N-back Training Task Performance: Analysis and Model

J. Isaiah Harbison (jiharb@umd.edu)
Center for Advanced Study of Language and Department of Psychology, University of Maryland
7005 52nd Avenue, College Park, MD 27642 USA

Sharona M. Atkins (satkins@psyc.umd.edu)
Neuroscience & Cognitive Science Program Department of Psychology, University of Maryland
Biology/Psychology Building, College Park, MD 27642 USA

Michael R. Dougherty (mdougherty@psyc.umd.edu)
Department of Psychology and Center for Advanced Study of Language, University of Maryland
Biology/Psychology Building, College Park, MD 27642 USA

Abstract

Despite the n-back task's apparent effectiveness as a working memory (WM) training task, its status as a WM assessment is questionable. We analyzed the accuracy and reaction time data of participants performing of an adaptive n-back training task and developed a computational model to describe this performance. Application of our model to n-back training data suggests that performance is consistent with a two-stage, familiarity and recollection account. Furthermore, our results suggest that interference resolution is an important determining factor for task accuracy, especially when responding to targets.

Keywords: working memory; executive functioning; working memory training; n-back; continuous performance task; computational model.

N-back and Working Memory

The n-back task has often been used as a working memory (WM) assessment (Owen et al., 2005) and has recently become popular as a WM training task (Jaeggi et al., 2008). Performance gains on n-back training transfer to tasks that are heavily reliant on WM. Nevertheless, prior work questions the validity of n-back as a measure of WM ability (Jaeggi et al. 2010; Kane et al., 2007) and n-back performance gains do not appear to transfer to complex WM span tasks (Jaeggi et al., 2008; Li et al., 2008).

Understanding how n-back is performed is important both for the purpose of evaluating the its validity as a measure of WM and for isolating the mechanisms that improve over the course of WM training. The present study provides an analysis of performance on an adaptive n-back training task and a model of n-back performance.

The N-back Task

In the n-back task, participants are presented with a sequence of stimuli (e.g., letters) one at a time and asked to compare the current stimulus to one presented *n* items prior in the sequence. When performing 2-back, the current stimulus is a target when it matches the stimulus presented two stimuli ago. So in the letter sequence "P-F-D-C…", the participant should respond "match" if the 5th letter in the sequence were a "D" because it would match the one occurring two prior, but respond "no match" otherwise.

The inter-relationships within a sequence of stimuli appear to be an important factor in determining how the task is performed. In particular, stimuli (i.e., lures) that match in locations n+1 or n-1 can change how the n-back task is performed (Kane et al, 2007). For example, if the 5th letter in the aforementioned sequence were an F, it would be considered a lure because it occurred n+1 stimuli ago, and the correct response is "non match". Lures are more difficult to reject than other non-lure/non-targets stimuli; participants are less accurate and take longer to respond to lures than to other non-targets (Gray, Chabris, & Braver, 2003; Kane et al, 2007; McCabe & Hartman, 2008; Oberauer, 2005).

Arguably, the presence of lures changes how participants perform the n-back task (Kane et al., 2007). Without lures, it would be possible to use familiarity alone as the basis for a correct response. Any stimulus re-occurring somewhat recently would be a target. However, when lures are included in the sequence recent re-occurrence is not enough to distinguish targets from non-targets. Instead, it is necessary to recollect either what stimulus occurred n items back or have a fine-grained estimate of when a familiar stimulus last appeared.

Given the suggested importance of lures, the current analysis focuses on comparing participant performance on targets, lures, and other non-targets.

Experiment: Training Data

Fifty-six participants completed ten sessions of an adaptive, n-back training task as part of a larger working memory training battery. This battery included a training version of running-span, letter-number sequencing, and block span (Atkins et al., 2009) tasks as well as four tasks provided by Posit Science inc. (Brain Fitness Program, Version 2.1; Insight, Version 1.1). For the present purposes, we will only note that many participants improved their performance on the training tasks, and specifically on the n-back training task. Furthermore, performance gains on the n-back training task correlated with gains in several remote tasks, including sentence ambiguity resolution (Novick et al., submitted).

