this.

Status Quo

SAM in its current form activates the representation of the cue word automatically upon presentation of the test stimulus then samples for the correct target using cue-to-target and context-to-target associations. As it is canonically written, stimulus-to-item strength is S_{qi} and

context-to-item strength is S_{ci} for cue word q, context c, and target i. Let j be the alternative, incorrect items. To let item similarity impact sampling odds, can we adopt the Gillund & Shiffrin (1984) version as a starting point and allow for an additional parameter m_{qi} to denote the degree of match between test stimulus q and item i. The probability of correctly sampling the target word is then:

$$\mathrm{P}_S(i|c,q) = \frac{m_{qi}S_{qi}S_{ci}}{m_{qi}S_{qi}S_{ci} + \sum_j m_{qj}S_{qj}S_{cj}} \tag{14}$$

Study times in Experiment 1, 4 s vs. 8 s, were long enough to prevent a substantial difference in context-to-item encoding between the two conditions (Malmberg & Shiffrin, 2005). Further, the critical manipulations were cue and associative strength, not target strength. We can therefore say $S_{ci} = S_{cj}$ is an acceptable approximation and simplify:

$$\mathrm{P}_S(i|c,q) = \frac{m_{qi}S_{qi}}{m_{qi}S_{qi} + \sum_j m_{qj}S_{qj}} \tag{15}$$

In other words, in the context of Experiment 1, SAM can account for effects stemming from manipulations of stimulus-to-item similarity (e.g. paired associates vs hybrid cued recall) and cue-target associations (study task), but not the strength of the cue under this implementation.

Updates to the Model

A two-sample reformulation might take the following form: let Q now be the test stimulus, q its representation in the long-term store, and k the alternative incorrect items to q. As above i denotes the target trace, and j the incorrect alternatives to i.

$$\mathrm{P}_S(q|c,Q) = \frac{d(m_{Qq}, S_{qq})S_{cq}}{d(m_{Qq},\ S_{qq})S_{cq} + \sum_k d(m_{Qk}, S_{kk})S_{ck}} \tag{16}$$

$$\mathrm{P}_S(i|c,q) = \frac{S_{qi}S_{ci}}{S_{qi}S_{ci} + \sum_j S_{qj}S_{cj}} \tag{17}$$

And the probability of sampling the correct cue trace given the stimulus and the correct target trace given the correct cue trace was sampled is $\mathrm{P}_S(q|c,Q)\mathrm{P}_S(i|c,q)$. The second equation is the original SAM sampling equation; the first equation is somewhat novel. In essence, the second equation sets the net sampling strength of an item as a function $d()$ of its degree of match to the test cue and its self-strength (controlled by study time and parameter c in SAM), times the strength of the item's binding to the context of test.

We combine degree of match and self-strength into a single function in order to account for differentiation (Shiffrin et al., 1990). The net strength $d\big(m, S(t)\big)$ during stimulus-to-cue sampling, as a function of study time and repetitions (lumped together here as t), depends on the similarity between the item in question and the test stimulus. The exact form of this function is a topic for future study and will involve discerning the relationship between differentiation and m. We can reasonably state from the prior literature some approximate functional forms for typical study times. When the test stimulus is the word, $d(m, S(t)) \sim t$ because in SAM $S_{qq} = ct$, more or less, where c is an encoding rate parameter. When a word is dissimilar from the test stimulus, $d(m, S(t)) \sim t^{-1}$ as was implied by (Shiffrin et al., 1990, fig. 1). Similar words fall on some continuum between these two and the form of the function will do the same. It is likely that, at some level of similarity, $d(m, S(t)) \sim k$, in other words the strength between it and the test stimulus would not increase or decrease from additional study.

As for parameter m: this value is fixed to parameters borrowed from Processing Implicit and Explicit Relations (PIER) models (Nelson et al., 1992, 1998, 2007, 2013). PIER predicts performance in extralist cued recall and item recognition tasks given the semantic relationships between words (as derived from the USF free association norms, Nelson, McEvoy, & Schreiber,

2004). The model is framed as an elaboration of the direct-access component of SAM (Gillund & Shiffrin, 1984; Nelson et al., 1992, p. 334; Raaijmakers & Shiffrin, 1981) and takes a similar functional form:

$$P_S(T'|Q) = \frac{S(Q,T')}{S(Q,T') + S(Q,D)} \tag{18}$$

$$S(Q,T') = S_{qt} + S_{tq} + S_{t'} + E \tag{19}$$

Where $P_S(T'|Q)$ is the probability of sampling the target T' given the test stimulus Q; $P_S(Q,T')$ gives the sampling strength; $S(Q,D)$ gives the net strength of the alternatives; $S_{qt} + S_{tq} + S_{t'}$ are parameters derived from free association norms data; and E is episodic memory strength. Parameter E is not developed (Nelson et al., 2013).

Thus, PIER lays out a detailed representation of how semantic relations drive episodic memory performance but heavily abstracts how learning during a study list influences retrieval. This is the converse of SAM, where performance due to learning the study list is accounted for in detail, but the impact of pre-existing knowledge is abstracted. Integration of these two models would be a natural next step if one wishes to generalize either. This would be entirely acceptable to the PIER theory, which believes the familiarity process to be compatible with it (Nelson et al., 1992, p. 334). It is thus tempting to simply state that $m = (S_{qt} + S_{tq} + S_{t'})$.

Adaptive Character of Thought – Rational (ACT-R)

The retrieval process in ACT-R is governed by sampling from net activations generated by baseline familiarity weighted activations from a cue set. The activation strength of an item A_i is written as:

$$A_i = B_i + \sum_j W_j S_{ji} \tag{20}$$

Where B_i is baseline activation from prior experience (word frequency), S_{ji} is derived from the probability that item i would need to be accessed given test stimulus j in the current context, and W_j weights the degree of activation from S_{ji}. Degree of match M_i is then found by subtracting from A_i a penalty P for task-dependent mismatch; for our purposes this just prevents a trace from sampling itself. The probability of retrieving item i from the set of items k is a weighted and exponentiated Luce choice rule:

$$\Pr(\text{retrieve } i) = \frac{e^{M_i/t}}{\sum_k e^{M_k/t}} \tag{21}$$

Where t is a derived parameter giving information about variance in the memory store. Exponentiating the sum gives ACT-R roughly the same sampling formula as SAM: for the purposes of sampling, net activation is a weighted product of the associations.

In keeping with the two-sample idea, ACT-R can retrieve the cue trace given the test stimulus and then retrieve the target trace given the cue trace. Our reading of the field suggests that this would be relatively unproblematic for the model.

Matrix Model

As it stands, the Matrix Model (Humphreys et al., 1989) has established methods for both extralist and paired associates cued recall. As demonstrated in Appendix A, the mechanism used to extract a target in paired associates cued recall can be written in the form of a single Luce Choice probability. However, the peculiarities of the retrieval procedure complicate the argument and are worthy of note. Interestingly, Matrix model implements extralist cued recall with two sampling processes and paired associates cued recall with one. This is due to the way the memory store is structured in the Humphreys et al. (1989) implementation. As a result, modeling the hybrid cued recall task with this implementation is rather difficult. Therefore, we

find a way to adjust the storage assumptions and retrieval mechanism so that paired associates cued recall samples twice and extralist cued recall once.

Status Quo

In the Matrix model, items are represented as vectors with a mean of 0 and some variance. For stimuli, the variance is fixed such that the dot product of a stimulus $\boldsymbol{s}$ to itself $(\boldsymbol{s} \cdot \boldsymbol{s}) = \|s\|^2 = 1$. The variance of an encoded vector parameterizes encoding strength and thus may depend upon factors such as study time. The presentation of a word pair is represented as the cross-product of the two word vectors. These data points are stored in a single composite memory store. The representation of a list of cues $\boldsymbol{a}$ and targets $\boldsymbol{b}$ in context $\boldsymbol{x}$ is written as

$$\boldsymbol{E} = \sum_i \boldsymbol{x}_i \boldsymbol{a}_i \boldsymbol{b}_i \tag{22}$$

For study events i. Additional study of $\boldsymbol{x}_i \boldsymbol{a}_i \boldsymbol{b}_i$ increases the magnitude of the array. The Matrix model also includes a representation of semantic knowledge $\boldsymbol{S}$ that is also probed during retrieval operations.

The Matrix model completes paired associates cued recall by jointly probing with cue word and study-test context, resulting in an extracted trace that resembles the target. Specifically, the model extracts representation $\boldsymbol{b}'$ of target $\boldsymbol{b}_t$ by multiplying the long-term episodic memory store $\boldsymbol{E}$ by the context and item cues $\boldsymbol{x}_t$ and $\boldsymbol{a}_t$:

$$\boldsymbol{x}_t \boldsymbol{a}_t \boldsymbol{E} = \boldsymbol{b}' = (\boldsymbol{x}_t \cdot \boldsymbol{x}_i)(\boldsymbol{a}_t \cdot \boldsymbol{a}_i)\boldsymbol{b}_i + \sum_{i \neq j} (\boldsymbol{x}_t \cdot \boldsymbol{x}_j)(\boldsymbol{a}_t \cdot \boldsymbol{a}_j)\boldsymbol{b}_j \tag{23}$$

This can be represented as a single Luce choice sampling process when quantifying the "success" of this extraction with the shared variance between correct response $\boldsymbol{b}_i$ and extracted information $\boldsymbol{b}'$ (Appendix A). This equation allows for cue strength, similarity between test stimulus and cue, and cue-target associative strength to factor into the quality of the extracted

target. A further elaboration to allow for intrusions from words semantically related to the test stimulus jointly probes the episodic and semantic stores: $\boldsymbol{x}_t\boldsymbol{a}_t(\boldsymbol{E}+\boldsymbol{S})=\boldsymbol{b}'$.

