Bipartite Structure of Working Memory

Adam Chuderski (achuders@apple.phils.uj.edu.pl)
Zbigniew Stettner (stettner@apple.phils.uj.edu.pl)
Jarosław Orzechowski (jarek@apple.phils.uj.edu.pl)
Institute of Psychology, Jagiellonian University
Al. Mickiewicza 3, 31-120 Krakow, Poland

Abstract

The problem whether working memory (WM) structure is unitary or bipartite is being analysed in the present paper. We modified an original unitary ACT-R theory introducing a model of WM that assumes its two distinct substructures: a focus of attention (FA) and an activated part of memory outside FA. The access to the focus of attention is serial while the retrieval from the activated part of memory is parallel. We present results from a recognition memory task and we simulate these data with the manipulation in one crucial parameter of our bipartite model: the capacity of FA. We suggest that explanation of our data would be difficult to obtain on a ground of the original unitary ACT-R theory and thus we claim that bipartite model of WM fit the data better.

Introduction

Both within psychological theories of mind (Engle & Kane, 2004) as well as within its computational architectures (e.g., ACT-R; Daily, Lovett, & Reder, 2001; EPIC; Kieras, Meyer, Mueller, & Seymour, 1999; Soar; Young & Lewis, 1999) the construct of working memory – the cognitive mechanism responsible for active maintenance of information during cognitive processing – is believed central to higher cognition. The crucial issue in theorizing about WM is its structure: is WM unitary (i.e., is there a single mechanism responsible for maintenance and access) or is it partitioned into more substructures, each implementing a different mechanism of information processing?

The aim of this paper is to provide evidence for bipartite structure of WM, as we think a unitary WM mechanism is too simplistic to fully explain WM functioning. We will try to support Nelson Cowan’s (1995) view that WM consists of processes operating on two distinct memory substructures: the focus of attention and the activated area outside the focus. We describe a model of WM that assumes two distinct modes of information processing within these two substructures. We present and simulate results of the recognition memory experiment that successfully dissociated WM processing. We show that capacity of the focus of attention is crucial for explaining the observed phenomenon.

Unitary Models of Working Memory

In the 1980s, in response to some dual models of working memory (e.g., Baddeley, 1986), which proposed different mechanisms for short-term (e.g., phonological loop, visual sketch-pad) versus long-term storage, unitary theories of memory were proposed (e.g., Crowder, 1982; theories reviewed in: Nairne, 2002).

They offered a more parsimonious view of memory, showing how both working and long-term memory phenomena may be explained with a single set of cognitive mechanisms. In general, they propose (Nairne, 2002) that the item presentation events are memorized as records of cues that relate to these events. Both, in delayed and immediate recall people use available cues to reconstruct the searched item from permanent memory. No short-term (or activated) memory store is needed. Memory is potentially capacity-unlimited, but as with the passage of time current cues become poor predictors of target items, items presented long ago are almost unavailable (unless strong links with other representations are formed).

Unitary memory mechanism is implemented in the most popular cognitive architecture: ACT-R system (Anderson, Bothell, Lebiere, & Matessa, 1998; Anderson, 2005). It is the theory of mind expressed as a central control structure operating with procedural knowledge (productions) on several specialized modules, including visual, manual, goal, and long-term declarative memory modules. Working memory in ACT-R may be defined in two ways: as a subset of highly active elements of declarative memory or as a process of spreading source activation from system’s single-item focus of attention (a so-called goal buffer) to declarative elements linked with the content of the focus. These two conditions are correlated: memory traces are active due to additional activation. Trace accessibility is also affected by a learning process and by decay of trace activation.

Within ACT-R there are two methods of trace retrieval from working memory (Anderson et al., 1998). In some specific conditions (like time-pressure), the most active memory element is retrieved and tested against a probe or reported. Target element may not be the most active one, so an error of omission is probable. Alternatively, the system may try to retrieve an element identical to the probe. The higher activation of successfully retrieved element, the lower latency of its retrieval. Due to the partial matching mechanism, an element similar but not identical to the probe may be retrieved causing an error of commission. If there is no target-like element above a certain threshold at all, retrieval failure with long latency occurs. With this simple memory mechanism the theory was able to explain successfully many short-term and long-term memory phenomena.