N-back Training Task Design

Similar to other training versions of n-back, our version adapted in difficulty based on participant performance. Two factors were manipulated to change the task difficulty. The first was the lure level. There were three levels of lures. The easiest level (level 0) consisted of no lures. At the next difficulty level (level 1) lures appeared in position n+1. In the most difficult lure level (level 2) lures appeared both in position n+1 and n-1. In addition to adapting lure level to participant performance, we also adapted difficulty by changing the value of n. N could range from 1 to 8.

Participants were presented 25-item sequences. In each sequence there were 5 targets, 0 or 5 lures and the rest were other non-targets (i.e., letters that had last occurred more than 10 letters prior). Participant performance on each sequence was used to determine whether and how the task difficulty should adapt on the subsequent sequence of 25. When participants were correct at least 85% of the time the task got more difficult; when they were correct less than or equal to 65% of the time, the task got easier. Otherwise, the task remained at the same difficulty level.

The difficulty level changed by first changing the lure level. If the difficulty needed to be increased and the lure level was less than 2, the lure level would increase. Once at the maximal lure level, n would increase and the lure level would be reset at zero. Similarly when the task needed to be made easier and the lure level was greater than 0, the lure level would be decreased by one level. If the lure level was already 0, then n would be decreased by one and the lure level would be reset to two. All participants started at 2-back with no lures (i.e., lure level of zero).

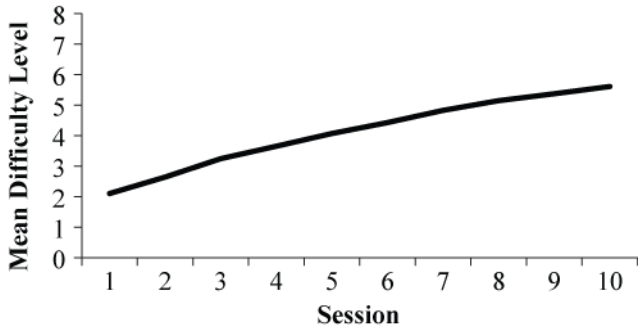

Figure 1: Mean Difficulty level reached by participants by training session.

General Findings

On average, participants showed marked improvement over the course of training. Figure 1 shows the mean difficulty level reached by participants across training sessions, where difficulty level is defined as the value of n reached plus 1/3 of the lure level or

$$D = n + \frac{LureLevel}{3}. \qquad \text{Eq.1}$$

Difficulty level can be taken as an indicator of overall performance, but it does not shed light on what cognitive processes were used to complete the task.. For that purpose we turn to accuracy and reaction times on the target, lures, and other non-targets individually.

Accuracy

Figure 2 shows the percent correct when the target, lure, and other non-target trials were shown in the third through 25th serial positions. Participants demonstrated pronounced and consistent primacy on target trials across serial positions. Little or no primacy was found for lures and other non-target trials.

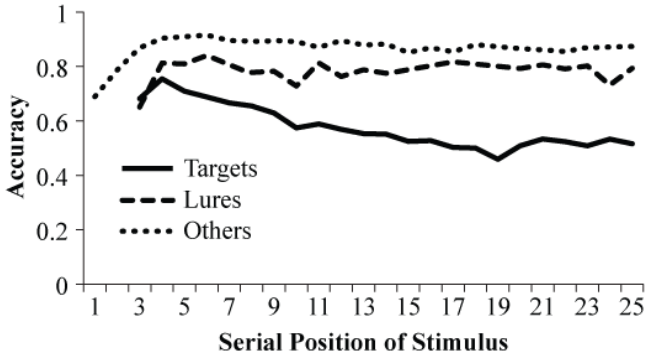

Figure 2: Mean Accuracy for Targets, Lures and Other non-targets across serial position in the stimulus sequence.