Extralist cued recall utilizes a different process to retrieve a target. The model first probes the episodic and semantic stores jointly to extract a vector that best matches the test stimulus:

$$(\boldsymbol{r}\boldsymbol{a}_t)(\boldsymbol{E}+\boldsymbol{S})=\boldsymbol{a}' \tag{24}$$

Where $\boldsymbol{r}$ is an empty vector used to preserve dimensionality. The model then jointly probes the extraction $\boldsymbol{a}'$ with context to ascertain its familiarity. Recovery operations occur if $\boldsymbol{a}'$ passes a recognition test, in other words if

$$[(\boldsymbol{x}\boldsymbol{a}_t)\boldsymbol{r}]\cdot(\boldsymbol{E}+\boldsymbol{S})-[\boldsymbol{r}\boldsymbol{a}_t\boldsymbol{r}]\cdot(\boldsymbol{E}+\boldsymbol{S})=\mu>\varepsilon \tag{25}$$

Where μ is a value giving the degree of match and ε gives the minimum necessary match to endorse the item as studied. In other words, the Matrix model finds a to-be-recovered target by sampling then recognizing, and furthermore uses a matrix of semantic associations during the sampling component. The Matrix model completes an extra step and utilizes a novel source of information to complete extralist cued recall (versus paired associates cued recall).

Interestingly, this suggests that the reaction times in extralist cued recall should be greater than that in paired associates cued recall, opposite of what the two-sample framework suggests. In paired associates cued recall, Matrix extracts from the store a word and then transforms that into a verbal response. Extralist cued recall adds an extra step in the middle by requiring recognition of the extracted target first. The two-sample framework would expect the opposite because extralist cued recall requires one sampling step, while paired associates cued recall requires two. This assumes, of course, that a similar recognition process does not occur in paired associates cued recall.

The limited response time data collected in Experiment 1 are consistent with the expectation of the two-sample framework and not the Matrix model: extralist cued recall has faster reaction times than paired associates cued recall, so much so that, in Experiment 1, the largest median RT for a study condition in extralist cued recall is smaller than the smallest median RT in paired associates cued recall. This is consistent with other studies that report RTs for paired associates and extralist cued recall. However, this could be confounded with task difficulty: a harder task can have lower accuracy and larger response times. Whether the prediction that extralist cued recall is slower after controlling for performance holds is a topic for future study, but at present (acknowledging the confounds), the data suggests the opposite is true. This is more in line with the two-sample framework we outline in this paper because more things needs to be done in paired associates cued recall and the hybrid task than in extralist cued recall.

As for the hybrid task in the Matrix Model: A process is not outlined in the literature; we consider two possible but unsatisfactory hybridizations of extralist cued recall and paired associates cued recall. The first is that hybrid cued recall be completed by completing the first sampling phase of extralist cued recall, then using that extracted information to jointly probe with context to yield $\boldsymbol{x}_t\boldsymbol{a}'\boldsymbol{E} = \boldsymbol{b}'$. Combining terms, $\boldsymbol{x}_t\big((\boldsymbol{r}\boldsymbol{a}_t)(\boldsymbol{E}+\boldsymbol{S})\big)\boldsymbol{E} = \boldsymbol{b}'$. This, however, suggests that hybrid cued recall will dissociate from paired associates cued recall. The cue strength gets factored in twice: first by cueing $(\boldsymbol{E}+\boldsymbol{S})$ to get $\boldsymbol{a}'$, and again by probing $\boldsymbol{E}$ for $\boldsymbol{b}'$. The magnitude of the cue strength is therefore squared (give or take) in hybrid cued recall, creating a dissociation.

Alternatively, hybrid cued recall could be completed in the same way as paired associates cued recall. Performance in the task would necessarily be worse than in paired associates cued recall because the test stimulus is dissimilar, reflected in a smaller cosine between test stimulus

and what was extracted from the long-term store than in paired associates cued recall. However, this is theoretically odd given the theory of the extralist cued recall procedure; the use of an extralist probe should require both $\boldsymbol{E}$ and $\boldsymbol{S}$ in both extralist cued recall and hybrid cued recall to complete. However, if this is the case, then splitting the hybrid task into two stages within an experiment— have participants recall the word resembling the extralist test stimulus, then recall the word studied alongside—may be fruitful in determining whether this is so.

Updates to the Model

One solution that is consistent with the proposed framework requires an adjustment to what information is encoded upon study. Here, we borrow a page from TODAM2 (Murdock, 1993) and the Composite Holographic Associative Recall Model (CHARM; Metcalfe, 1990) and say that studying a list of word pairs includes representations of both items and the associations. These are represented by storing the cross product of each item to itself as well as the cross product of the two items to each other:

$$\boldsymbol{E} = \sum_{i} \boldsymbol{x}_i\boldsymbol{a}_i\boldsymbol{a}_i + \boldsymbol{x}_i\boldsymbol{a}_i\boldsymbol{b}_i + \boldsymbol{x}_i\boldsymbol{b}_i\boldsymbol{a}_i + \boldsymbol{x}_i\boldsymbol{b}_i\boldsymbol{b}_i \tag{26}$$

Both $\boldsymbol{xab}$ and $\boldsymbol{xba}$ are included to allow for adirectional or directional associations, as necessary (Kato & Caplan, 2017). The inclusion of $\boldsymbol{xaa}$ and $\boldsymbol{xbb}$ allows $\boldsymbol{xa}$ to probe $\boldsymbol{E}$ and extract a representation of $\boldsymbol{a}$:

$$\boldsymbol{x}_t\boldsymbol{a}_t\boldsymbol{E} = \boldsymbol{a}' = (\boldsymbol{x}_t \cdot \boldsymbol{x}_i)\big((\boldsymbol{a}_t \cdot \boldsymbol{a}_i) + (\boldsymbol{a}_t \cdot \boldsymbol{b}_i)\big)[\boldsymbol{a}_i + \boldsymbol{b}_i] + \text{noise} \tag{27}$$

The extracted information $\boldsymbol{a}'$ gives a representation of the encoded information that best matches $\boldsymbol{a}_t$. This can be transformed into a response to complete an extralist cued recall task. The dot product $\boldsymbol{a}' \cdot \boldsymbol{a}_t$ gives a measure of whether the representation of $\boldsymbol{a}'$ is sufficiently strong to probe for b or to respond with. The same dot product $\boldsymbol{a}' \cdot \boldsymbol{a}_t$ also allows us to inhibit stored

information about $\boldsymbol{a}_t$ in order to efficiently extract $\boldsymbol{b}' \cong \boldsymbol{b}_i$. To find $\boldsymbol{b}'$, we can probe again with $\boldsymbol{a}'$ in context, inhibiting the representation of the cue thusly:

$$\boldsymbol{x}_t\boldsymbol{a}'(\boldsymbol{E} - (\boldsymbol{a}' \cdot \boldsymbol{a}_t)\,\boldsymbol{x}_t\boldsymbol{a}_t\boldsymbol{a}_t) = \boldsymbol{b}' \cong (\boldsymbol{x}_t \cdot \boldsymbol{x}_i)\big((\boldsymbol{a}' \cdot \boldsymbol{a}_i) + (\boldsymbol{a}' \cdot \boldsymbol{b}_i)\big)\boldsymbol{b}_i + \text{noise} \tag{28}$$

The model may then transform $\boldsymbol{b}'$ into a response via some recovery mechanism.

Temporal Context Model (TCM), Context Maintenance and Retrieval 2 (CMR2)

This model class has been applied almost exclusively to free recall tasks; a control process for cued recall has not been defined in publications. Unlike the previous models where we suggest updates to extant control processes, we instead suggest a plausible cued recall process that maintains the spirit of the model class but is consistent with the present findings.

Status Quo

The core of the TCM class of models (Howard & Kahana, 2002; Lohnas et al., 2015; Polyn et al., 2009) is that context information is used to find items and is in turn updated by both retrieved and presented items. The process of learning and retrieval can be described as a cycle of item information updating context and context then activating items, which then updates context, and so on. The model thus has two layers, a feature layer $\boldsymbol{f}$ representing item activations and a context layer $\boldsymbol{c}$ representing temporal context. Activation or updating of one layer given another is mediated by the 2D matrices $\boldsymbol{M}^{FC}$ and $\boldsymbol{M}^{CF}$; these are the memory store. $\boldsymbol{M}^{FC}$ is used to transform the feature layer into the context layer, and $\boldsymbol{M}^{CF}$ is used to activate items on the feature layer given the current context. These transformations are simple matrix multiplications:

$$\boldsymbol{f}_t = \boldsymbol{M}^{CF}\boldsymbol{c}_t \tag{29}$$

$$\boldsymbol{c}_t = \boldsymbol{M}^{FC}\boldsymbol{f}_t \tag{30}$$

It is presumed that there is some pre-existing context $\boldsymbol{c}_{t-1}$ at the beginning of any study or test trial. During study, presentation of items, represented by $\boldsymbol{f}_t$, updates context:

$$\boldsymbol{c}_t = \rho_t \boldsymbol{c}_{t-1} + \beta \boldsymbol{M}^{FC} \boldsymbol{f}_t \tag{31}$$

This updating occurs in this fashion any time anytime the feature layer is activated at study or at test, although parameter values may change between the two test phases. Study also updates the memory store, such that

$$\boldsymbol{M}^{FC}{}_t = M^{FC}{}_{t-1} + \boldsymbol{c}_{t-1} \boldsymbol{f}_t{}^{\mathrm{T}} \tag{32}$$

$$\boldsymbol{M}^{CF}{}_t = M^{CF}{}_{t-1} + \boldsymbol{f}_t \boldsymbol{c}_{t-1}{}^{\mathrm{T}} \tag{33}$$

Context-to-item matrix $\boldsymbol{M}^{CF}$ includes semantic relations: $\boldsymbol{M}^{CF}$ stores the knowledge that some items are often experienced together in the same context over the course of a lifetime.