The results usually taken as an effect of functioning of the special short-term store are explained within ACT-R theory as an interaction of universal mechanisms of both contextual activation division and of activation decay (e.g., near perfect recall of a few most recent items; Taatgen, 2001).
Bipartite Models of Working Memory

Early models of short-term memory represented it as a box-like buffer, which contained a few slots for temporal storage of information (Atkinson & Shiffrin, 1971). If all the slots where filled and a new chunk of information had to be temporarily stored, one of the previous chunks had to be removed. Recently, converging evidence emerges from both experimental work (Unsworth & Engle, 2007) and computational modeling (Burgess & Hitch, 2006) that immediate storage of information results from both complex interaction between active maintenance of information and its contextual retrieval from permanent memory. Although many detailed models of specific memory phenomena have been developed, less effort has been devoted to examination and modeling of a general structure of working memory.

Cowan (1995, 2005) has developed a very influential theoretical model of WM. The model postulates bipartite WM structure. Information comes from two hierarchically embedded faculties: currently activated subset of long-term (or permanent) memory and a set of highly activated items—the focus of attention. The activated memory is prone to decay and interference, and processing in this structure is passive, reflecting results of incoming stimuli and previous cognitive operations. Information in the focus of attention is maintained actively (thus it is resistant to decay) but this structure is capacity-limited. Cowan (2001) estimates that from three to five items can be held simultaneously in the focus. Besides from being exogenously driven, the focus can also be voluntarily controlled. For example, it can be switched by a control process to a novel stimulus or adjusted to task requirements (e.g., zoomed-in if it is important to hold just one goal in mind in face of interference).

Davelaar, Goshen-Gottstein, Ashkenazi, Haarmann, and Usher (2005) proposed a neurocomputational model of free recall consistent with Cowan’s ideas. The model includes two components: a capacity-limited activation buffer (a focus of attention analogue) that can hold a few self-activating but laterally inhibiting items, and a long-term episodic system encoding contexts in which items occurred. When information is recalled from memory, first contents of active buffer are uploaded, and then context cues are used for retrievals from long-term memory.

Chuderski, Stettner, and Orzechowski (2007) have recently published the model based on Cowan’s assumptions on structure of WM. The model was used to simulate both position effects and group differences in memory recognition task (Sternberg, 1966). Below, we present a modified version of the model and we test it against some new experimental data. We show that our bipartite model predicts observed results very well and offers probably a more parsimonious and a more plausible explanation than the unitary ACT-R memory theory the model is based on. Simulation results suggest that capacity-limited WM buffer is an important factor in WM functioning.

Two-Phase Model of Working Memory Search

In the well known paper, Saul Sternberg (1966) examined the way people search their short-term memory (STM). Participants were shown 1 up to 6 digits (a memory set), and then another digit (a probe), which could have belonged or not to the memory set. Sternberg found that RT rose with increasing memory set size. He concluded that STM had been searched serially and participants needed approximately 38 ms to scan one MS element. Subsequent experiments showed that in case of larger memory sets RTs did not rise so rapidly. Both parallel, capacity-limited models (Murdock, 1971) and two-phase (serial-parallel) models of STM scanning were proposed (Theios, 1975).

Chuderski et al. (2007) proposed a two-phase model of STM search process. In our model of Sternberg task performance, which is implemented in ACT-R cognitive architecture, we used above cited ideas of Cowan. To ACT-R declarative memory, which closely reflects Cowan’s concept of activated part of long-term memory, we added a capacity-limited but prone to decay and interference focus of attention. Technically, the focus was implemented as a few additional slots in the ACT-R goal buffer. The processing of the current version of the model is as follows.

During encoding the stimuli from memory set respective chunks are added to declarative memory. If time allows, updating production is fired. It retrieves the most active chunk from declarative memory (usually, due to decay process, the most recent one) and inserts it into the least recently updated slot while removing the chunk that resides in this slot. When a probe is presented, the model starts a search process. First, it serially scans slots of the focus of attention. Each scan takes one production to fire. The more slots are scanned (this is controlled by a model parameter: focus of attention capacity), the longer this phase of search takes. If the probe is not found in the focus of attention, the model starts the parallel phase of search: regular retrieval from ACT-R declarative memory. The latency of this phase does not depend on the number of items in the memory set. If nothing is found after both phases, the model guesses the answer. The probability of responding “yes” increases with memory set size (this is the second specific model’s parameter). This mechanism is based on the rationale that the more items are presented to a participant in a trial, the more he or she is aware that a lack of recognition could result from the fact that he or she had forgotten the presented target.

