An Integrated Working Memory Model For Time-Based Resource-Sharing

Joseph J. Glavan and Joseph W. Houpt

Department of Psychology, Wright State University

Keywords: working memory; time-based resource-sharing; ACT-R; attentional refreshing; articulatory rehearsal; computational modeling

Author Note

Joseph J. Glavan, Department of Psychology, Wright State University; Joseph W. Houpt, Department of Psychology, Wright State University. Correspondence concerning this article should be addressed to Joseph J. Glavan, Department of Psychology, Wright State University, Dayton, OH 45435. Email: glavan.3@wright.edu

Abstract

The time-based resource-sharing (TBRS) model envisions working memory as a rapidly switching, serial, attentional refreshing mechanism. Executive attention trades its time between rebuilding decaying memory traces and processing extraneous activity. To thoroughly investigate the implications of the TBRS theory, we integrated TBRS within the ACT-R cognitive architecture, which allowed us to test the TBRS model against both participant accuracy and response time data in a dual task environment. In the current work, we extend the model to include articulatory rehearsal, which has been argued in the literature to be a separate mechanism from attentional refreshing. Additionally, we use the model to predict performance under a larger range of cognitive load than typically administered to human subjects. Our simulations support the hypothesis that working memory capacity is a linear function of cognitive load and suggest that this effect is less pronounced when articulatory rehearsal is available.

Keywords: working memory; time-based resource-sharing; ACT-R; attentional refreshing; articulatory rehearsal; computational modeling

An Integrated Working Memory Model For Time-Based Resource-Sharing

Working memory (WM) is vital for elemental cognitive processing. It provides the cognitive system with the means for manipulating and retaining information, such as the intermediate solution to an algebra problem, for short periods of time. WM is often thought of as a storage system with a structurally limited capacity (Case, Kurland, & Goldberg, 1982), similar to a physical container. However, evidence is growing for the interpretation of WM as a mechanism whose active maintenance yields a functional capacity (Barrouillet & Camos, 2015). That is to say that the actions of the WM mechanism over time cause more items to remain accessible than it can strictly store at a single time (e.g. in a buffer).

The time-based resource-sharing (TBRS) model (Barrouillet, Bernardin, & Camos, 2004; Barrouillet & Camos, 2015) advocates this view, proposing that WM is a rapidly switching, serial mechanism that focuses executive attention on a single memory trace at a time to rebuild its activation (i.e. accessibility). This process, called attentional refreshing, counteracts the continuous temporal decay experienced by items in memory. While all memory items decay, only one item can be refreshed at a time because of a central bottleneck. To maintain access to a set of items, the items must share their time (i.e. take turns) in the focus of attention.

A popular paradigm for studying WM is a dual task where the subject must memorize a list of items while responding to some distractor task. The distractor task is intended to prevent the use of higher-level strategies (i.e. mnemonics) so that the number of items recalled, *span*, reflects a more pure WM capacity. Barrouillet et al. (2004) recognized that this approach may reduce the subject's ability to engage in target maintenance, but it cannot completely eliminate it. Instead, they sought to carefully control and quantify the interference from the distractor activity as cognitive load (CL).

In support of TBRS, Barrouillet, Camos, and colleagues (e.g., Barrouillet et al., 2004; Barrouillet, Bernardin, Portrat, Vergauwe, & Camos, 2007) have repeatedly demonstrated

that the essence of CL is time away from maintenance[1], rather than the number of distractors or their inherent difficulty. By using distractor task response times as a proxy for distractor processing time, they demonstrated that observed WM span is an approximately linear function of CL. Specifically, if CL is measured as the fraction of the total time on task dedicated to the distractor task, then

$$\text{span} = k(1 - \text{CL}).$$

In this formulation, k is a free parameter reflecting an individual's "pure" WM capacity. A meta-analysis of 14 experimental conditions (Barrouillet, Portrat, & Camos, 2011) indicated excellent fit for this function ($R^2 = .98$),

$$\text{span} = -8.33\text{CL} + 8.13,$$

which closely agrees with Miller's Magical Number 7 ± 2 (Miller, 1956).

One of the goals of this paper is to explore whether the linear relationship between CL and span holds for more extreme values. The mean CLs of the experimental conditions from which the preceding regression comes range from approximately .25 to .65. By using measurements from the middle of the CL continuum, it is possible that we have only observed the locally linear, middle section of a function that is actually non-linear (e.g., ogival). In other words, testing WM span at the extremes of CL would provide a stronger test of the linear function hypothesis. We explore this possibility by using our model to simulate the full range of CL.

Oberauer and Lewandowsky (2011) took a similar approach with their connectionist model, TBRS*, and simulated a much more complete range of the CL continuum. While their model did find support for a linear effect of CL on WM span, it made a simplifying assumption that may limit its correspondence with TBRS; TBRS* abstracts away the

[1]Interestingly, Lebiere (2001) also used processing time as a measure of cognitive workload in an earlier, unrelated ACT-R model. See Cummings and Guerlain (2007) for a similar metric applied to the cognitive workload of pilots.

distractor processing portion of the task. To simulate each episode of cognitive load, it samples a response time from a distribution determined post-hoc from empirical observations. We argue that this abstraction is inappropriate because the critical insight of TBRS is that observed WM capacity is a product of the interaction of maintenance and processing, and therefore a computational model of TBRS must be integrated to explain both the memory and processing tasks. Later in the paper we will demonstrate an unexpected prediction for the relationship between WM span and CL that is only revealed by an integrated process model.