At retrieval, context activates the feature layer which is then used to output an item. The process by which an item is output given the feature activations differs from version to version of the model class. To conceptually simplify, we use the exponentiated Luce choice rule proposed by TCM. The feature layer is compared to the features representing each word i using the dot product $a_i = \boldsymbol{f}_t \cdot \boldsymbol{f}_i$. The probability of selecting item i among the k possible items is

$$\Pr(\text{output } i) = \frac{e^{a_i}}{\sum_k e^{a_k}} \tag{34}$$

Proposed Extension to Cued Recall

As always, the test trial opens with the current temporal context $\boldsymbol{c}_0$. The test stimulus $\boldsymbol{f}_{stim}$ is presented. To sample for a cue, $\boldsymbol{f}_{stim}$ updates the context $\boldsymbol{c}_0$, which in turn activates items that are or resemble $\boldsymbol{f}_{stim}$ and that were experienced on the study list:

$$\boldsymbol{c}_{stim} = \boldsymbol{M}^{FC} \boldsymbol{f}_{stim} \tag{35}$$

$$\boldsymbol{c}_1 = \rho \boldsymbol{c}_0 + \beta \boldsymbol{c}_{stim} \tag{36}$$

$$\boldsymbol{f}_{cue} = \boldsymbol{M}^{CF}\boldsymbol{c}_1 \tag{37}$$

In extralist cued recall, the contents of $\boldsymbol{f}_{cue}$ are then used to determine what item on the study list to output. The response $\boldsymbol{f}_{out}$ then updates context, $\boldsymbol{c}_2 = \rho\boldsymbol{c}_1 + \beta\boldsymbol{M}^{FC}\boldsymbol{f}_{out}$, which is the new context of test on the following test trial.

In paired associates and hybrid cued recall, the feature activation $\boldsymbol{f}_{cue}$ is instead used to re-probe and resample for a target item. The cue features update context, which in turn activates features representing the target word:

$$\boldsymbol{c}_2 = \rho\boldsymbol{c}_1 + \boldsymbol{M}^{FC}\boldsymbol{f}_{cue} \tag{38}$$

$$\boldsymbol{f}_{target} = \boldsymbol{M}^{CF}\boldsymbol{c}_2 \tag{39}$$

The vector $\boldsymbol{f}_{target}$ is then used to sample for the target item to output via Luce choice after excluding from the possible outcome space the test stimulus and cue word. The response $\boldsymbol{f}_{out}$ then updates context, $\boldsymbol{c}_3 = \rho\boldsymbol{c}_2 + \beta\boldsymbol{M}^{FC}\boldsymbol{f}_{out}$, which is the new context of test on the following test trial.

To see how this generates differential influence of additional cue study and stronger cue-target associations, consider how additional study increases the relative activations given each test stimulus. Additional cue study modifies $\boldsymbol{M}^{FC}$ such that the retrieved context $\boldsymbol{M}^{FC}\boldsymbol{f}_{cue}$ looks more like the context of study-test and $\boldsymbol{M}^{CF}$ such that the context $\boldsymbol{c}_0$ is more likely to activate the cue. However, unless the cue word receives additional study with the target word, this additional activation does not help activate target features. If cue and target are studied together, this makes the contextual profiles of the two words look more like one another (although not necessarily more like the general study-test context), increasing the odds of activating the target features given the cue features. Contextual similarity to study-test and contextual similarity to one another are thus differently influenced by the study manipulation. Extralist cued recall does not

access contextual similarity of one item to another, thus allowing for differential influence of cue study and study task on extralist cued recall versus the other two cued recall tasks.

Appendix A: Quality of Extracted Target in the Matrix Model as a Luce-choice Probability

In this appendix, we demonstrate within the Matrix Model that the variance shared between the target trace and the trace extracted from probing with the cue-context composite is expressible as a Luce-choice probability. The proof relies on the fact that, when $\boldsymbol{y} = \boldsymbol{v_1} + \boldsymbol{v_2} + \boldsymbol{v_3}$ ..., $r^2(\boldsymbol{y}, \boldsymbol{v_1}) = \frac{\mathrm{Var}(\boldsymbol{v_1})}{\mathrm{Var}(\boldsymbol{y})}$ after centering $\boldsymbol{y}$ and $\boldsymbol{v_1}$. This demonstration can be performed in other composite memory models with this as a starting point.

In the Matrix Model (Humphreys, Bain, & Pike, 1989) the long-term episodic store $\boldsymbol{E}$ is a 3-dimensional array representing the sum of N experienced events. Let $\boldsymbol{x}$, $\boldsymbol{a}$, and $\boldsymbol{b}$ represent the study-test context, cue, and target traces respectively:

$$\boldsymbol{E} = \sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{a}_i \boldsymbol{b}_i \tag{A.1}$$

Where the magnitudes of $\boldsymbol{x}_i$, $\boldsymbol{a}_i$, and $\boldsymbol{b}_i$ can be influenced by (among other factors), time studied. In a paired associates cued recall test, the context and cue word are used to probe the store such that an approximation of the target word is extracted:

$$\boldsymbol{x}_i \boldsymbol{a}_i \boldsymbol{E} = \boldsymbol{b}' \cong \boldsymbol{b}_i \tag{A.2}$$

For the i^{th} event. Expanding on $\boldsymbol{b}'$:

$$\boldsymbol{x_i a_i E} = \boldsymbol{b}' = (\boldsymbol{x}_i \cdot \boldsymbol{x}_i)(\boldsymbol{a}_i \cdot \boldsymbol{a}_i)\boldsymbol{b}_i + \sum_{i \neq j} (\boldsymbol{x}_i \cdot \boldsymbol{x}_j)(\boldsymbol{a}_i \cdot \boldsymbol{a}_j)\boldsymbol{b}_j \tag{A.3}$$

We want to know the fraction of variance shared, or r^2, between $\boldsymbol{b}'$ and $\boldsymbol{b}$. A perfectly extracted trace will have $r^2 = 1$. Observe that $\boldsymbol{b}'$ is effectively a sum of the target vectors weighted by the encoding strength of the cue and context and the quality of the match between them. If we treat the product of the dot products as random variables, then the variance of $\boldsymbol{b}'$ is the sum of the covariances of each i^{th} and j^{th} experiences times their encoding strengths. Let $\mathrm{M}(i, j) =$

$(\boldsymbol{x}_i \cdot \boldsymbol{x}_j)(\boldsymbol{a}_i \cdot \boldsymbol{a}_j)$ be the product of the dot products of the cue-context probe i and cue-context representation j and $\mathrm{Cov}(i,j)$ be the covariance between target vectors i and j:

$$\mathrm{Var}(\boldsymbol{b}') = \sum_{j,k} \mathrm{M}(i,j)\mathrm{M}(i,k)\mathrm{Cov}(\boldsymbol{b}_j, \boldsymbol{b}_k) \tag{A.4}$$

The absolute variance contributed by the correct item i is sum of the covariance terms between i and the other items. Separating the terms stemming from the correct experience i,

$$\mathrm{Var}(\boldsymbol{b}') = \sum_{j} \mathrm{M}(i,i)\mathrm{M}(i,j)\mathrm{Cov}(\boldsymbol{b}_i, \boldsymbol{b}_j) + \sum_{j\neq i, k\neq i} \mathrm{M}(i,j)\mathrm{M}(i,k)\mathrm{Cov}(\boldsymbol{b}_j, \boldsymbol{b}_k) \tag{A.5}$$

The shared variance r^2 between two vectors $\boldsymbol{b}'$ and $\boldsymbol{b}$ is the fraction of the variance of $\boldsymbol{b}'$ contributed by $\boldsymbol{b}$. That contribution is the i^{th} term from (A.5), the shared variance is the proportion of total variance from that term:

$$r^2(\boldsymbol{b}', \boldsymbol{b}_i) = \frac{\sum_j \mathrm{M}(i,i)\mathrm{M}(i,j)\mathrm{Cov}(\boldsymbol{b}_i, \boldsymbol{b}_j)}{\sum_j \mathrm{M}(i,i)\mathrm{M}(i,j)\mathrm{Cov}(\boldsymbol{b}_i, \boldsymbol{b}_j) + \sum_{j\neq i, k\neq i} \mathrm{M}(i,j)\mathrm{M}(i,k)\mathrm{Cov}(\boldsymbol{b}_j, \boldsymbol{b}_k)} \tag{A.6}$$

If we define $\mathrm{S}(n) \equiv \sum_{j,k} \mathrm{M}(n,j)\mathrm{M}(n,k)\mathrm{Cov}(\boldsymbol{b}_j, \boldsymbol{b}_k)$ we can condense the fraction to:

$$r^2(\boldsymbol{b}', \boldsymbol{b}_i) = \frac{\mathrm{S}(i)}{\sum_j \mathrm{S}(j)} \tag{A.7}$$

In other words, the variance explained between the extracted trace and the correct trace is equivalent to the unweighted Luce-choice probability of sampling i based upon $\mathrm{S}(i)$.