The previous version of the presented model has successfully replicated both observed accuracy and latency position curves, indicating a relation between a position of a probe in a fixed-size memory set and a participants’ performance (Chuderski et al., 2007). Accuracy of recognition fell rapidly from near perfect for the most recent probe position (-1) to a random level for the least recent probe positions (-6, -7), rising a little for the first position (-8; a primacy effect). The latency rose rapidly from the most recent position (675 ms) to the position -4 (850 ms), when the latency curve became almost flat (~900 ms; a small but significant primacy effect was observed). High $R^2 = .970$ was obtained for prediction of latency data when focus of attention capacity parameter was randomly varied from one to five elements.

The most interesting observation concerned individual differences found in the latency data. We found that some participants had almost flat latency position curves, speeding up responses only for probes on position -1 and -2, while the other participants’ RTs rose steadily for less
recent probe positions. Both groups of participants did not differ significantly in accuracy. We assumed that steep-curve participants exploited their focus of attention to greater extent, updating more items than flat-curve participants, who relied mostly on parallel access to activated memory outside the focus. When we run two separate simulations, one with low (1-2) and the other with high (4-5) values of the capacity parameter, we were able to replicate \( R^2 = .943 \) flat- and steep-curve group data, respectively.

However, a post hoc identification of the both groups is a quite weak method. In the present study we aim to get two different patterns of latency data (i.e., flat vs. steep position curves) with a stronger method of experimental manipulation. This manipulation should encourage or discourage (depending on an experimental condition) participants to use and update their attentional focus of working memory. We used performance feedback as a experimental factor. One group of participants was informed about latency of responding. We expected that in order to speed up their responses, these participants would rely on parallel access to activated memory while avoiding updating and searching the focus. The participants from the other group, informed about accuracy of responding, would exploit their focus and carefully update as many items in the focus as possible.

In the present study, apart from aiming to test further the new version of our memory search model, we are focused on more general aim of examination of the working memory structure. We will try to support Cowan’s idea of bipartite working memory by showing that a manipulation with focus of attention capacity parameter (i.e., a crucial feature of the crucial substructure of bipartite memory) allows for explanation of an important experimental result, which is difficult to be explained in another way.

The Experiment

Based on our previous results, we expected position effects both for accuracy as well as for latency of responses. The most important expectation, however, concerned feedback manipulation. We expected that accuracy feedback group would have steeper latency curve than latency feedback group (no difference in accuracy curves was predicted). We also wanted to computationally replicate these differences with a manipulation to only one model’s parameter: the capacity of the focus (i.e., its maximal number of slots). Higher value would represent larger involvement of attention in the accuracy feedback group.

Method

63 university students participated. A computerized memory search task was used, with a pool of 16 consonants as stimuli, each 2×1.5 cm in size. In each trial, eight random stimuli (none could be repeated) were presented sequentially in the center of the screen for 800 ms apiece. An asterisk presented before the first stimulus served as a fixation point. After the eighth stimulus a mask was presented for 500 ms and then a probe was displayed inside a rectangle. If participants decided that a probe was presented in the current trial, they had to press key “Z” (“M” in the opposite case). Cues on computer screen helped to remember proper response keys. The time for response was limited to 2000 ms.

40 trials for each target position (320 trials in total) and another 60 trials in negative condition (i.e., when memory set did not include any probe) were used randomly. All 380 trials were divided into five equal-length sessions with short breaks in between. The relative position of a probe in the memory set (-8...-1) was the first (repeated measure) independent variable.

The feedback bar (18×1 cm in size; black and white) shown on the top of the screen indicated participants’ performance. At the beginning of the session, black part of the bar constituted 50% of its length. In one group the black part became longer if a participant responded correctly and it shrank if she or he committed errors. In the other group the black part became longer if a participant speeded up her or his responses and it shrank if she or he slowed down. Several most recent trials were taken into account when calculating an index of participants’ performance. Participants were assigned randomly to one of the two groups. The type of feedback (i.e., accuracy vs. latency) was the second (between-subjects) independent variable.

