N-back Performance: Comparing Assessment and Training Performance

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 (smatkins@umd.edu)
Neuroscience & Cognitive Science Program and 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 its frequent use, much is unknown about how the n-back task is performed and how it relates to working memory. We conducted a detailed analysis of the accuracy and reaction time data from a 4-back version of the task and compared the results with previous results from an adaptive training version of the task. The experiment was also designed to test the novel predictions of a computational model of n-back performance. The assessment results were largely consistent with both the training data and the model predictions.

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

N-back and Cognition

The n-back task is used both to measure (Owen et al., 2005) and improve (Jaeggi et al., 2008) working memory (WM). It is considered a memory updating task, and updating is thought to be a core component of working memory (Miyake et al., 2000). However, the task is not consistently or strongly correlated with performance on complex working memory span tasks, such as operation span or reading span (Kane et al., 2007). Furthermore, despite transfer to measures of fluid intelligence, n-back training has not been found to transfer to other measures of WM (Jaeggi et al., 2008; Li et al., 2008).

To better understand n-back performance and its relation to WM, the present study provides a detailed analysis of 4-back data. This study builds on a previous analysis of an n-back training task (Harbison, Atkins, & Dougherty 2011) by testing if the results from an adaptive, training version of the n-back task are replicated in a non-adaptive, assessment version of the task. The present study also tests new predictions made by the computational model of n-back performance based on that training data (Harbison et al., 2011).

The N-back Task

In the n-back task participants are presented with a sequence of stimuli (e.g., letters). As each stimulus is presented, participants are asked to compare the current stimulus with the stimulus that occurred *n* items prior in the sequence. For example, in the 4-back version of the task, participants might be presented with the letter sequence "H-G-S-M-L-T-…". If the next letter in the sequence is "S" then participants should respond "target" as the current letter matches the letter occurring four letters prior. If the next letter is anything else, then the correct response is "non-target". Not all non-matching letters are the same in terms of difficulty. Lures, stimuli that match an item near to but not at the target location, are more difficult than fillers (stimuli that are neither lures nor targets). Participants are less accurate and take longer to respond to lures relative to fillers (Gray, Chabris, & Braver, 2003; Harbison et al., 2011; Kane et al., 2007; McCabe & Hartmen, 2008; Oberauer, 2005). From the example, the letters "H", "G", "M", and "L" are lures. They match the 6th, 5th, 3rd, and 2nd letter back, respectively, but not the 4th letter back. Letters such as "F", "P", and "R" are fillers.

In the training version of the n-back task the level of n varies as a function of participant performance. The n level is increased when participants perform well and decreased when participants perform poorly at their current n level. In contrast, assessment versions of the task are non-adaptive; participants are given a set number of trials at predetermined levels of n.

Previous Results

Performance on the n-back task is not often the focus of the experiments in which the task is used. Instead, the n-back task is either used to measure or to improve WM. Therefore, despite its frequent use, there remains a lack of detailed data on n-back task performance (for exceptions see Gray et al., 2003; Kane et al., 2007; McCabe & Hartmen, 2008; Oberauer, 2005).

Previously, we (Harbison et al., 2011) identified four results that characterize n-back training task performance. First, accuracy for target trials varies as a function of serial position. Figure 1a shows the results for sequences of 4-back from the training data; participants demonstrated primacy for target trials whereas this effect was weak to non-existent for lure and filler trials. Here the lures were one position away from the target, so they matched either the 3rd or 5th back stimuli. Second, in the reaction time (RT) data, we found that participants were faster making correct than

incorrect responses on lure and filler trials. This was not found for target trials. Figure 2a shows the mean RT data from 4-back sequences of the training data. Third, correct responses to targets and lures were made at approximately the same rate. Fourth, and perhaps least surprising, we found that participants made correct responses more quickly to filler stimuli than to either targets or lures. While only the results from 4-back are shown, the results are generally consistent across n levels of 3- to 7-back in the training data, with minor discrepancies at 1-, 2-, and 8-back.