When accuracy is examined separately for each level of n, the same basic relationship is found. There is an initial drop in target performance down to an asymptote; the lowest level of the asymptote is negatively correlated with n. The top panel of Figure 3 shows representative results from the 4-back task.

Reaction Times

Participants responded correctly to both lures and targets significantly more slowly than to other non-targets. As shown in Figure 4, the mean correct reaction time (RT) to targets and lures were both approximately 380 ms (380.5 and 379.8 respectively). The RT to other non-targets was 343.4, significantly quicker than both other trials types as determined by within participant t-tests (p's < 0.001, note that other significance values are also from within participant t-tests). This same pattern is found when analyses are performed separately for each level of n. The target and lure RTs did not differ significantly for any value of n. In contrast, for all n values except 8 other non-targets were responded to more quickly than lures and for all n values except 2 other non-targets were responded to more quickly than targets (p's < 0.05).

A different pattern was found for incorrect response RTs. Participants were significantly faster at responding incorrectly to targets than to lures ($p < 0.05$) and other non-targets ($p < 0.01$). When examined at each level of n, the

results are largely consistent. For n's of three through eight, incorrect target responses were quicker than incorrect lure and incorrect other non-target responses. However, likely due to the small number of incorrect lure and other non-target responses, these differences were only significant four times.

Comparing correct to incorrect response RTs, no significant difference was found for targets. However, correct responses were significantly quicker than incorrect responses for both lures ($p < 0.01$) and other non-targets ($p < 0.001$).

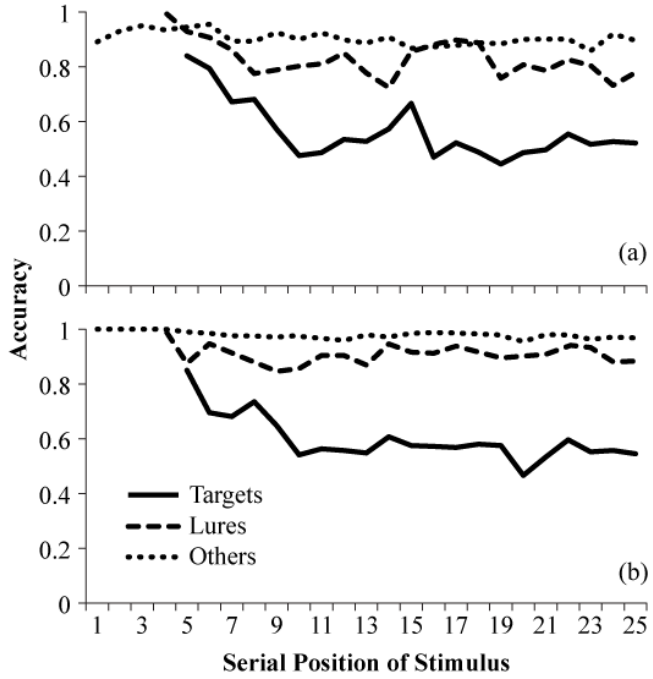

Figure 3: Participant (Panel A) and Model (Panel B) Accuracy across serial positions for 4-back.

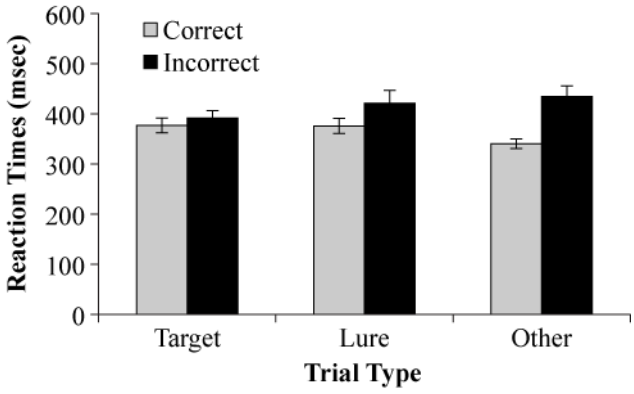

Figure 4: Mean Reaction Time for Targets, Lures and Other non-targets for Correct and Incorrect Trials.