The second goal of our paper is to integrate articulatory rehearsal into the computational TBRS model. While the original TBRS model regarded attentional refreshing as the primary mechanism behind memory maintenance, more recent work has highlighted the importance of other mechanisms (Camos & Barrouillet, 2014; Barrouillet & Camos, 2015). In particular, the effect of articulatory rehearsal (cf. the phonological loop; Baddeley & Hitch, 1974) on WM span depends on the verbal/non-verbal nature of the stimuli used. To our knowledge, no formal description of how attentional refreshing and articulatory rehearsal work together to maintain information in WM currently exists in the TBRS literature. Here, we propose one potential method and investigate how its predictions differ from attentional refreshing alone.

We built our computational model[2] within the cognitive architecture, ACT-R (Anderson, 2007). The comprehensive nature of the architecture allows for modeling an entire task from start to finish. Such a capability is imperative for our goal of a fully integrated model and resolves our earlier criticism of TBRS*. The architecture is composed of various modules that specialize in different areas of processing, from more central functions like goal planning to more peripheral, perception and action functionality (see

[2]The latest version of ACT-R is available for download at `http://act-r.psy.cmu.edu/`. We built our model using an older version of ACT-R 6.0, but we expect it to be compatible with newer versions. All the code needed to reproduce our results, including the model and older version of ACT-R, are available for download at `https://github.com/JosephGlavan/ACT-R-TBRS/tree/master/TopiCS/`.

Anderson (2007) for a review of the neural imaging work associating each of ACT-R's modules with particular brain regions). The modules communicate with each other through limited capacity buffers that are each able to hold a single piece of information called a chunk. Chunks are computational data structures with slots that represent their features (e.g. the context under which the chunk was created), which are themselves typically other chunks. A production system, operating on if-then rules, coordinates the behavior of the modules, which work in parallel of each other. Only a single production may fire (i.e. be selected and executed) at a time. This limit, combined with the modules' buffers, comprises the bottleneck assumed by TBRS.

One module of ACT-R in particular, the declarative module, is especially useful for our purposes. It operates on retrieval requests where the production system sends the module a set of features as a cue. The module compares the request to the chunks in declarative memory, computes their activations using Equation 1, and places the chunk with the highest activation (if it is above a threshold) into its buffer. In the activation equation,

$$A_i = B_i + I_i + P_i + \beta + \varepsilon, \tag{1}$$

β is the base-level constant (a model parameter) and ε is logistically distributed random noise. The other terms will be explained in turn later; however, it is worth noting now that base-level learning (B_i), implements the time-based decay and attentional refreshing assumptions of TBRS. In brief, the more time that has passed since a chunk was last retrieved, the less its activation will be, and the more times a chunk has been retrieved, the greater its activation will be.

A Benchmark Task

We chose an existing experiment from the literature, the third experiment from Barrouillet et al. (2007), as a benchmark for evaluating our model. Using pre-existing data allows us to develop the model around less contentious criteria than data collected

expressly for the purpose of fitting the model. It also affords us with the opportunity to determine the model's novel predictions *before* conducting new experiments.

Subjects completed a variant of the common complex span task in which they were asked to memorize a set of letters while performing a secondary task. In this experiment, a target letter was presented followed by a sequence of digits. The secondary task was to report whether each digit was even or odd (in one condition), or displayed above or below the midline of the monitor (in a separate condition). After the sequence of letters with interspersed digits, the subject was asked to verbally report the list of target letters. For each list length, a subject completed three trials. Each subject started with a single letter and would be given progressively longer lists as long as they correctly reported all letters in order for at least one of the three trials. Every target was presented for 1500 ms, followed by a 500 ms delay, and then 6400 ms of distractors. CL was manipulated by changing the number of distractors (4, 6, and 8) presented during this fixed interval, effectively changing the pacing of the task. The two versions of the secondary task, the parity task and the spatial task, further elicited differential amounts of CL. In the parity condition, subjects were asked to respond whether the distractor was odd or even. In the spatial condition, subjects were asked to respond whether the distractor appeared in the top or bottom half of the display. Before the experiment began, subjects trained with 96 distractor judgments (with feedback) and three trials each of one- and two-item lists in the manner and at the pace of their assigned experimental condition. WM span scores were calculated by dividing the number of correctly recalled trials for all lists by three (all-or-nothing unit scoring; Conway et al., 2005). Each of the six conditions entails a different demand on the subject, so we expect different mean WM spans across the conditions. The prediction of TBRS is that while the raw spans may differ, they should be the same once each condition is equated in terms of CL, and indeed this was what Barrouillet et al. (2007) found. Furthermore, this particular experiment serves as a robust test of the model by forcing it to explain quantitatively and qualitatively different task manipulations.

The Model

The general behavior of the model can be organized into essentially three levels of priority. The highest priority productions support perception, the intermediate-priority productions perform task-specific behavior (e.g., creating a new list, encoding a target, processing a distractor, initiating recall), and the lowest priority productions implement maintenance.[3]

Whenever a stimulus appears on the (virtual) display, the model moves its visual attention to encode it (i.e. the visual module creates a new chunk representing the stimulus with slots that contain its features such as position, color, etc.). Once the model attends to a stimulus (i.e. a new chunk is available in the visual module's buffer), it encodes the stimulus by retrieving the semantic information associated with the stimulus from declarative memory. The type of the chunk retrieved dictates which of the intermediate-priority routines are followed next.