By establishing that the shared variance is expressible as a Luce choice probability, we demonstrate competitiveness and that $\mathrm{S}(i)$ is an amalgamation of information. Let's consider the term $\mathrm{M}(i,j) = (\boldsymbol{x}_i \cdot \boldsymbol{x}_j)(\boldsymbol{a}_i \cdot \boldsymbol{a}_j)$. If the i^{th} vector is the test stimulus, then the vectors j are representations of encoded events with strength equal to the mean of the vector. We can rewrite this term to $\mathrm{M}(i,j) = \|\boldsymbol{x}_i\|\|\boldsymbol{x}_j\| \cos\theta_x \, \|\boldsymbol{a}_i\|\|\boldsymbol{a}_j\| \cos\theta_a$ where $\|\boldsymbol{x}\|$ gives a magnitude of vector $\boldsymbol{x}$ and $\cos\theta$ gives the similarity between vectors i and j. The model sets the magnitude of

stimuli to be constant, let that magnitude be equal to one. Then $\mathrm{M}(i,j) = \|\boldsymbol{x}_j\| \cos\theta_x \|\boldsymbol{a}_j\| \cos\theta_a$ and $\mathrm{M}(i,i) = 1$. Additional study time increases magnitude; similarity is a property of the stimulus. We can therefore say that additional study time of a cue word studied alongside the target increases the quality of the extracted target representation. One property of this is that the co-occurrence of both items together increases the probability of sampling the target more than studying either item separately.

Appendix B: Output Interference Analysis of Experiment 1

Output Interference

Output interference, the observation that performance generally worsens as a function of test trial or output position, is a noteworthy phenomenon in multiple tasks (Criss, Malmberg, et al., 2011; Roediger, 1973; Tulving & Arbuckle, 1963). Accounting for the decline in correct response rate over test trial in cued recall, free recall, and category-cued free recall of categorized lists was one of the early successes of the SAM model (Raaijmakers & Shffrin, 1981). Although reports and accounts of paired associates cued recall output interference can be found in the literature (Raaijmakers & Shffrin, 1981; Roediger & Adelson, 1980; Tulving & Arbuckle, 1963, 1966; Wilson et al., 2020), the same cannot be said for extralist cued recall or hybrid cued recall.

Here, we analyze the Experiment 1 data, considering changes in the rates of correct responses, intrusions, and response failures as a function of test bin in paired associates, extralist, and hybrid cued recall. In paired associates cued recall, output interference takes the form of a decline in correct responses and intrusions and an increase in response failures over test (Wilson et al., 2020). In SAM, output interference is driven by learning at test and a response filter that prevents words from being recalled twice in a test list. The three kinds of cued recall considered here differ somewhat in what kind of information is used to retrieve. Deviations in this pattern may indicate what kind of learning at test drives changes with test bin or what information is used to filter responses.

Analysis

We use Bayesian order-constrained inference to analyze change in correct responses, intrusions, and response failures as a function of test bin (6 bins, 3 trials each) for each cued

recall test (Figure B1). Output interference instantiates as an increase in response failures and a decrease in correct responses and intrusions over response bin, although sometimes intrusions remain flat (Wilson et al., 2020). We therefore test a model where response failures increase and correct responses and intrusions both decrease with test bin. We accept Bayes Factors (BFs) greater than 10 or less than 1/10 as meaningful evidence for or against the hypothesized pattern. The prior probability of the model is analytically complex and so we instead simulate the prior ($N = 10^7$ samples). Overall, we find evidence for output interference in all three tasks (extralist: $BF = 700$, paired associates: $BF = 4*10^5$, hybrid: $BF = 30$).

Discussion

We observe output interference in the same form generally across the three tasks. According to SAM, output interference in cued recall is driven by both learning at test and a process by which double recalls are inhibited (Raaijmakers & Shffrin, 1981; Wilson et al., 2020). The later component—the retrieval filter—should be general to all three tasks as all three recover items. However, it need not have been the case that learning during test would play the same role in the tasks—if output interference was absent in extralist cued recall this would have indicated that what was learned during test played a role in the cue-to-target sampling phase and not stimulus-to-cue sampling. The generality of the pattern we observe suggests that output interference plays a role in stimulus-to-cue sampling and recovery. This is consistent with the explanation that output interference occurs because item information is confusable (Malmberg et al., 2012).

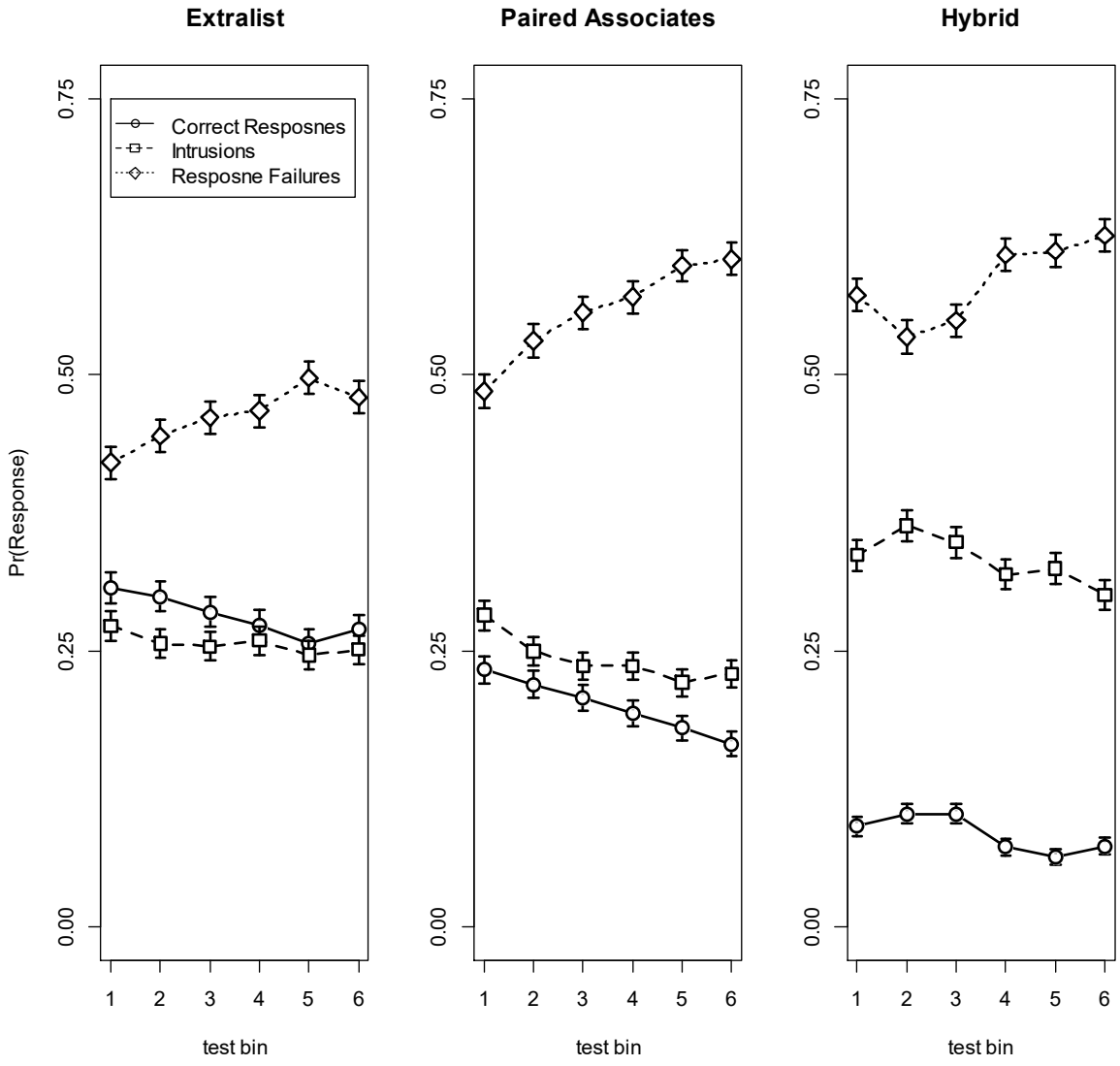

Figure B.1. Response rates by test bin. Means +/- 1 SEM computed under binomial assumptions. Left: extralist cued recall. Center: paired associates cued recall. Right: hybrid cued recall. Circles: correct response rates. Squares: intrusion rates. Diamonds: response failure rates.

Appendix C: Tests of Encoding Strength as a Function of Confusability

We argue that the pattern of data observed in Experiment 2 cannot be fully explained by confusable items receiving additional study, as might be predicted by an extension of the proactive interference and facilitation paradigm. Here, we directly test for effects of similarity on pair encoding through conceptual replications of Experiment 2 that test memory using item recognition.

Experiment 3

The purpose of this Experiment is to test whether confusability impacts pair strength. To that end, we manipulate, within subject and between list, whether or not a pair member is confusable, and test memory with yes-no single item recognition.

Participants

$N = 45$ undergraduates from Syracuse University participated for course credit.

Materials & Design

Words used in this experiment were drawn from the same word pool as in Experiment 2. Participants studied and were tested on four lists of word pairs. Two lists of 36 word pairs consisted of distinct words. On the other two lists of 36 word pairs, one member of the pair was distinct, and the other pair confusable. The list was set up such that the two confusable words were on opposite halves of the study list. Each pair was presented for 2.5s with a 0.5s interstimulus interval. Participants were instructed to place the words in each pair into a sentence, as done in the prior two experiments. Memory was tested after each list following a 60s running addition distractor task.