Summary of Results

The RT results are consistent with previous research. Lures were expected to take longer to reject than other non-targets. Similarly, responses to lures were expected to be less accurate than responses to other non-targets. However, the primacy found in targets trials was surprising. The number of items that it is necessary to track, namely n, is constant across the entire sequence. Despite this, the accuracy for early targets in the sequence is greater than for later targets.

Follow-up analyses indicated that the obtained primacy was not due to a decrease in the probability of responding "match" due to the number of prior "match" responses. The probability of responding "match" to a target did not vary within a sequence, and remained constant at about 58%.

One explanation for the observed primacy is that participants were less than perfect at removing stimuli from consideration that were not longer relevant. Irrelevant stimuli, stimuli that occurred greater than n positions prior, may have been maintained in addition to and potentially at the expense of the relevant stimuli. Removal of irrelevant information has previously been indicated as important to performance in the n-back task (Oberauer, 2005).

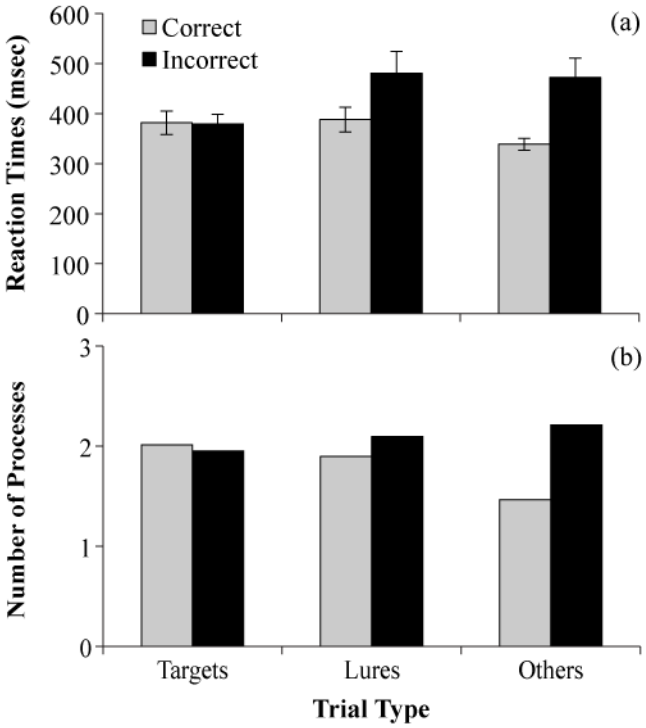

Figure 5: Participant Reaction time data (Panel A) and Model predictions for 4-back.

Modeling n-back Performance

A computational model of n-back performance was developed based on prior work describing n-back performance. Specifically, the model implemented a two-stage decision process, which includes a familiarity and a recollection process. It also implemented imperfect removal

of irrelevant information from the set actively maintained in WM. Both of these assumptions were based on Oberauer's (2005) account of n-back performance. In addition, to allow the irrelevant information maintained in WM to impact performance, we implemented forgetting as due to interference between items actively maintained in WM (Oberauer & Lewandowsky, 2008).

Model Implementation

These theoretical assumptions were implemented within an existing model of familiarity/probability judgment and recall/recollection, HyGene (Thomas et al., 2008). While this model has previously only been applied to hypothesis generation and judgment, it is based on a model of recognition memory, Minerva2 (Hintzman, 1988) and is therefore well equipped to handle familiarity judgments. It also utilizes sampling and retrieval dynamics based on successful models of recall, making it capable of recollection as well. To apply HyGene to the n-back task it was necessary to: (1) Elaborate on its WM processes, (2) Add a multi-stage recognition process, and (3) Represent time.