If the chunk retrieved denotes that this stimulus is an asterisk, which signals the beginning of a new list, the model creates a new chunk to represent the new list-context in the goal module's buffer and sets a new goal to encode the next letter presented as the head of the list by encoding it with an additional feature slot designating it as such. If a letter (the targets in our task) is instead presented, the imaginal module, which is responsible for creating and modifying mental representations (i.e. chunks), creates a new chunk in its buffer to be the episodic representation of the target. The imaginal module also gives this chunk feature slots that contain the current temporal-context and list-context. For simplicity, we use the actual model time at the moment of encoding as the temporal-context, but in the future ACT-R's temporal module (Taatgen, Van Rijn, & Anderson, 2007) could be recruited to create this. If a number (the distractors in our task) is instead presented, then the model begins to determine the correct parity or spatial

[3]For a more thorough motivation and explanation of each of the model's assumptions and components, see Glavan (2017).

judgment, a process we discuss in more detail in the next paragraph. If the stimulus is the word "recall", the model begins recalling and verbally reciting the target list. We further describe the details of the recall process after we discuss maintenance because the processes are similar.

Distractor processing resembles a binary decision tree. At each step, the model decides between responding and retrieving more information. These decisions represent the elementary information processing steps required for the level of abstraction assumed in the model. For example, when making parity judgments, the model may first blindly guess or wait until it has fully encoded the digit. Next, it may respond based on the number alone or retrieve the parity of the number. Then it may respond based on the number's parity or retrieve the response rule that connects the semantic concepts with their associated keyboard response. Once a response rule has been retrieved, the model faithfully uses it to make a key press. Thus, the model only makes errors by responding with insufficient information. More sophisticated decision making mechanisms could be included in the future, but this simple strategy is sufficient for our current purposes. The model uses the same strategy in the spatial condition but requires fewer retrievals; in fact it only requires retrieving the associated response rule. The disparity in necessary processing steps is one reason why the spatial condition is easier than the parity condition (i.e. takes less time and induces less CL).

The model learns the appropriate response mappings through utility learning during training. If the model responds incorrectly, it may learn to retrieve more information and respond later; if the model does not respond in time, it will be penalized and be more likely to respond earlier because a guess has a better chance of being correct than a lapse, which is always incorrect. In this way, the model should exhibit a speed-accuracy trade-off (observed in Barrouillet et al., 2007). In order to bias the model toward initially using its knowledge of the task instructions over learning by purely random guessing, we give the information-seeking productions greater starting utility than the early-response

productions. Additionally, the model may continue to learn after training through production compilation, which may eliminate retrievals altogether by combining compatible productions. For example, the model may learn to associate a number's parity with its appropriate response directly, bypassing the need to retrieve a response rule (and reducing CL).

The model engages in maintenance, either attentional refreshing or articulatory rehearsal, whenever it is not busy. We discuss attentional refreshing first as it forms the core of the model and address later how we incorporate articulatory rehearsal. Attentional refreshing is implemented as simple retrieval from declarative memory. It uses two productions that leverage a remotely linked representation of the list (explained in the next paragraph). The first production initiates maintenance by requesting any target chunk, which should retrieve the currently strongest item associated with the current list. The second production uses the just retrieved item as the cue for a subsequent retrieval with the intention that the cue biases retrieval toward the next item in the list, although this is not guaranteed due to transient variations in activation caused by random noise or differences in retrieval history. Until it is interrupted by a higher priority routine, the second production repeats itself, implementing the attentional refreshing loop. In a similar fashion, recall begins by first requesting the head of the list, and then a second production vocalizes the just retrieved target while using it as a cue to retrieve the next item.

We had to make some ancillary assumptions because the TBRS theory does not specify how the list structure is represented in declarative memory. The ongoing debate in the literature regarding the exact nature of list representation suggests that it may at least partially depend on the structure of the task. Although positional coding has been used in previous ACT-R models of serial recall (e.g. Anderson, Bothell, Lebiere, & Matessa, 1998), that task design intentionally biased subjects toward an index-based representation by spatially cueing the position in the list to encode or recall. The present task avoided such positional artifacts, so we used a symbolic representation inspired by Burgess and Hitch

(2006), the assumptions of which Oberauer and Lewandowsky (2011) used to implement their connectionist TBRS* model. There are two major differences between our models. First, Burgess and Hitch (2006) assumed that items and their contexts are always available, but their association can be learned or forgotten. In our model, we assume that items and their contexts are forever joined in a single episodic representation, and this integrated representation, or chunk, itself is what is strengthened or forgotten. Second, Burgess and Hitch (2006) and Oberauer and Lewandowsky (2011) cue the retrieval of the next item by activating the next position's context representation, whereas our model uses the context of the previous item to cue the context of the next item in a manner reminiscent of chaining models (e.g. Solway, Murdock, & Kahana, 2012). We invite the reader to see Glavan (2017) for a more thorough discussion of these issues.

Two terms in the activation equation, partial matching (P_i) and base-level inhibition (I_i), enable the maintenance and recall productions to successfully iterate the list by using the just retrieved item as a cue for the next item. First, partial matching penalizes chunks based on episodic dissimilarity. In the current model, episodic dissimilarity is synonymous with temporal dissimilarity: chunks that were encoded at more disparate times receive a greater penalty; however, one could imagine episodic similarity being expanded to include other contextual information (e.g., spatial proximity, affective state, etc.) in a more general model. We quantify the penalty using

$$P_i = -\eta \cdot \ln(1 + |\epsilon_i - \epsilon_{\text{requested}}|) \tag{2}$$

where ϵ_i is the temporal context (i.e. time) at which chunk i was encoded and η is the episodic selectivity parameter.