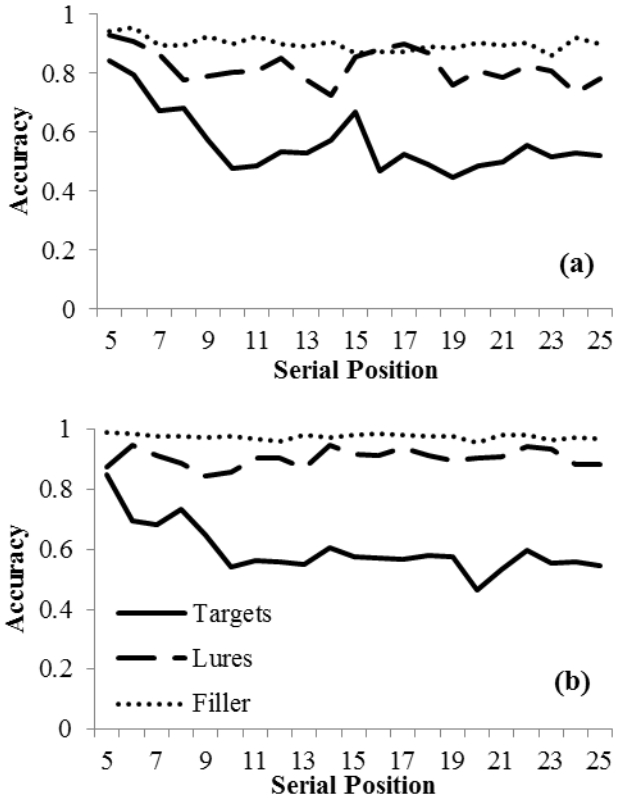

Figure 1. Participant (a) and Model (b) accuracy across serial position from the 4-back performance in a training experiment (Harbison et al., 2011).

We developed a two-process model of recognition to account for these accuracy and RT results (Harbison et al., 2011). The model assumes that when each stimulus in a sequence is presented, participants first generate an estimate of familiarity. If the stimulus is not familiar, the response is "non-target". If the stimulus is familiar, then an attempt is made to determine if it does indeed match the stimulus occurring n items back through the process of recollection. If the recollected item matches the current stimulus, then a "target" response is made. If the recollected item does not match, then the "non-target" response is made. Finally, if recollection fails, the model guesses. RT predictions are based on the number of processes necessary to respond (familiarity = 1, familiarity and recollection = 2, familiarity, recollection, and guessing = 3). The model's performance on the 4-back training stimuli is shown in Figures 1b and 2b. The model captures the main qualitative patterns observed in the participant data. For example, according to the model the observed primacy for targets is due to the interference of previous items in the sequence on the maintenance of subsequent items (i.e., proactive interference). While both targets and lures are reliant on the same processes, familiarity and recollection, primacy is predicted more for targets as participants are expected to be much more likely to guess "non-target" than "target" when recollection fails as targets are much less frequent (only 20% of the stimuli are targets) Therefore, guesses are most likely to lead to correct responses for lure trials and incorrect responses for target trials.

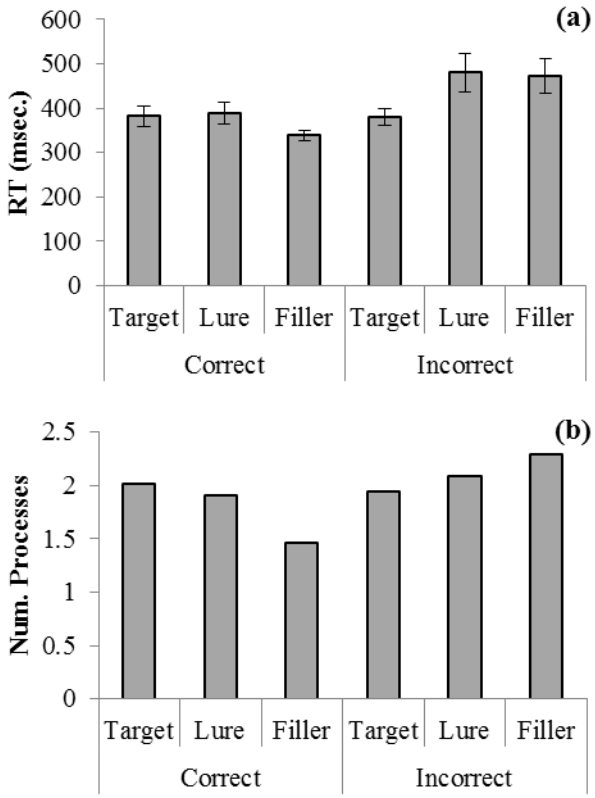

Figure 2. Participant (a) and Model (b) reaction time data from 4-back performance in a training experiment (Harbison et al., 2011).

N-back Model Details

The n-back model is implemented within the HyGene framework (Thomas et al., 2008) and consists of three components: a representation of the current stimulus, the active subset in memory, and semantic memory. Stimuli are represented by a concatenation of an item vector and the current context vector. The elements in both the item and context vectors can take on the value of 1, -1, or 0. Here 0 represents lack of information about a feature, 1 indicates a feature's presence and -1 its absence. Each item's representation and the initial context vector are generated

randomly. However, the current context changes with each new stimulus. Specifically, when a new stimulus is presented each element of the current context has some probability of randomly changing to a new value. This probability is a parameter in the model (pDrift).