WM Processes We assumed that while performing the n-back task, participants try to maintain the last n items in an active subset of memory. Once the item is more than n stimuli old, the model attempts to remove that item from the active subset. The probability of successfully removing the no longer relevant item on each time step is determined by a new parameter in the model, pRemove. In addition, items in the active subset compete with one another. Each feature can only be maintained by one item in the active subset (Oberauer & Lewandowsky, 2008), therefore the competition for features between active items causes interference.

Recognition Process The model completes up to three processes when responding in the n-back task. The initial step is determining the familiarity of the current stimulus. If the stimulus is not sufficiently familiar, then the current stimulus is judged as a non-match and no further processing steps are taken. However, if the current stimulus is sufficiently familiar, an attempt to recall or recollect the n-th back item is made. If the retrieved item matches the current stimulus, the response is "match". If the retrieved item does not match the current stimulus, then the response is "non-match". If retrieval fails, that is the activation of the to-be-retrieved items is less than a threshold tRetreival, then the model guesses whether or not that stimulus is a target. The RT predictions from the present simulations are based on the simplifying assumption that each process (familiarity judgment, recollection, and guessing) takes a single unit of time.

Time Contextual drift was used to represent time. With each time step the representation of the current context was modified with probability pDrift. This allowed the model to search for the n-th back stimulus by probing memory with the n-th back context. However, we assumed that the n-th back stimulus is only probabilistically reinstated. Specifically, each item of the n-th back context is reinstated with probability pReinstate.

The current, modified version of HyGene does not use any of the standard HyGene parameters (L, A_C, Act_{MinH}, TMAX). Instead, as indicated in the model modification description it introduces four new parameters. These parameters and their values for the reported simulations are shown in Table 1.

Table 1. Parameters

Name	Sim. Value
pRemove	.15
pDrift	.33
pReinstate	.75
tRetrieval	.10

Model Details

There are three components used in the modified model: the probe, the active subset of memory, and semantic memory. Each stimulus in the active subset of memory is represented as a trace, a combination of an item (e.g., letter) and the context in which the item appeared. Each item is represented as a unique, randomly generated vector of 1's, -1's, and 0's. Ones represent the presence and negative ones represent the absence of some abstract feature. A zero indicates that the presence or absence of a feature is unknown or lost. For each simulation run, a new randomly generated vector is created for each of the letters used in the experiment. The collection of unique letter vectors constitutes the semantic memory of the model.

While the initial context vector is generated randomly, like the item vectors, each subsequent context was generated based on the previous context vector and a random drift factor. Each element in a new context is the same as each element in the previous context with probability (1-pDrift). With pDrift, that element is set to a random value (i.e., -1, 0, 1).

As each stimulus is processed, a vector representing that stimulus and the vector representing the current context are stored as a trace in the active subset of memory. Once the active subset has more than n traces, the model attempts to remove the traces of the items that occurred more than n stimuli prior from the active subset. The probability of removing the extra traces at each time step is pRemove. The maintenance of items in the active subset has a cost. Specifically, every trace competes with every other trace for each of its shared features. When a new item enters the active subset, there is a 50% chance that it loses each feature it shares with an item already in the active subset and a 50% chance that it keeps that feature and that the item already in the active subset loses it.

Familiarity is accessed by probing the active subset with the item portion of the current vector. To determine familiarity, the first step is to calculate the similarity of the current item to the items in the active subset by

$$S_i = \frac{\sum_{j=1}^{M} P_j T_{ij}}{N_i}, \qquad \text{Eq. 2}$$

where P_j is jth element in probe P and T_{ij} is the jth element in memory trace i. N_i is the number of elements that are non-zero in either the probe or the trace. M is the number of traces in the active subset.

The activation of each trace, A_i, is the cube of its similarity value. The echo intensity of the active subset to the probe is the sum of all these activations:

$$I = \sum_{i=1}^{M} A_i , \qquad \text{Eq. 3}$$

where M is the number of traces in the active subset. If the I is greater than 0, then the stimulus is considered familiar. Otherwise, the response is "non-match".