Without compensation, the episodic dissimilarity penalty, which imposes a remotely linked structure on the list representation, creates two potential problems. First, when the just retrieved item is used as the cue for the next retrieval, it is maximally similar to itself and would therefore be continuously retrieved again, preventing iteration through the list. Second, the episodic dissimilarity penalty is effectively symmetrical in the sense that it does

not bias retrieval toward the item in the next position over the item in the previous position (assuming approximately constant intervals between encoding episodes). These problems are usually solved in other serial recall models by completely suppressing items after they are retrieved, but most of these models overlook how such a suppression mechanism is released (Glavan, 2017). Instead, we use base-level inhibition (Equation 3; Lebiere & Best, 2009) to penalize chunks based on how recently they were retrieved. In the equation below, t_i is the time at which chunk i was last retrieved and γ is the inhibition decay parameter. The base-level inhibition penalty decays over time, making the release of inhibition a function of the retrieval that initiated it rather than an arbitrary time set by the modeler.

$$I_i = -\ln\left(1 + t_i^{-\gamma}\right) \tag{3}$$

Notice that while these mechanisms promote unidirectional refreshing, they do not force it. If an item was skipped, the transposition error may be followed by a fill-in error because the skipped item has not been recently inhibited.

The base-level learning mechanism (Equation 4) in the activation equation implements TBRS's assumptions regarding attentional refreshing. It treats every time a chunk is cleared to declarative memory (e.g. following retrieval) as a separate memory trace that decays as a power law. A chunk's total base-level activation is therefore the log-sum of its decaying traces:

$$B_i = \ln\left(\sum_j t_{ij}^{-.5}\right) \tag{4}$$

where j indexes the times of past retrievals or encodings. Thus, the longer it has been since a chunk was retrieved, the less likely it is to be retrieved again (recency effect), and the more often it has been retrieved, the more likely it is to be retrieved again (frequency effect).

Note that we do not assume an explicit maintenance strategy (i.e. some fixed pattern like always starting from the beginning of the list). Although others have begun to investigate the predictions of different strategies (e.g. Lemaire, Pageot, Plancher, &

Portrat, 2018), there is not sufficient evidence at this time that attentional refreshing, which can be covert and outside conscious awareness, follows a clearly defined schedule. In light of this, we allow the temporal dynamics of decay, refreshing, and inhibition to determine which item will be retrieved to resume maintenance after interruption.

We propose that articulatory rehearsal supports maintenance by establishing a more stable signal for cueing retrieval than the temporal dynamics of attentional refreshing allow. The audicon (cf. phonological store, Baddeley & Hitch, 1974; echoic memory, Cowan, 1984; see also Huss & Byrne, 2003) is managed by the aural module and is able to hold sensory information that the model heard in the last 3 seconds in the crude order it was heard. We suggest that the cognitive system can leverage the alternative temporal and structural constraints of this sensory memory register to improve retention without a complex, CL-inducing strategy. The articulated target information can be left in its raw sensory form (i.e. without encoding a new episodic representation) and used to cue the retrieval of the corresponding target's episodic representation directly, bypassing the episodic similarity penalty.

Maintenance with articulatory rehearsal engaged is essentially an asynchronous cooperation of two (attentional and articulatory) parallel loops. The attentional refreshing loop, operating on declarative memory retrievals, continues to reactivate memory traces as previously described. Meanwhile, the articulatory loop is rehearsing the phonological representations of the items. When the two loops align (i.e. the attentional loop is ready to make a new retrieval at the same time that the articulatory loop has an item ready to be rehearsed), a retrieval request is made using the phonological representation from the articulatory loop as the retrieval cue instead of the previously retrieved item, which is what is usually used as the cue in attentional refreshing as we've previously described it. This method of cueing is direct and more reliable than similarity-based requests, which we hope will explain the commonly found advantage for recall when the articulatory system is unimpeded (e.g. in the absence of suppression).

For the reader interested in the specific details of how these two loops interact in terms of production rules, we now step through a typical maintenance cycle in the full model. It begins in the same way as the previously described attentional refreshing mechanism: a retrieval for any target chunk is requested. After one target has been retrieved the next step depends on whether there is a chunk available to be rehearsed in the aural buffer (provided that a higher priority subroutine has not interrupted maintenance). If there is not, then another retrieval using the just retrieved item is requested, similar to attentional refreshing. If the vocal module is not currently busy articulating something, then the just retrieved item is also subvocalized (i.e. the model silently repeats the item to itself). The model is able to hear subvocalizations just like external sounds, and when the aural buffer is empty it pulls the oldest sound from the audicon. If however there is a chunk available in the aural buffer when the model is ready to make a new memory retrieval and if the vocal module is also free, then a proper rehearsal is performed by subvocalizing the aural chunk and requesting a retrieval of its corresponding target from declarative memory. If the declarative module is busy (e.g. with the distractor task) while there is a chunk available in the aural buffer and the vocal module is free, then the aural chunk is subvocalized without making a retrieval request. This ensures that the articulatory loop remains full and ready for when the declarative module returns from a period of CL.