Participants' memory for distinct words was tested using yes-no single item recognition. The test consisted of 36 target words—one distinct word from each word pair—and 36 distinct

unstudied foils from the same pool of distinct words. Between 0.5 s interstimulus intervals, participants would be shown each word on the screen and instructed to identify whether it was on the most recent study list. Between the testing phase of one block and the study phase of the next, participants were informed that they would not study or be tested on those words again and that they should try to forget them.

Results

We analyzed both hits and false alarms in Bayesian and frequentist 2 (old vs new) x 2 (distinct + distinct vs. distinct + confusable) repeated measures ANOVAs. We observed a mirror effect: distinct words studied alongside other distinct words were associated with greater hit rates and lower false alarm rates than distinct words studied alongside confusable words, $F(1, 44) = 18.1$, $p < .001$, $BF_{exclude} = 0.236$. Distinct words and their foils were equally likely, in aggregate, to be endorsed as old on either list, $F(1, 44) < 1$, $BF_{exclude} = 5.64$.

Discussion

In all, distinct words studied alongside confusable words are less well-remembered than those studied alongside other distinct words. We can rule out the possibility that confusable items are prompting stronger encoding of the entire pair—both cue and target item. Were whole pairs being strengthened, we might expect increased discriminability for distinct items studied alongside confusable words than otherwise. We observe the opposite effect. In all, we can say that it is insufficient to account for the patterns of data we have observed.

There are, in all, three plausible accounts for this recognition data. First, introduction of confusable items leads to the creation of "triads" in the long-term episodic memory store that include the distinct item and both confusable items creating, effectively, a list length effect. Second, paring a distinct item with a confusable item selectively weakens the encoding of the

distinct items, and the criterion used during testing moved in a compensatory fashion. The third possibility is that, by some means, pairing distinct items with one another leads to stronger encoding of both items than when a distinct item is paired with a confusable item—either distinctiveness drives additional item encoding or confusability inhibits item encoding across the board. In Experiment 4, we assess these accounts.

Experiment 4

In this experiment, we use the same manipulation, but within list. The triad formation and increased strength for confusable items explanations both predict no within-list effects. If a triad forms, then the number of items in the representation of the study list grows. In other words, triads increase list length. This explanation predicts no within-list effects of confusability because two items on the same list are on lists of the same length. Strengthened encoding for confusable items does not alter the degree of encoding for distinct items, so no within list effects are predicted by this explanation, either. Selective weakening of items studied alongside confusable words would result in the same outcome within list as observed between list. Thus, a replication of the effect in Experiment 3 would be consistent with selective weakening, while the elimination of the effect would be consistent with the other two explanations.

Participants

$N = 58$ undergraduate students from Syracuse University completed the experiment for course credit.

Materials & Design

Participants studied and were tested upon four lists of 36 word pairs. Half of the word pairs had distinct members. The other half contained a word that was confusable with another word studied on the list.

All other details are the same as Experiment 3. The list was set up such that the two confusable words were on opposite halves of the study list. Each pair was presented for 2.5 s with a 0.5 s interstimulus interval. Participants were instructed to place the words in each pair into a sentence, as done in the prior three experiments. Memory was tested after each list following a 60 s running addition distractor task. Participants' memory for distinct words was tested using yes-no single item recognition. The test consisted of 36 target words—one distinct word from each word pair—and 36 distinct unstudied foils from the same pool of distinct words. Between 0.5 s interstimulus intervals, participants would be shown each word on the screen and instructed to identify whether it was on the most recent study list. Between the testing phase of one block and the study phase of the next, participants were informed that they would not study or be tested on those words again and that they should try to forget them.

Results

We analyzed the hit rates for the two conditions in Bayesian and frequentist paired samples *t* tests. The false alarm rate (M ± SEM = .248 ± .020) was not analyzed because it is impossible in this paradigm to distinguish which false alarms "belong" to one condition or another. The hit rate for distinct items was unchanged by the presence of a confusable word in the pair, $t(57) = 0.524$, $p = .602$, $BF_{01} = 6.11$.

Discussion

If confusable words impacted, within list, the strength of distinct words, it would have been observed here. We found that it did not impact performance, evidence against strength differences. Between these findings and those of Experiment 3, we can conclude that, if the presence of confusable words leads to a difference in distinct word encoding strength, it must apply across the entire study list.

Discussion of Experiments 3 and 4

In all, two possible accounts exist for this data. First, introduction of confusable items leads to the creation of "triads" in the long-term store that include the distinct item and both confusable items, effectively creating a list length effect. The second is that, by some means, pairing distinct items with one another leads to stronger encoding of both items than when a distinct item is paired with a confusable item across the entire study list—either distinctiveness drives additional item encoding or confusability inhibits item encoding across the board. This second account lacks theoretical backing and, in either case, would generate effects between- rather than within-list.

The concept of forming of triads comes from the proactive interference and facilitation literature. Under the explanation, the similar words are encoded twice in separate traces, once upon presentation, and again when the similar word is presented (with the similar word being encoded alongside the first word as well). The dissimilar targets are not re-encoded in this account: if they were, we would see strengthening, not weakening, of the target items. This creates something akin to a list-length effect. In Experiment 3, when distinct words were paired with confusable words in an entire list, it would be as if 108 words were studied, not 72. Implementation of this in the REM model produces a small mirror effect in the correct direction, although the effect it predicts is smaller than the mirror effect we see in the data. In either case, the list length effect manipulations occur between lists or between subjects. It is not a within list effect. It therefore cannot explain the Experiment 2 data.

Appendix D: A REM Model of the Free Association Task

The free association task implementation we use is simple. We assume that the lexical-semantic store includes 17,000 words (D'Anna et al., 1991; Goulden et al., 1990), each with 20 features generated randomly from the geometric distribution with parameter $g = .4$. Upon presentation of a test stimulus i, the model computes the λ for the test stimulus to each of the 16,999 possible responses k and responds with item j probabilistically:

$$\Pr(j|i) = \frac{\lambda_{ij}}{\sum_k \lambda_{ik}} \tag{B.1}$$

The comparisons λ are made with parameters $g = .4$ and $c = .7$. All parameters operating within the model are either fixed either by convention (g, c, number of features per word) constrained to values observed in the literature (number of words known). Thus, any degree of similarity between two stimuli is a product of random chance. However, because each word is compared to 16,999 others any given word will likely be similar to some other words.

To compare the outcomes of this model with the USF free association norms data, we ran the model approximately 149 times per word (the exact number of time is taken from of the samples in the USF free association norms) for the 4935 of the probes (which is the number of probes in the norms with logHAL entries in the English Lexicon Project (Balota et al., 2007)). Our critical points of comparison between the model and the data were the effects of word frequency, set size, and response entropy. We chose word frequency because implementation of word frequency in REM has been developed. We chose set size and entropy because they are distributional properties REM is capable of simulating.

To compare effects of word frequency, we took the maximum likelihood estimate of the g parameter for each stimulus and treated that as the frequency of the item. Response frequency is simply the average frequency of the response set (observed or simulated) given a test stimulus.

The simulations are remarkably consistent with the norms (Figures D.1 & D.2). Responses are generally of higher frequency than their probes and by roughly the same amount, and the correlation coefficient between test stimulus frequency and average response frequency is equivalent (r = .58 in both simulations and observations).

To measure forward set size, we counted the number of unique responses that were elicited by a given test stimulus. To measure backward set size, we counted the number of probes presented that elicited a given response. Entropy was computed in an analogous fashion: forward entropy is computed with the probabilities of responding given a test stimulus, and backward entropy is computed from the probability that a test stimulus was presented given a response.

Finally, we compared the total set of biserial correlations between test stimulus word frequency, response word frequency, forward entropy / set size, and backward entropy / set size that could be computed from the model and from the USF norms (Tables D.1, D.2). The model captures the properties within the norms that can be attributed to components of the test stimulus (correlations between test stimulus word frequency and the word frequency, entropy, and set size of the response word) remarkably well, considering the lack of free parameters in the model. This suggests that some higher-level semantic properties of stimuli, pre-experimental set size, can naturally arise from the low-level properties of the stimuli in the set.

The model is not capturing the properties within the norms related to the target. This could be expected given the simplicity of the model design. Most semantic models attempt to fit within a network the empirical relationships between specific words, which we do not do here. The major takeaway from this simple model is that the relationship between test stimulus word frequency and other semantic properties can naturally arise from REM without attempting to fit

stimuli into a known structure. To account for the target relationships, two other factors must first be accounted for. The method by which the cue set is selected should be synchronized between model and norms. Further, a more elaborate generative procedure for the list of stimuli is in order.

There is a mismatch between how probes are selected in the free association norms and how probes are selected in these simulations. The REM simulations offered here selected test stimulus words randomly, in that the semantic content of the probes was a byproduct of their randomly sampled features. In the Nelson, et al. (2004) norms, these probes were chosen *ad hoc* as interest in some combination of words or semantic structures arose. Their sample of probes is non-random. This is critical for reproduction of the backward properties because these are conditionalized upon the set of probes selected to be used during the norms. Given a response, what are the properties of the probes? The answer to this question depends upon what test stimulus words have been chosen by the experimenters. It would be difficult or impossible to reproduce the selection process Nelson et al used to decide test stimulus sets here. Alternatively, this could be resolved by renorming with a truly random sample of test stimului.