If the item is familiar then the recollection or recall process is initiated to determine if the current stimulus matches the stimulus n-back. This requires the n-th back context be reinstated. Each element in the current context is converted to the n-th back context with probability pReinstate. The reinstated context is used to probe the active subset by again cubing the results from Equation 2. This time, however, the context is used as the probe and activations are not used to determine the echo intensity but instead the echo content by

$$C = \sum_{i=1}^{M} A_i T_{ij} . \qquad \text{Eq. 4}$$

The echo content is a noisy version of the items most activated by the reinstated context. C will not be an exact match of any particular item. Therefore, C is disambiguated following the procedure used to disambiguate hypotheses in HyGene. This is done by recalling items from semantic memory based on their activation to C.

Semantic memory is the collection of the vectors representing each of the items used as stimuli. C is first normalized and then it is used to probe semantic memory. Once more Equation 2 is used to determine the activation but this time of semantic memory instead of the active subset. Retrieval from semantic memory is based on the activation of each item vector. The probability of sampling semantic vector i is

$$P_i = \frac{A_i}{\sum_{j=1}^{W} A_j}, \qquad \text{Eq. 5}$$

where W is the number of vectors in semantic memory.

The first item sampled from semantic memory is considered the n-th back stimulus. However, to be successfully retrieved the activation of the to-be-retrieved vector must be greater than the retrieval threshold, tRetrieval, otherwise retrieval fails and the model guesses whether or not the stimulus is a target. The probability of the model guessing target is set to the actual probability of targets in the sequence, 0.2 in the current experiment.

If retrieval is successful then the retrieved item is compared with the current stimulus. If the current stimulus matches the retrieved item, then the response is "match". If the retrieved item does not match the current stimulus, then the response is "non-match".

Familiarity, recollection, and guessing each take time. Here we assume that each take a single unit of time. Therefore, the RT predictions are completely determined by the average number of processes required to correctly and incorrectly respond to the targets, lures and other non-targets.

Simulations Results

The model was run once on each stimulus sequence given to participants at each level of n. The second panel of Figure 3 shows simulation results for 4-back. The model produces primacy, especially for targets. It also shows the same pattern of RT results as shown by participants, as shown in the second panel of Figure 5. Specifically, correct responses are made to targets and lures at approximately the same speed but responses to other non-targets are faster. Incorrect responses to other non-targets and lures are slower than incorrect responses to targets. While the detailed results are only shown for 4-back, the model predictions, like participant performance, is consistent across levels of n. The only change being that as n increases, the asymptotic level of accuracy for targets decreases for both participants and the model.

Primacy is predicted by the model due to the interference between the items maintained in the active subset of memory. Specifically, it is due to the number of other items that any given item must compete with before that item can be used to make a response. For example, when performing 4-back, the first item of the sequence only competes with the three items added after it. After the third subsequent item is added, the first item will be the n-th back stimulus to be used to make the next response. However, the fourth item in the sequence competes with at least the three items that preceded it into the active subset and the three items that followed it. The amount of interference is increased when items that are no longer relevant remain in the active subset. However, even with perfect removal of irrelevant items some degree of primacy is found.

As mentioned above, the RT predictions are completely driven by the number of processes used to make a response. For example, normally two processes are necessary to make a correct or incorrect response to a target: familiarity and recollection. Correct responses to other non-targets are quicker because they can usually be identified as non-matches by the results of the familiarity process alone. In contrast, incorrect responses to other non-targets occur primarily when the stimulus is judged as familiar but recall fails and an incorrect guess of "match" is made. Like

targets, correct lure responses often involve both familiarity and recollection, but incorrect lure responses are sometimes the result of false recollection and sometimes the result of guessing.