Simulation Study

As noted earlier, the bulk of the support for WM span as a linear function of CL comes from studies where spans were probed in the middle of the CL scale. To further test this hypothesis, we simulated the benchmark task (Barrouillet et al., 2007) in two different ways. In the first method, we forced specific amounts of CL on the model by preventing any refreshing or rehearsal productions from firing during the first portion of the inter-letter interval (e.g., if CL was fixed at .30, then the model would be unable to conduct maintenance for the first 2070 ms following the presentation of a target but would

be able to do so during the last 4330 ms before the presentation of the next target). We still allowed the version of the model with articulatory rehearsal to continue repeating items within the vocal-aural loop during this period as if the central bottleneck was obstructed by a non-verbal task, providing a potential advantage to this version of the model because the phonological code used to access items during rehearsal may remain intact in the articulatory loop.

In the second method, the model completed the parity and spatial judgment tasks to induce CL. Recall that Barrouillet et al. (2007) kept the inter-letter interval constant while using 4, 6, or 8 distractors to create three levels of CL; we add levels with 1, 2, 3, 9, 10, 11, and 12 distractors. In this version of the simulation, we estimate CL as the sum of distractor response times per inter-letter interval (Barrouillet et al., 2007). We expect there to be small differences between the two approaches because the predictor variables are not exactly the same (e.g., the RT-derived estimate of CL is slightly inflated from including the time to physically execute a response), but the overall trend should persist. Lastly, because we are using lower levels of CL, some conditions should produce WM spans greater than those observed in the original study; therefore, we allowed the model to encounter target lists as long as 10 items, whereas the original study stopped at 7.

The model has six free parameters: reward, inhibition decay, episodic selectivity, base-level constant, and two retrieval latency scaling parameters. All other parameters were fixed at architectural defaults. We fixed the starting utility of the information-seeking productions (see distractor processing above) to the reward parameter to avoid adding an additional free parameter.

We used MindModeling@Home (Harris, Gluck, Mielke, & Moore, 2009) to speed up our extensive simulation of the large parameter space. We began fitting the model's free parameters by first enumerating a coarse grid of the six-dimensional parameter space with 10 repetitions of each condition. We used a weighted linear composite score from Glavan (2017) that uses multiple dependent variables (accuracy, mean response time, mean total

processing time, mean span, and the slope of the regression of span on CL) to quantify the model's misfit to the human data. In this step, we only considered the six experimental conditions common to the model and the human study. Based on these results, we fixed the reward, latency exponent, and latency factor parameters at 7.0, 0.3, and 0.7, respectively, because they reasonably approximated most of the accuracy and response time trends in the human data but did not seem to interact with the remaining three parameters when predicting the span-related outcomes[4]. We then conducted a more focused, fine-grained enumeration of the three-dimensional parameter space using 20 repetitions. We used the RMSE of the predicted CL and span to the regression line from Barrouillet et al. (2011) to determine the best parameterization because in this paper we are more concerned with the model's ability to fit the proposed function of CL and because the remaining free parameters did not seem to affect accuracy or response time. It turned out that the results from the attentional refreshing and articulatory rehearsal versions of the model agreed on the best-fitting parameterization of inhibition decay (5.7), episodic selectivity (6.0), and base-level constant (15), so we used this final parameterization[5] to simulate 1000 runs of the model in the two CL and the two maintenance versions of the model. The accuracy, mean response time, mean total processing time per inter-letter interval (a proxy for CL because the inter-letter interval was constant), and mean WM span predicted by the attentional refreshing-only and articulatory rehearsal versions of the model are shown in Figure 1 with standard errors from (Barrouillet et al., 2007) shown in red. Figure 2 shows the WM span as a function of CL.

[4]For a more detailed exploration of the relationships among parameters, see Glavan (2017).

[5]The value for the base-level constant parameter is quite high compared to many ACT-R models, but this is because it is compensating for both the partial matching (episodic dissimilarity) and base-level inhibition penalties, which are not commonly used together in this way.

Discussion

As expected, the distractor results were similar regardless of whether articulatory rehearsal was included (Figure 1). Accuracy was high and in the range expected given human performance (i.e. Barrouillet et al., 2007). It began to noticeably decrease at higher distractor paces as the time available to respond approached mean response latencies (discussed further in a later paragraph). Response times were approximately 100 ms longer for the articulatory rehearsal version because the model is repeating targets to itself while the declarative module is busy processing distractors, and on average this requires two productions to fire, each taking 50 ms. Thus, simply making articulatory rehearsal available as a cognitive strategy may increase CL.

The primary difference between these two models is in WM span. When CL is fixed (Figure 2, top panel), span is approximately linear for both models. The articulatory rehearsal version predicts a smaller effect of CL and resulted in more variation around the regression line. This shallower function may reflect the slower but more stable nature of the articulatory rehearsal signal for attentional refreshing. The maximum span at lower CL may be lower than the attentional refreshing version because the refreshing loop can cycle faster than the articulatory loop, but the span at higher CL may be higher than the attentional refreshing version because the direct retrieval of targets via the articulatory cue, which does not decay with increasing (non-verbal) CL is more able to refresh items than when the episodic similarity penalty is involved.