Secondly, this first pass at free association does not put much thought into the generative process for stimuli. Semantic networks derived from text-based copora possess several mathematical properties. They are, for example, “small world” networks, in that the average path length (the minimum number of jumps needed to go from one word to another) is short, usually less than 5 (Griffiths et al., 2007). Such a network can be generated via one of several algorithms. In the model, our stimuli are independent random samples from the geometric distribution. Suffice it to say, randomly sampling from a geometric distribution does not produce the kind of similarity space that is described as a small-world network. It is also unlikely that the

configural properties of real-world words can be accurately described as independent random samples; some process led to their arrangement in the semantic space. Implementing a process for procedural generation of stimuli may help account for the mismatch in the backward set sizes and entropies.

Table D.1.

Bivariate correlations of measures from the USF free association norms and REM

			Pearson Product-Moment Correlation r					
V.	*Measure*		1	2	3	4	5	6
1	*Word Frequency*	*Probe*	---	0.576	0.131	0.557	0.165	0.433
2		*Target*	0.574	---	-0.084	0.183	-0.043	0.149
3	*Entropy*	*Forward*	0.127	-0.317	---	0.132	0.913	0.038
4		*Backward*	0.591	-0.004	-0.011	---	0.148	0.748
5	*Set Size*	*Forward*	0.160	-0.252	0.893	0.003	---	0.076
6		*Backward*	0.098	-0.132	-0.017	0.833	0.054	---

Note: Upper triangle: correlations from the USF norms, $N = 4936$. Lower triangle: correlations from the model, $N = 4872$.

Table D.2.

Difference in bivariate correlations of measures from the USF free association norms and REM

			Difference in Pearson Product-Moment Correlation r				
V.	*Measure*		1	2	3	4	5
1	*Word Frequency*	*Probe*	---				
2		*Target*	0.003	---			
3	*Entropy*	*Forward*	0.004	0.234***	---		
4		*Backward*	-0.035**	0.187***	0.144***	---	
5	*Set Size*	*Forward*	0.006	0.209***	0.021***	0.146***	---
6		*Backward*	0.336***	0.281***	0.055**	-0.085***	0.022

Note: Difference in correlation coefficients *r*, data - model. Significant differences from Fisher's r-to-z transformation: *p<.05, **p<.01, ***p<.001

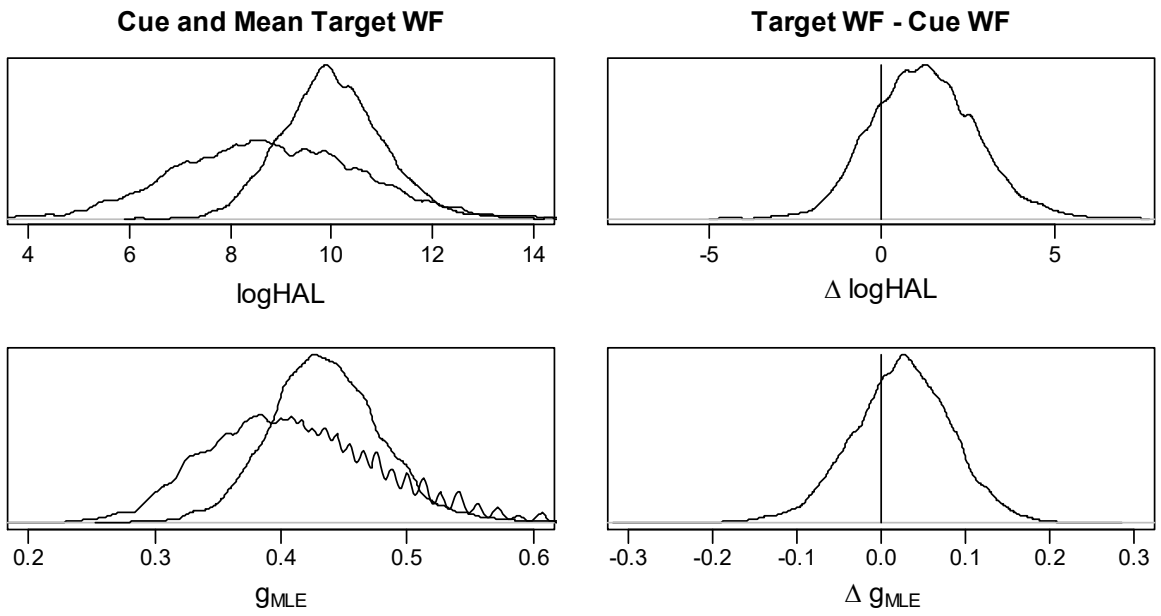

Figure D.1: Observed versus simulated word frequencies (WF) for cues and their average targets in free association. Cues and targets without entries in the English Lexicon Project database were excluded, $N = 4935$ cues.

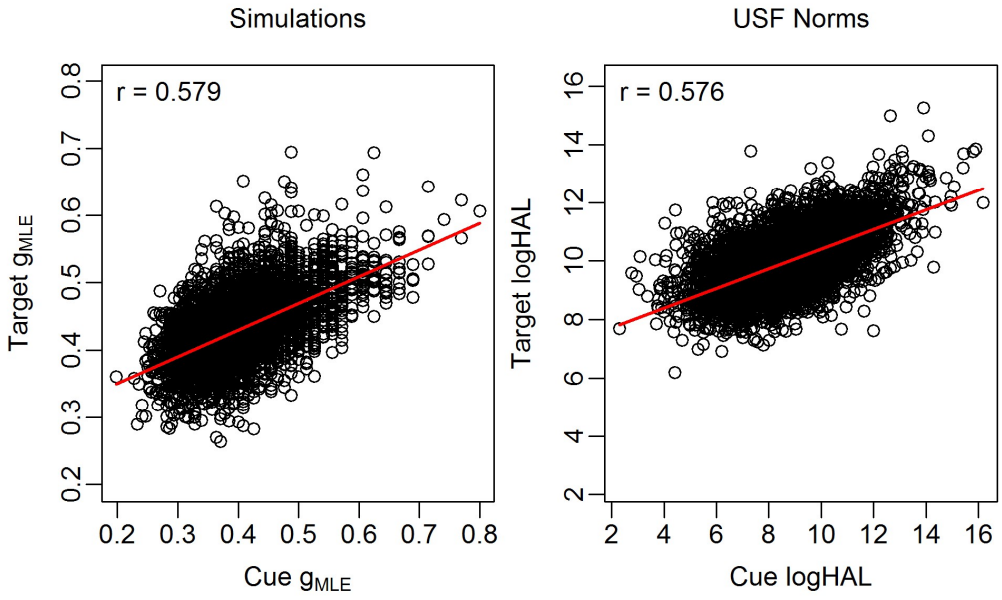

Figure D.2: Observed versus simulated word frequencies (WF) for cues and their average targets in free association. Cues and targets without entries in the English Lexicon Project database were excluded, $N = 4935$ cues.

References

Anderson, J. A., Sliverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review*, *84*(1), 415–451.

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review*, *79*(2), 97–123.

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(5), 1063–1087.

Atkinson, R. C., & Shiffrin, R. M. (1965). Mathematical models for memory and learning. *Technical Report No. 79*.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human Memory: A proposed system and its control processes. In *Psychology of Learning and Motivation—Advances in Research and Theory* (Vol. 2, pp. 89–195).

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

Bahrick, H. P. (1970). Two-phase model for prompted recall. *Psychological Review*, *77*(3), 215–222.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchinson, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*, 445–459.

Brainerd, C. J., Reyna, V. F., Wright, R., & Mojardin, A. H. (2003). Recollection rejection: False-memory editing in children and adults. *Psychological Review*, *110*(4), 762–784.

Brysbaert, M., & New, B. (2009). Moving Beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990.

Criss, A. H., Aue, W. R., & Smith, L. (2011). The effects of word frequency and context variability in cued recall. *Journal of Memory and Language*, *64*, 119–132.

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

D'Anna, C. A., Zechmeister, E. B., & Hall, J. W. (1991). Toward a meaningful definition of vocabulary size. *Journal of Reading Behavior*, *23*(1).

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

Dunn, J. C. (2008). The dimensionality of the remember-know task: A state-trace analysis. *Psychological Review*, *115*(2), 426–446.

Dunn, J. C., & Kalish, M. L. (2018). *State-Trace Analysis* (1st ed.). Springer International Publishing.

Ebbinghaus, H. (1895). *Memory: A Contribution to Experimental*.

Feigenbaum, E. A. (1966). Information processing and memory. *Proceedings of the Fifth Berkley Symposium on Mathematical Statistics and Probability*, *IV*.

Feigenbaum, E. A., & Simon, H. A. (1962). Simulation of human verbal learning behavior. *Communications of the ACM*, 223.

Gardiner, J. M. (1988). Recognition failures and free-recall failures: Implications for the relation between recall and recognition. *Memory & Cognition*, *16*(5), 446–451.

Gillund, G., & Shiffrin, R. M. (1981). Free recall of complex pictures and abstract words. *Journal of Verbal Learning and Verbal Behavior*, *20*, 575–592.

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

Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics*, *11*(4), 341–363.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244. Scopus. https://doi.org/10.1037/0033-295X.114.2.211

Hintzman, D. L. (1968). Explorations with a discrimination net model for paired-associate learning. *Journal of Mathematical Psychology*, *5*(1), 123–162.

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

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*, 2554–2558.

Howard, M., & Kahana, M. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*(3), 269–299.