General Discussion

A detailed examination of n-back performance supports the claim that lures are necessary for making the task more than a familiarity judgment task (Kane et al., 2007). However, the difference in RTs between other non-targets and the two trial types in which recollection is necessary, targets and lures, indicated that the presence of lures in a stimulus sequence does not necessarily change how participants respond to the other non-target trials. The present model accounts for this RT data by assuming that the familiarity of a stimulus determines whether or not a recollection is attempted. If a stimulus is not sufficiently familiar, then the stimulus is immediately labeled a non-target. Therefore, according to the present model, correct responses on non-target trials can be accounted for exclusively by familiarity whether or not the stimulus sequence also contains lures. Only lures and targets, the trial types likely to be familiar due to their occurrence approximately n stimuli ago are likely to trigger recollection.

Other non-targets make up at least 50% of the trials in most applications of n-back, so an overall n-back score could mostly reflect the ability to discriminate familiar items. Therefore, according to the present analysis the score does not primarily reflect a participant's ability to recognize the reoccurrence of the n-th back item, but instead familiarity judgment. This is one potential reason for the low correlation between the n-back task and standard working memory assessments (e.g., operation span and reading span) in which recall is necessary.

WM is often conceptualized as having a capacity or span component as well as an executive function or attentional control component. The present modeling effort suggests that the span component of WM is not necessary to account for n-back performance, as this aspect of WM is not implemented within the model. Instead the executive function or attentional control aspect alone might be sufficient. Attentional control was implemented here as the ability to remove irrelevant information from attention (pRemove) and the ability to conduct controlled memory search (pReinstate). This might also differentiate n-back from other WM assessments, as the other tasks might rely more heavily on capacity or span.

Acknowledgments

This research was supported by the University of Maryland Center for Advanced Study of Language with funding from the Department of Defense. The authors thank Michael Bunting, Jared Novick, Scott Weems, Erika Hussey, Susan Teubner-Rhodes, and Barbara Forsyth for their contributions to the design and implementation of the experiment.

References

Atkins, S. M., Harbison, J. I., Bunting, M. F., Teubner-Rhodes, S., & Dougherty, M. R. (2009, November). *Measuring working memory with automated block span and automated letter-number sequencing*. Poster presented at the 50th Annual Meetin of the Psychonomic Society.

Gray, J. R., Chabris, C. F., & Braver, T. S. (2003). Neural mechanisms of general fluid intelligence. *Nature Neuroscience, 6,* 316-322.

Hintzman, D. L. (1988). Judgments of frequency and recognition memry in a multiple-trace memory model. *Psychological Review, 96,* 528-551.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences of the United States of America, 105,* 6829-6833.

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., Meier, B. (2010). The concurrent validity of the N-back task as a working memory measure. *Memory, 18,* 394-412.

Kane, M. J., Conway, A. R. A., Miura, T. K., & Colflesh, G. J. H., (2007). Working memory, attention control, and the n-back task: A question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 615-622.

Li, S. C., Schmiedek, F., Huxhold, O., Röcke, C., Smith, J., & Lindenberger, U. (2008). Working memory plasticity in old age: Practice gain, transfer, and maintenance. Psychology of Aging. 23, 731-742.

McCabe, J., & Hartman, M. (2008). Working memory for item and temporal information in younger and older adults. *Aging, Neuropsychology, and Cognition, 15,* 754-600.

Novick, J. M., Hussey, E., Teubner-Rhodes, S., Dougherty, M. R., Harbison, J. I., & Bunting, M. F. (submitted). Clearing the garden path: Improving sentences processing through executive training.

Oberauer, K. (2005). Binding and inhibition in working memory: Individual and age differences in short-term recognition. *Journal of Experimental Psychology: General, 134,* 368-387.

Oberauer, K., & Lewandowsky, S. (2008). Forgetting in immediate serial recall: Decay, temporal distinctiveness, or interference? *Psychological Review, 115,* 544-576.

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping, 25,*46-59.

Thomas, R. P., Dougherty, M. R., Sprenger, A., & Harbison, J. I., (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review, 115,* 155-185.