When CL was not fixed, and the distractor task performance was simulated (Figure 2, bottom panel), span results largely agreed with the forced-CL version. An interesting prediction of the model is that mean span may not always continue to decrease with increased processing pace as suggested by previous descriptions of TBRS. Consider the conditions of Figure 1 with the most distractors per inter-letter interval. Rather than predicting continuously increasing total processing time, the *integrated* model predicts that when the presentation rate of distractors is too high, the agent changes its strategy in order

to maximize performance. Whereas the many lapses (i.e. no response given before the next distractor onset) that would be incurred by sticking to a slower but more accurate distractor processing strategy in the faster distractor pace conditions would yield accuracies close to zero, accuracy in these conditions only trends toward chance (50%) performance (Figure 1, left-most panels), which is in line with switching to a guessing strategy. Although there appears to be only minor evidence for it here, we have observed the effect across various parameterizations of the model not presented here. Critically, the model predicts that WM span can actually increase with regard to distractor pace in these conditions because the strategy change reduces the CL (Figure 2, lower right). While the majority of TBRS-related studies have assumed a homogenous processing strategy across conditions, our integrated model demonstrates that the relationship between processing and storage proposed by TBRS holds even when this assumption is violated. This task-induced, strategy-mediated tradeoff cannot be predicted by Oberauer and Lewandowsky's TBRS* model.

Our results give support to the hypothesis that WM span is a linear function of CL with the following caveats. Figure 2 only reflects the model's ability to fit the regressions of Barrouillet et al. (2007, 2011). We cannot guarantee that other parameterizations of the model could not fit alternative span functions (e.g., S-shaped, etc.). A full analysis of the model's flexibility is needed (Roberts & Pashler, 2000). Second, one could reasonably argue that WM span in Figure 2 (top left) begins to asymptote. From the current efforts, it is difficult to determine if this is noise. If WM span does level off at maximum CL, we hypothesize that this minimum will depend on the overall task time. Under full CL, no maintenance takes place, so recall should be a function of only time-based decay.

A significant difference between the forced-CL and task-driven versions of the model is the temporal distribution of CL between to-be-remembered stimuli. In the forced-CL version, all of the CL occurs during the first portion of the inter-letter interval, whereas in the task-driven version, CL is dispersed across the entire interval. TBRS does not

distinguish between these conditions because it treats CL as the overall proportion of processing time over total task time. In process models, such as ours, it becomes clear that the distribution of CL may interact with the observed WM span. For example, when CL is forced toward the first half of task time, span measures likely reflect the amount of information that could still be reinforced following the period of CL. When CL is forced toward the second half of task time, span likely reflects the surviving activation of items following their consolidation. De Schrijver and Barrouillet (2016) showed that free time (for maintenance) coalesced into a single interval has an advantage in terms of mean span over free time that is distributed over distractors, which may help to explain the overall lower spans observed in the task-driven version compared to the forced-CL version. Our larger point, that CL is also a function of time, as opposed to a summary proportion like that proposed by Barrouillet et al. (2004), demands further exploration of the effect of the distribution of CL on WM span.

Our modeling results indicate some potential pitfalls that may be encountered when trying to measure human WM spans along the full CL scale. Prior TBRS research has likely not studied the effects of small amounts of CL in humans because these conditions would be so similar to traditional serial recall or "simple span" paradigms. If CL is nearly absent such that the task is effectively a "dual" task in name only, then the trick will be retaining the carefully controlled timing of the continuous span task while discouraging the use of high-level mnemonics. At the other extreme, the model suggests that it may be impossible to actually test humans at full CL because they will change their strategy to adapt to the task demands. A further concern is that human data may reflect a mixture of refreshing and rehearsal span functions (Figure 2). People may adaptively use rehearsal so that they benefit from the higher spans available from attentional refreshing only at lower CL but then switch to articulatory rehearsal at higher CL so as to protect a larger amount of information.

We tried to further explore the increased variability of the model compared to the

human data, and while the model appears sensitive to the stimulus and subject variance, we have not been able to reproduce span functions with as little variance as the human data. One explanation may be the all-or-nothing scoring of span; Conway et al. (2005) has suggested that partial-unit scoring is less sensitive to individual variability. Indeed, the more recent literature seems to have begun to use the percentage of items correctly recalled over this measure of WM span (e.g. Camos & Barrouillet, 2014).

We hope to expand the model in the future to use ACT-R's spreading activation mechanism (i.e. association) and more sophisticated representations of context instead of partial matching. This may enable the model to produce intrusion errors and clustering effects (Farrell, 2012; see also Glavan, 2017). It is interesting to note that Lovett, Reder, and Lebiere (1999) implicated one of the parameters used in ACT-R's spreading activation calculation as a source of individual differences in WM capacity. Glavan (2017) demonstrates how the partial matching function used in this model (Equation 2) is analogous to the spreading activation equation and suggests that the episodic selectivity parameter, which plays the same role as Lovett et al.'s parameter of interest in its respective equation, may also reflect individual differences. This would be in line with Unsworth and Engle's (2006) assertion that the ability to effectively use temporal-contextual information is responsible for individual differences.

In conclusion, we provided a process-level, computational model that supports TBRS's prediction that WM span is a linear function of CL. We also proposed a novel interpretation of articulatory rehearsal as a stable signal to guide attentional refreshing. We hope that this model will inspire future process-level considerations of WM. Real world cognition takes place on a dynamic, continuous timescale. We know that WM, its constraints, and concurrent cognitive load are each related to various forms of higher-level cognition (e.g., decision making, reasoning, and problem solving; Conway et al., 2005; Payne, Bettman, & Johnson, 1988). Our work lays the foundation for coordinating each of these disciplines within a single, integrated computational process model.