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

Jacoby, L. L. (1998). Invariance in automatic influences in memory: Toward a user's guide for the process dissociation procedure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(1), 3–26.

JASP Team. (2020). *JASP* (0.14.1) [Computer software]. https://jasp-stats.org

Jones, M. N., Gruenenfelder, T. M., & Recchia, G. (2018). In defense of spatial models of semantic representation. *New Ideas in Psychology*, *50*, 54–60. Scopus. https://doi.org/10.1016/j.newideapsych.2017.08.001

Kato, K., & Caplan, J. B. (2017). Order of items within associations. *Journal of Memory and Language*, *97*, 81–102.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *3*, 490–517.

Kucera, H., & Francis, W. N. (1967). *Computational Analysis of Present-day American English*. Brown University press.

Lehman, M., & Malmberg, K. J. (2013). A buffer model of memory encoding and temporal correlations in retrieval. *Psychological Review*, *120*(1), 155–189.

Lohnas, L. J., Polyn, S. M., & Kahana, M. J. (2015). Expanding the scope of memory search: Modeling intralist and interlist effects in free recall. *Psychological Review*, *122*(2), 337–363. https://doi.org/10.1037/a0039036

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. Wiley.

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

Malmberg, K. J., Criss, A. H., Gangwani, T., & Shiffrin, R. M. (2012). Overcoming the negative consequences of interference that results from recognition memory testing. *Psychological Science*, *23*, 115–119.

Malmberg, K. J., Lehman, M., Annis, J., Criss, A. H., & Shiffrin, R. M. (2014). Consequences of testing memory. In *Psychology of Learning and Motivation—Advances in Research and Theory* (Vol. 61, pp. 285–313).

Malmberg, K. J., & Shiffrin, R. M. (2005). The "one-shot" hypothesis for context storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 322–336.

McEvoy, C. L., & Holley, P. E. (1990). Aging and the stability of activation and sampling in cued recall. *Psychology and Aging*, *5*(4), 589–596.

McEvoy, C. L., & Nelson, D. L. (1990). Selective access in cued recall: The roles of retrieval cues and domains of encoding. *Memory & Cognition*, *18*(1), 15–22.

Medin, D. L. (1975). A theory of context in discrimination learning. *Psychology of Learning and Motivation*, *9*, 263–314.

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

Mulligan, N. W. (2012). Differentiating between conceptual implicit and explicit memory: A crossed double dissociation between category-exemplar production and category-cued recall. *Psychological Science*, *23*(4), 404–406.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, *89*(6), 609–626.

Murdock, B. B. (1993). TODAM2: A model for the storage and retrieval of item, associative, and serial-order information. *Psychological Review*, *100*(2), 183–203. https://doi.org/10.1037/0033-295X.100.2.183

Nelson, D. L., Bajo, M. T., & Cañas, J. (1987). Prior knowledge and memory: The episodic encoding of implicitly activated associates and rhymes. *Journal of Experimental Psychology: Leaning, Memory, and Cognition*, *13*(1), 54–63.

Nelson, D. L., Bennett, D. J., Gee, N. R., Schreiber, T. A., & McKinney, V. M. (1993). Implicit memory: Effects of network size and interconnectivity on cued recall. *Journal of Experimental Psychology: Human Learning & Memory*, *19*(4), 747–764.

Nelson, D. L., Cañas, J., & Bajo, M. T. (1987). The effects of natural category size on memory for episodic encodings. *Memory & Cognition*, *15*(2), 133–140.

Nelson, D. L., & Friedrich, M. A. (1980). Encoding and cuing sounds and senses. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(6), 717–731. Scopus. https://doi.org/10.1037/0278-7393.6.6.717

Nelson, D. L., & Goodmon, L. B. (2002). Experiencing a word can prime its accessibility and its associative connections to related words. *Memory & Cognition*, *30*(3), 380–398.

Nelson, D. L., Goodmon, L. B., & Ceo, D. (2007). How does delated testing reduce effects of implicit memory: Context confusion or cuing with context? *Memory & Cognition*, *35*(5), 1014–1023.

Nelson, D. L., Kitto, K., Galea, D., McEvoy, C. L., & Bruza, P. D. (2013). How activation, entanglement, and searching a semantic network contribute to event memory. *Memory & Cognition*, *41*(6), 797–819.

Nelson, D. L., & McEvoy, C. L. (1979). Encoding context and set size. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(3), 292–314.

Nelson, D. L., & McEvoy, C. L. (1984). Word fragments as retrieval cues: Letter generation or search through memory? *American Journal of Psychology*, *97*(1), 17–36.

Nelson, D. L., & McEvoy, C. L. (2000). What is this thing called frequency? *Memory & Cognition*, *28*(4), 509–522.

Nelson, D. L., McEvoy, C. L., & Bajo, M. T. (1984). Retrieval processes in perceptual recognition and cued recall: The influence of category size. *Memory & Cognition*, *12*, 498–506.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1990). Encoding context and retrieval conditions as determinants of the effects of natural category size. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(1), 31–41.

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

Nelson, D. L., McKinney, V. M., Gee, N. R., & Janczura, G. A. (1998). Interpreting the influence of implicitly activated memories on recall and recognition. *Psychological Review*, *105*(2), 299–324.

Nelson, D. L., Schreiber, T. A., & McEvoy, C. L. (1992). Processing implicit and explicit relations. *Psychological Review*, *99*(2), 322–348.

Nelson, D. L., & Zhang, N. (2000). The ties that bind what is known to the recall of what is new. *Psychonomic Bulletin & Review*, *7*(4), 204–617.

Nelson, T. O., & Narens, L. (1980). Norms of 300 general-information questions: Accuracy of recall, latency of recall, and feeling-of-knowing ratings. *Journal of Verbal Learning and Verbal Behavior*, *19*(3), 338–368. Scopus. https://doi.org/10.1016/S0022-5371(80)90266-2

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 104–114.

Perfect, T. J. (2004). The role of self-rated ability in the accuracy of confidence judgements in eyewitness memory and general knowledge. *Applied Cognitive Psychology*, *18*(2), 157–168. Scopus. https://doi.org/10.1002/acp.952

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, *116*(1), 129–156. https://doi.org/10.1037/a0014420

Postman, L. (1975). Tests of the generality of the principle of encoding specificity. *Memory & Cognition*, *3*(6), 663–672.

Raaijmakers, J. G., & Shffrin, R. M. (1981). Order effects in recall. *Attention and Performance*, *9*, 403–415.

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

Reifer, D. M., & Rouder, J. N. (1992). A multinomial modeling analysis of the mnemonic benefits of bizzare imagery. *Memory & Cognition*, *60*(6), 601–611.

Roediger, H. L. (1973). Inhibition in recall from cueing with recall targets. *Journal of Verbal Learning and Verbal Behavior*, *12*, 644–657.

Roediger, H. L., & Adelson, B. (1980). Semantic specificity in cued recall. *Memory & Cognition*, *8*(1), 65–74.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255.

Ross, B. H., & Bower, G. H. (1981). Comparison of models of associative recall. *Memory & Cognition*, *9*(1), 1–16.

Schacter, D. L., Israel, L., & Racine, C. (1999). Suppressing False Recognition in Younger and Older Adults: The Distinctiveness Heuristic. *Journal of Memory and Language*, *40*(1), 1–24. Scopus. https://doi.org/10.1006/jmla.1998.2611

Schweickert, R. (1993). A multinomial processing tree model for degradation and redintegration in immediate recall. *Memory & Cognition*, *21*(2), 168–175.

Shiffrin, R. M. (1970). Forgetting: Trace erosion or retrieval failure? *Science*, *168*, 1601–1603.

Shiffrin, R. M. (1973). Visual free recall. *Science*, *180*(4089), 980–982.

Shiffrin, R. M., & Atkinson, R. C. (1969). Storage and retrieval processes in long-term memory. *Psychological Review*, *76*(2), 179–193.

Shiffrin, R. M., Ratcliff, R., & Clark, S. E. (1990). List-strength effect II: Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(2), 179–195.

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

Shiffrin, R. M., & Steyvers, M. (1998). The effectiveness of retrieval from memory. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 73–95). Oxford University Press.

Smith, A. D. (1971). Output interference and organized recall from long-term memory. *Journal of Verbal Learning and Verbal Behavior*, *10*, 400–408.

Sternberg, S. (1966). High-speed scanning in human memory. *Science*, *153*, 652–654.

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2005). Word Association Spaces for Predicting Semantic Similarity Effects in Episodic Memory. In *Experimental cognitive psychology and its applications* (pp. 237–249). American Psychological Association. https://doi.org/10.1037/10895-018

Tulving, E. (1972). *Episodic and semantic memory* (E. Tulving & W. Donaldson, Eds.). Academic Press.

Tulving, E. (1974). Recall and recognition of semantically encoded words. *Journal of Experimental Psychology*, *102*(5), 778–787.

Tulving, E., & Arbuckle, T. Y. (1963). Sources of intertrial interference in immediate recall of paired associates. *Journal of Verbal Learning and Verbal Behavior*, *1*, 321–334.

Tulving, E., & Arbuckle, T. Y. (1966). Input and output interference in short-term associative memory. *Journal of Experimental Psychology*, *72*(1), 145–150.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*(5), 352–373.

Tulving, E., & Wiseman, S. (1975). Relation between recognition and recognition failure of recallable words. *Bulletin of the Psychonomic Society*, *6*(1), 79–82.