Results and Discussion

Data from seven participants, who made errors in more than one third of trials, were excluded from analysis. This resulted in 30 participants in the accuracy feedback group and 26 subjects in the latency feedback group.

In case of accuracy dependent variable, both main effects were significant. Accuracy feedback group responded more correctly than latency feedback group (.84 vs. .77, respectively), \( F(1, 54) = 13.00, p < .001, \text{MSE} = 0.047 \). This effect validates experimental manipulation with feedback. The accuracy dropped significantly from .95 for the most recent probe position to .65 for position -7, with a primacy effect for position -8 (.72), \( F(7, 378) = 73.24, p < .001, \text{MSE} = 0.009 \). Both factors did not interact significantly (\( p > .1 \)).

In latency data analysis, latencies shorter than 200 ms and longer than mean plus three median absolute deviations were excluded (4.82% of trials). Only the position effect was significant, \( F(7, 378) = 72.84, p < .001, \text{MSE} = 2121 \), and this factor interacted with the feedback factor, \( F(7, 378) = 2.93, p = .005 \). As we expected, the accuracy feedback group generated much longer response latencies (in comparison to the latency feedback group) for the least recent probe positions (-8...-6) than for its most recent positions. Response latencies in negative condition were longer than averaged latencies in positive condition (722 ms vs. 632 ms, respectively), \( F(1, 54) = 95.59, p < .001, \text{MSE} = 2520, \) but feedback factor was neither significant nor interacted with positive vs. negative condition. Latency data for all probe positions and for negative condition are presented in solid lines in Fig. 1. Latencies for both feedback groups are presented in solid lines in Fig. 2.

The most important conclusion from the presented data is that the accuracy feedback group had steeper latency position curve than the latency feedback group. Both groups had comparable latencies of three most recent stimuli recognition, but it took the accuracy feedback group longer than the latency feedback group to recognize the less recent stimuli. This probably indicates that the former group relied more on serial memory scanning than the latter one.
Simulation of Working Memory Search

We aimed at precise simulation of chronometric patterns in the observed data, putting accuracies aside (however, simulated pattern of accuracy data resembled observed data, $R^2 = .852$). We optimized a fit of simulated latencies to data averaged from both groups, searching for the best values in the parameters space. Three ACT-R parameters were set to values different than default ACT-R values: activation threshold (0.76), activation noise (0.08), and latency factor (0.40). The latencies of both a production responsible for perceiving stimuli and for emitting motor responses were set to widely used values of 150 ms. Four specific model parameters were also set. The most crucial one, the capacity of the focus of attention parameter was varied randomly from 1 to 3 slots. Latency of a production responsible for updating process, which usually consists on silent rehearsal (Chuderski et al., 2007), was set to 250 ms. Latency of a production which simulates access to the focus of attention was set to 40 ms (a bit shorter than default production latency in ACT-R which is 50 ms). Probability of guessing ‘yes’ in case of a retrieval failure during the second phase of search was set to .264 (i.e., 8 items × .033) on the basis of our previous simulations. With these parameter values set, the model, in 666 Monte Carlo runs, was able to predict 98.7% of variance in latency of search in nine experimental conditions (see: Fig. 1, dashed lines). The error of prediction was very small, $RMSD = 12.8$. The model was able to replicate the exact shape of the position curve as well as the effect of longer latencies of negative responses than the longest positive responses (i.e., for probes on position -7).

Fig. 1: Observed (solid lines) and simulated (dashed lines) mean RTs (ms) for all target positions relative to a probe.

However, the crucial test of the model concerns the question of whether the model is able to replicate patterns of two experimental groups without any parameter modification. On the basis of our hypothesis that participants from both groups differed in the way they exploited their focus of attention, we used all simulations with the capacity parameter equal to 1, and a random half of the simulations with this parameter equal to 2, to predict data from latency feedback group which, as we assumed, exploited the focus of attention to less extend. Simulations with capacity parameter equal to 3, and a random half of the simulations with this parameter equal to 2, were used to predict data from the accuracy feedback group, which probably relied on the focus of attention to greater extend. We did not change any model’s parameter. Simulated data fitted very well to 18 observed data points ($R^2 = .947$, $RMSD = 16.8$) and each prediction fell in respective 95% confidence interval. The simulation was also consistent with observed accuracy differences: the accuracy feedback group simulation yielded higher average accuracy than simulation of latency feedback group (81.83% vs. 80.05%, respectively).