References

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* Oxford University Press.

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language*, *38*(4), 341–380.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. *Psychology of Learning and Motivation*, *8*, 47–89.

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. *Journal of Experimental Psychology: General*, *133*(1), 83.

Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., & Camos, V. (2007). Time and cognitive load in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 570.

Barrouillet, P., & Camos, V. (2015). *Working memory: Loss and reconstruction.* Psychology Press.

Barrouillet, P., Portrat, S., & Camos, V. (2011). On the law relating processing to storage in working memory. *Psychological Review*, *118*(2), 175.

Burgess, N., & Hitch, G. J. (2006). A revised model of short-term memory and long-term learning of verbal sequences. *Journal of Memory and Language*, *55*(4), 627–652.

Camos, V., & Barrouillet, P. (2014). Attentional and non-attentional systems in the maintenance of verbal information in working memory: The executive and phonological loops. *Frontiers in Human Neuroscience*, *8*, 900.

Case, R., Kurland, D. M., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Journal of Experimental Child Psychology*, *33*(3), 386–404.

Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786.

Cowan, N. (1984). On short and long auditory stores. *Psychological bulletin*, *96*(2), 341.

Cummings, M. L., & Guerlain, S. (2007). Developing operator capacity estimates for supervisory control of autonomous vehicles. *Human Factors*, *49*(1), 1–15.

De Schrijver, S., & Barrouillet, P. (2016). *Consolidation and refreshing in working memory.* (Poster presented at the 57^{th} Annual Meeting of the Psychonomics Society)

Farrell, S. (2012). Temporal clustering and sequencing in short-term memory and episodic memory. *Psychological Review*, *119*(2), 223.

Glavan, J. J. (2017). *Exploring the time-based resource-sharing model of working memory through computational modeling* (Unpublished master's thesis). Wright State University. (`http://rave.ohiolink.edu/etdc/view?acc_num=wright1496099678023 64`)

Harris, J., Gluck, K. A., Mielke, T., & Moore, L. R. (2009). Mindmodeling home... and anywhere else you have idle processors. In A. Howes, D. Peebles, & R. Cooper (Eds.), *Proceedings of the ninth international conference on cognitive modeling.* Manchester, United Kingdom: University of Manchester.

Huss, D., & Byrne, M. (2003). An ACT-R/PM model of the articulatory loop. In *Proceedings of the fifth international conference on cognitive modeling* (pp. 135–140).

Lebiere, C. (2001). A theory-based model of cognitive workload and its applications. In *Proceedings of the 2001 interservice/industry training, simulation and education conference.*

Lebiere, C., & Best, B. J. (2009). Balancing long-term reinforcement and short-term inhibition. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 2378–2383).

Lemaire, B., Pageot, A., Plancher, G., & Portrat, S. (2018). What is the time course of working memory attentional refreshing? *Psychonomic Bulletin & Review*, *25*(1), 370–385.

Lovett, M. C., Reder, L. M., & Lebiere, C. (1999). Modeling working memory in a unified architecture: An ACT-R perspective. In A. Miyake & P. Shah (Eds.), *Models of working*

memory: Mechanisms of active maintenance and executive control (pp. 135–182). Oxford University Press.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*(2), 81.

Oberauer, K., & Lewandowsky, S. (2011). Modeling working memory: A computational implementation of the time-based resource-sharing theory. *Psychonomic Bulletin & Review*, *18*(1), 10–45.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 534.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? a comment on theory testing. *Psychological Review*, *107*(2), 358.

Solway, A., Murdock, B. B., & Kahana, M. J. (2012). Positional and temporal clustering in serial order memory. *Memory & Cognition*, *40*(2), 177–190.

Taatgen, N. A., Van Rijn, H., & Anderson, J. R. (2007). An integrated theory of prospective time interval estimation: The role of cognition, attention, and learning. *Psychological Review*, *114*(3), 577.

Unsworth, N., & Engle, R. W. (2006). A temporal–contextual retrieval account of complex span: An analysis of errors. *Journal of Memory and Language*, *54*(3), 346–362.