Watkins, M. J., & Tulving, E. (1975). Episodic memory: When recognition fails. *Journal of Experimental Psychology: General*, *104*(1), 5–29.

Wilson, J. H., & Criss, A. H. (2017). The list strength effect in cued recall. *Journal of Memory and Language*, *95*, 78–88.

Wilson, J. H., Kellen, D., & Criss, A. H. (2020). Mechanisms of output interference in cued recall. *Memory & Cognition*, *48*(1), 51–68.

Wiseman, S., & Tulving, E. (1975). A test of confusion theory of encoding specificity. *Journal of Verbal Learning and Verbal Behavior*, *14*, 370–381.

Yntema, D. B., & Trask, F. P. (1963). Recall as a search process. *Journal of Verbal Learning and Verbal Behavior*, *2*, 65–74.

Zeelenberg, R., Pecher, D., Shiffrin, R. M., & Raaijmakers, J. G. W. (2003). Semantic context effects and priming in word association. *Psychonomic Bulletin & Review*, *10*(3), 653–660.

CURRICULUM VITAE

Jack Wilson

Syracuse University

Department of Psychology

jhwilson@g.syr.edu

Education

Doctoral Candidate Cognition, Brain, and Behavior Program, Syracuse University

Advisor: Amy H. Criss, Ph.D.

M.S. *Experimental Psychology*, Syracuse University, 2015

B.S. *Psychology*, College of Charleston, 2011

B.S. *Biology*, College of Charleston, 2011

Publications

Russo, N., Kaplan, E. A., Wilson, J. H., Criss, A. H., & Burack, J.A., (in press). Choices, challenges, and constraints: a pragmatic examination of the limits of mental age matching in empirical research. *Development and Psychopathology.*

Wilson, J. H., Kellen, D., & Criss, A. H. (2019). Mechanisms of output interference in cued recall. *Memory & Cognition.*

Wilson, J. H., Criss, A. H., Spangler, S.A., Walukevich, K., & Hewett, S. (2017). Analysis of acute naproxen administration on memory in young adults: A randomized, double-blind, placebo-controlled study. *Journal of Psychopharmacology.* DOI: 10.1177/0269881117724406

Wilson, J. H., & Criss, A. H. (2017). The list strength effect in cued recall. *Journal of Memory and Language,* 95, 78-88.

Wilson, J. H. (2016). Extralist cues and the recall process. (Unpublished Qualifying Exam document). Syracuse University, Syracuse, NY.

Wilson, J. H. (2015). The list strength effect in cued recall: Estimation, implications, and models. (Unpublished master's thesis). Syracuse University, Syracuse, NY.

Awards, Honors, and Grants

Department of Psychology Travel Awards (by Academic Year):	2013-2014
	2014-2015
	2015-2016
	2016-2017
Graduate Student Organization Travel Awards (by Academic Year):	2014-2015
	2015-2016
Travel Award from the Society of Mathematical Psychology	2015, 2016

Fellowships

Syracuse University Department of Psychology Fellowship Active Fall 2012-present

Conference Activity

Wilson, J. H., Russo, N., Kaplan, E. A., Criss, A. H., & Burack, J. A. (November 2020). Choices, challenges, and constraints: A pragmatic examination of the limits of mental age

matching in empirical research. Affiliate Meeting of the Society for Mathematical Psychology at the Annual Meeting of the Psychonomic Society, online.

Wilson, J. H. & Criss, A. H. (July 2019). Evidence for global matching during memory recovery. Annual Meeting of the Society for Mathematical Psychology, Montreal, Quebec, Canada.

Wilson, J. H. & Criss, A. H. (May 2019). Evidence for global matching during memory recovery. Context in Episodic Memory Symposium, Philadelphia, PA.

Wilson, J. H. & Criss, A. H. (November 2018). Sources of interference in a 3-pahse cued recall framework: One sampling process is insufficient to jointly account for paired associates and extralist cued recall. Annual Meeting of the Psychonomic Society, New Orleans, LA.

Wilson, J. H. & Criss, A. H. (August 2018). Sources of interference in a 3-phase cued recall framework. Annual Meeting of the Society for Mathematical Psychology, Madison, WI.

Wilson, J. H. & Criss, A. H. (May 2018). Evaluation of noise sources in a 3-phase cued recall framework. Context in Episodic Memory Symposium, Philadelphia, PA.

Wilson, J. H., & Criss, A. H. (November 2017). Evaluation and determination of noise sources in a three-phase cued recall framework. Annual Meeting of the Psychonomic Society, Vancouver, BC, Canada.

Wilson, J. H., & Criss, A. H. (May 2017). Output interference and release in cued recall. Context in Episodic Memory Symposium, Philadelphia, PA.

Chen, S., Wilson, J. H., & Criss, A. H. (November 2016). Investigation of the source of list strength effect within the Retrieving Effectively from Memory (REM) framework: Types of cue vs levels of competition. Annual Meeting of the Psychonomic Society, Boston, MA.

Wilson, J. H., & Criss, A. H. (August 2016). The list strength effect in cued recall won't be pushed around. Annual Meeting of the Society for Mathematical Psychology, Brunswick, NJ.

Wilson, J. H., & Criss, A. H. (May 2016). The (null) list strength effect in cued recall. Context in Episodic Memory Symposium, Philadelphia, PA.

Wilson, J. H., & Criss, A. H. (November 2015). Release from output interference in cued recall. Annual Meeting of the Psychonomic Society, Chicago, IL.

Wilson, J. H., & Criss, A. H. (July 2015). Release from output interference in cued recall: cue-target dissociations. Annual Meeting of the Society for Mathematical Psychology, Newport Beach, CA.

Aue, W. R., Criss, A. H., & Wilson, J. H. (July 2015). Evaluating the robustness of output interference. Annual Meeting of the Society for Mathematical Psychology, Newport Beach, CA.

Criss, A. H., Aue, W. R., & Wilson, J. H. (2015). (lack of) Output Interference in Semantic Memory. Annual Summer Interdisciplinary Conference, Mammoth, CA

Wilson, J. H., & Criss, A. H. (May 2015). Release from output interference in cued recall: Dissociations between cues and targets. Context in Episodic Memory Symposium, Philadelphia, PA.

Wilson, J. H., & Criss, A. H. (July 2014). The Retrieving Effectively from Memory Model and the list strength effect in cued recall. Summer School for the Computational Modeling of Cognition, Laufen, Bavaria, Germany.

Wilson, J. H., & Criss, A. H. (July 2014). The list strength effect in cued recall. Annual Meeting of the Society for Mathematical Psychology, Québec City, Canada.

Wilson, J. H., Aue, W. R., & Criss, A. H. (November 2013). The effects of cue and target strength in cued recall. Annual Meeting of the Psychonomic Society, Toronto, Canada.

Campus Talks

Wilson, J. H. (October 2018). The latent dimensionality of cued recall-tasks. Proseminar. Syracuse University, Syracuse, NY.

Wilson, J. H. (March 2017). A deeper look into sample-recovery models of memory. Proseminar. Syracuse University, Syracuse, NY.

Wilson, J. H. (September 2016). Towards a model of extralist and paired associates cued Recall. Proseminar. Syracuse University, Syracuse NY.

Wilson, J. H., White, C.N., Kalish, M.L., & Criss, A. H. (March 2016). The replication crisis and "Science 2.0". Proseminar. Syracuse University, Syracuse, NY.

Wilson, J. H., & Criss, A. H. (November 2015). (More) release from output interference in cued recall. Proseminar. Syracuse University, Syracuse, NY.

Wilson, J. H., & Criss, A. H. (April 2015). Release from output interference in cued recall. Proseminar. Syracuse University, Syracuse, NY.

Wilson, J. H., & Criss, A. H. (October 2014). Context, REM and the list strength effect in cued recall. Proseminar. Syracuse University, Syracuse, NY.

Wilson, J. H. & Criss, A. H. (April 2014). Is there really a list strength effect in cued recall? Or: how I learned to stop worrying and love null effects. Proseminar. Syracuse University, Syracuse, NY.

Wilson, J. H., & Criss, A. H. (September 2013). Differential list strength effects (LSEs) for cued and targets in cued recall? Proseminar. Syracuse University, Syracuse, NY.

Wilson, J. H., & Criss, A. H. (April 2013). The effects of cued and target repetition in cued recall. Proseminar. Syracuse University, Syracuse, NY.

Research Experience

Graduate Student, Memory Modeling Lab

Fall 2012 – Present

PI: Amy H. Criss, Ph.D., Syracuse University

Teaching Experience

Teaching Assistant:

Statistical Methods II	Fall 2016 — Spring 2015
Introduction to Psychology	Fall 2013 — Spring 2015
Cognitive Psychology	Fall 2019 — Spring 2020

Adjunct Professor:

Experiments in Cognitive Psychology	Fall 2020 – present

Service Positions

The Graduate Student Organization of Syracuse University

Senator	Fall 2016 — Present
President (paid position)	Summer 2017 — Summer 2019

and Graduate Student Representative to the Syracuse University Board of Trustees

National Association of Graduate Professional Students

Assistant Director, Northeast Region	April 2018 — April 2019
Director, Northeast Region	April 2020 — December 2020
Employment Concerns Director, Northeast Region	March 2021 — Present

Psychology Action Committee

Cognition, Brain, and Behavior Area Representative	Fall 2012 — Spring 2016
Events Committee Chair	Fall 2016 — Spring 2017
Treasurer	Fall 2019 — Spring 2020

Affiliations

Psychonomic Society

Society for Mathematical Psychology

National Association of Graduate Professional Students