Fig. 2: Observed (solid lines) and simulated (dashed lines) mean RTs (ms) in both groups, for all target positions relative to a probe.

Alternative Unitary Explanations

Could the observed between-group feedback effect be explained by a more parsimonious mechanism of memory, for example by the ACT-R unitary memory mechanism which includes neither the multi-slot focus of attention nor the serial phase of its search? Anderson et al. (1998) proposed a ACT-R model of the Sternberg task. In this model, memo- rized items are coded directly in declarative memory. Their activation is being boosted by a rehearsal process. An additional activation is propagated from a probe to the item associated with this probe. The most activated item is retrieved and compared with the probe. Owing to the additional propagated activation it is this associated item that is most probably retrieved. If retrieved item matches the probe the model responds ‘yes’, if the item mismatches the probe (what would be the case in negative condition, when no item is associated with the probe), the model responds ‘not’. This model nicely replicated (Anderson et al., 1998) both the recency and primacy effects, and position effect as well. It simulated longer latencies in negative than positive condition, as no additional activation can be propagated in the former case. Below, we discuss whether an explanation of the observed effect of steep versus flat position curves would be possible for this model. For two reasons we decided not to take directly Anderson et al.’s model code running respective simulations, but to conduct theoretical analysis of ACT-R working memory unitary theory instead. First, this model simulated different experimental situation: words presented for 1.5 sec. were used as stimuli, while our
model encodes shortly presented letter stimuli. It is much harder to associate single letters with long-term memory contents (huge proactive interference is being built quickly, as almost the same letters are used in consecutive trials) than to do the same with words. Second, if our simulation would fail to produce the expected flat vs. steep position curve effect, this in fact would not be a proof against Anderson et al’s theory, but it might result from specific model parameterization. Theoretical analysis of a model allows for escaping such possibility.

Is ACT-R theory then able to explain the observed effect in so elegant and parsimonious way as our model is (i.e., with change in only one model parameter)? What parametric manipulation would increase accuracy of response in accuracy feedback group simulation (as compared to simulated latency feedback group) while increasing at the same time average latency for responses to the less recent target positions, but leaving recognition latencies for other positions intact?

The crucial ACT-R parameter controlling efficiency of its working memory is the amount of activation that can be propagated from a goal to memory representations. Manipulations made to this value (i.e., $W$ parameter) yielded apt predictions of individual differences in working memory task performance (Daily et al., 2001). The higher $W$ value is, the more probable are correct memory retrievals. What if we assumed that participants in the accuracy feedback group adjusted their $W$ values (e.g., concentrated their attention)? Higher baseline plus propagated activation of target memory item indeed would yield higher accuracy of responses (as observed in our experiment), but at the same time it would make (contrary to the experiment results) these representations retrieved faster (as item retrieval latency is a decreasing function of its activation). Thus, $W$ parameter cannot be considered as the only one responsible for observed effects.

Another memory parameter is a retrieval threshold: a level of activation below which memory items cannot be retrieved. Lowering this ($r$) parameter would adjust probability of correct retrievals. It would also result in (observed) increase in time of negative response (retrieval failures, whose latency is a decreasing function of $r$). However, latency of successful retrievals is not related to $r$ parameter, so no difference in latency position curve would be observed between simulations done with high and low $r$ values.

It might be possible that accuracy feedback group has better memory retention, which means that items decay more slowly ($d$ parameter would be lower). But less decayed items would be retrieved faster as they would retain more activation.

Lowering activation noise ($s$ parameter) would probably increase response accuracy, but it would unlikely increase their latency. Manipulations in other ACT-R parameters (e.g., in mismatch penalty or in base level activation) are even less likely to bring the expected results.