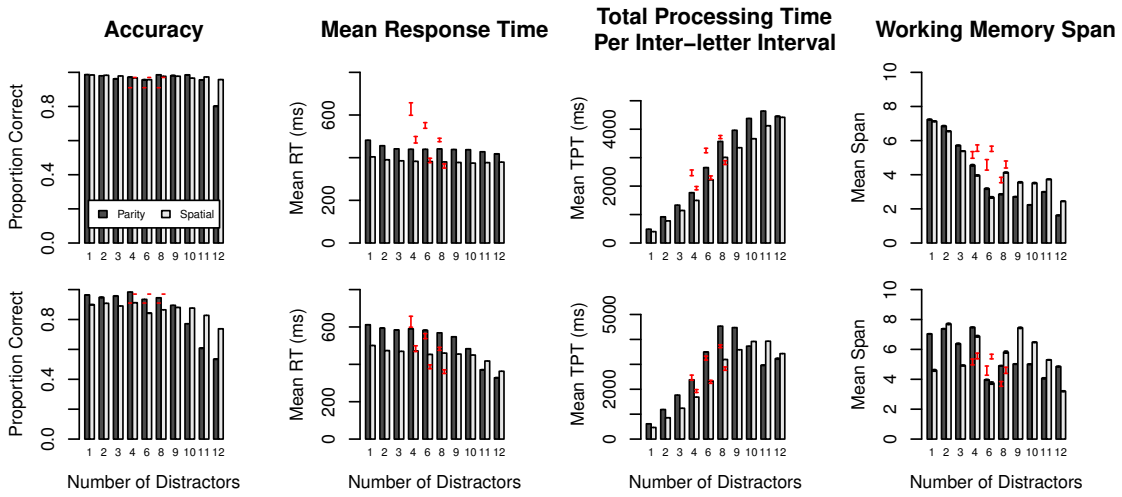

Figure 1. The three leftmost panels show accuracy, mean response time, and mean total processing time per inter-letter interval for the parity (darker bars) and spatial judgments (lighter bars). The far right panels show mean WM span. The top row comes from the attentional refreshing-only version of the model, and the bottom row comes from the version with articulatory rehearsal. Standard errors of the means from Barrouillet et al. (2007) are overlaid in red; note that Barrouillet et al. did not report accuracy variance or means by number of distractors so we plot only the mean difference by task.

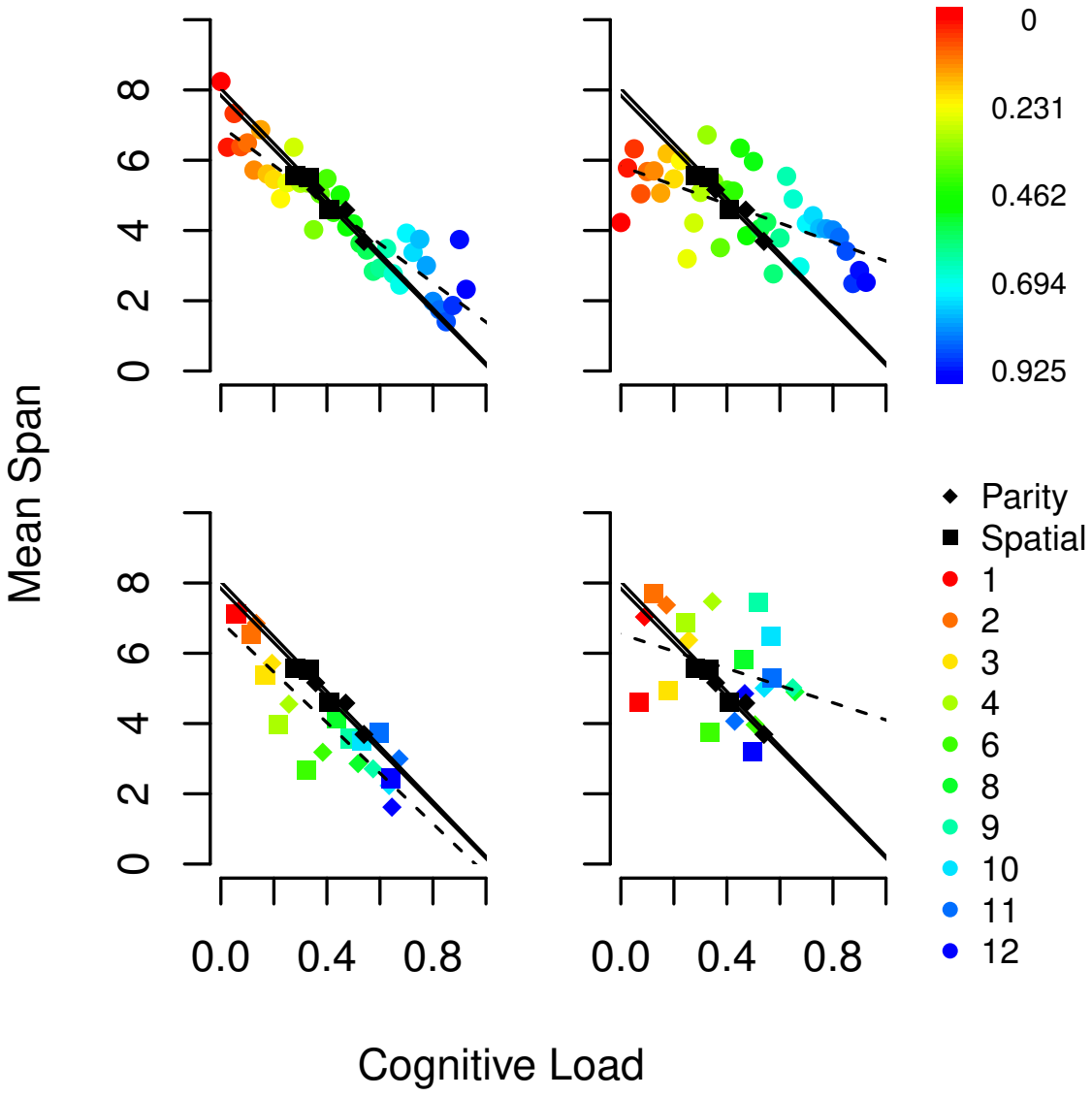

Figure 2. Mean WM span as a function of cognitive load. The attentional refreshing-only version of the model is presented on the left, and the version with articulatory rehearsal is presented on the right. The top two panels come from forcing specific levels of CL on the model, whereas CL in the bottom panels is estimated from performance on the parity (diamonds) and spatial (squares) judgment tasks. The color of the plotting symbols indicates the expected degree of CL; in the top panels this is known, but in the bottom panels this is enacted by the number of distractors processed between the presentation of each target. Results from Barrouillet et al. (2007) are plotted in black. Regression lines (solid for the human data, dashed for the model) are also included.