Probably the only interesting possibility of theoretically grounded parameter manipulation that would bring intergroup results similar to observed in the experiment, would be an increase in $W$ parameter in simulation of accuracy feedback group (as compared to simulation of latency feedback group data) while at the same time increasing a latency factor ($F$), that scales relation between item activation and its retrieval time. Probably, when paying attention to accuracy feedback participants might become more cautious in the way they accessed their memory.

Of course, we do not claim that ACT-R theory cannot predict our data in other way. Probably, some specific differences in modeled strategy (e.g., differences in procedural knowledge or in memory associations between groups) would result in successful simulation of the observed effect. Anyway, we suggest that our bipartite model of working memory, which assumes two different substructures and two sets of processes (related to each structure), offers a very natural and parsimonious explanation of a non-trivial effect that we observed. Explanation of main effects (e.g., set size, recency, or position effects), successful for ACT-R theory, is a very important way of validating the working memory models. However, also an explanation of inter-individual differences (e.g., caused by between-group experimental manipulation) should be accounted easily by such models (Chuderski et al., 2007; Lewandowsky & Heit, 2006). We believe that some memory phenomena (e.g., flat vs. steep position curves in recognition memory tests) are natural effects of discontinuity in working memory structure and processes. For example, the presented differences in latency patterns of recognition in two feedback groups probably result from different mixtures of two memory processes (e.g., serial search of the focus and parallel access to activated memory outside the focus) in these groups. We believe that unitary theories of memory are less capable of capturing such discontinuities (for similar argument concerning a different nature of recency effect see: Davelaar et al., 2005).

**Summary and Conclusions**

We observed that when people are encouraged with a proper feedback to respond more accurately, they tend to prolong only responses for the less recent target positions, while for other target positions they respond with the same latency as people encouraged to respond more quickly. The proposed explanation of this effect, which follows Cowan’s model, is based on the assumption of bipartite structure of working memory (focus of attention + activated memory) and of two phases of search (the serial scan of the focus and the parallel access to the activated memory). The model we proposed offers a parsimonious explanation of this phenomenon. Taking subsets of the crucial model parameter (attentional focus capacity) allowed us to replicate different patterns of recognition latency with a very good fit to the empirical data.

We concluded that original ACT-R theory of working memory cannot offer as parsimonious and plausible explanation as our model can. We believe that the bipartite working memory model proposed by Cowan offers a reasonable compromise between elegant and parsimonious unitary theories of memory (which, in our view, skip some important differences between short-term and long-term memory mechanisms), and multiple-component models (e.g., the one proposed by Baddeley), which seem to be too modality focused and too specific. Cowan’s model abstracts from specific mechanisms (e.g., verbal vs. visual mnemonic strategies), but it accounts for many effects in working memory (see: Cowan, 2001, for an extensive
review of these effects) that seem to result from involvement and interaction of the two functionally and biologically distinct cognitive structures that are probably out of the scope of unitary theories of working memory.

Our work resolves also some problems posed by critiques of Cowan’s model. For example, Jou (2001) noticed that if Cowan uses the subitizing phenomenon (i.e., up to four visual stimuli are counted in parallel) as an argument in favor of the capacity-limited focus of attention in WM, he should also explain why the same effect is not observed in short-term memory search (i.e., usually recognition time increases as memory set size increases from one to four items). Our work shows that if serial access to the focus of attention is assumed (at least when its contents are not perceptually available), steep increase in latency with increasing memory set size should be expected.

We noticed that the last version of ACT-R (Anderson, 2005), owing to the introduction of the imaginal buffer, has the potential of accounting for bipartite structure of the system representing information crucial for on-line processing. This construct, which is responsible for decay-prone maintenance of the contextual information relevant to the current problem, seems to be a kind of analogy of the focus of attention construct from Cowan’s theory. For example, both theories localize the problem-state active-maintenance subsystem in the parietal region of the brain.

Our work points out that the capacity of the focus of attention (either constitutionally determined or task induced) is a crucial cognitive parameter. The way people exploit their attention probably influences their cognitive performance to the great extent.

Acknowledgments
This work was supported by grant N106 2857/32 (‘Computational model of working memory search’) sponsored by Polish Ministry of Science and Higher Education (years 2007-2008). We thank Jaroslaw Kwiatkowski for programming the experimental task.

References