Familiarity The first step in processing a new stimulus is judging its familiarity to items in the active subset by

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

where S_i is the similarity of the probe (P) and the i-th trace in memory (T_i). j is the index of the element in the item representation for both the probe and the trace. N_i is the number of elements that are non-zero in the trace, the probe, or both. M is the number of traces in the active subset. The similarity is cubed to calculate the activation (A_i) of each trace. Finally, the activations of all the traces in the active subset are summed to get the echo intensity for the probe. If the echo intensity is less than or equal to 0, then the item is unfamiliar and the "non-target" response is made. Otherwise, the model moves to the recollection process.

Recollection The model attempts to recollect the stimulus that occurred in the n-th back location when the current stimulus is familiar. This is performed by first trying to reinstate the n-th back context. Each element in the current context is changed to the n-th back context with some probability, pReinstate. This is the second parameter in the model.

Next, the (partially) reinstated context is used to probe the active subset. Equation 1 is again used but now the context portion of the representation serves as the probe instead of the item portion. Also, instead of summing the activations to get the echo intensity, the activations are used to create an echo content, a noisy representation of the item that occurred with the n-th back position by

$$C = \sum_{i=1}^{M} A_i T_{ij} \ . \qquad \text{Eq. 2}$$

To identify the item from the noisy representation, the model uses the item representation from the echo content as the probe for activating the item representations stored in semantic memory. Again the results of equation 1 are used to generate the similarity and the activation, but this time semantic memory is probed and instead of using the activations to generate echo intensity or echo content, the activations are used to determine the probability of sampling and recovery from semantic memory. Specifically, the probability of sampling an item in semantic memory is calculated by

$$p(\mathrm{Sample_i}) = \frac{A_i}{\sum_{j=1}^{W} A_j}, \qquad \text{Eq. 3}$$

where W is the number of items in semantic memory. Therefore, the probability of sampling an item in semantic memory is equal to its relative activation. After sampling an item, an attempt is made to recover that item. Recovery is successful if the activation of the sampled item is greater than the threshold tRetrieval, the third parameter in the model. If the recovered item matches the current stimulus, than the response is "target". If it does not match, the response is "non-target". If retrieval fails then the model guesses.

Guessing The probability of guessing target is equal to the base rate probability of targets in the sequence. This probability was .2 in both the training study and in the present experiment.

Encoding After a response is made the current stimulus is encoded by the model. The representation of the item and the current context are stored in the active subset of memory. Each item in the active subset competes with every other item. Specifically, each feature in an item's representation can only be non-zero for one item in the active subset. This assumption is based on the process of overwriting (Oberauer & Lewandowsky, 2008). To reduce competition, the model attempts to remove irrelevant items. In the case of 4-back, any item that occurred more than 4 items prior, from the active subset is irrelevant. Each time a new stimulus is encoded an attempt is made to remove all the irrelevant items currently in the active subset of memory. The probability of removing irrelevant items is the final parameter of the model, pRemove.

Limitations of Previous Results

The results from the previous training study provided a starting point but there are a number of reasons why a replication and extension is needed. The present study is motivated by a desire to get cleaner data than is acquired from training studies. In training versions of the n-back task, the level of n fluctuates as a function of participant performance. Therefore, the amount of data that each participant provides for each level of n can vary substantially. For example, in the previous training study some participants never reached 4-back (i.e., were never successful enough at 2- and 3-back to reach 4-back). Some participants quickly advanced past 4-back to get to higher levels of n. Finally, some participants were stuck at 4-back for a while, as their accuracy was not high enough for n to increase or low enough for n to decrease. More generally, at lower levels of n, the majority of data is from participants that have the most difficulty performing the task. At higher levels of n, there is only data from participants that either excelled at the task from the beginning or participants that improved and are near the end of their training.

Another limitation of the reported training data was that it was drawn from a larger WM training study in which participants performed a number of different WM and WM-related training tasks and assessments. Extensive practice on

these tasks might have changed how they approached the n-back task.

In addition, the n-back model makes a number of predictions that are not tested by the previous data. First, it predicts gradual improvement in accuracy as lures move further from the target position. Lures one away from the target position (3- and 5-back when n is 4) should be more difficult than lures two positions away (2- and 6-back). Furthermore, lures the same distance from the target position are predicted to have the same approximate difficulty (n+2 lures = n-2 lures, n+1 lures = n-1 lures). The predictions are shown in Figure 3b. These predictions, like all other predictions presented, are made using the same parameter values as used in Harbison et al. (2011) for matching the training data (pDrift = .33, pRemove = .15, pReinstate = .75, tRetrieval = .10)

Second, unlike accuracy predictions, RT predictions are not symmetric around the target position. RTs for lures closer to the current stimulus should take longer to respond to correctly than lures further away from the current stimulus. That is, lures that match the 2-back position should take longer to reject than lures in the 6-back position. In contrast, the time it takes to make incorrect responses to 2-back and 6-back lures should not differ. These predictions are shown in Figure 5b.

We conducted a new experiment in which all participants had extensive experience at a moderately high level of n, 4-back. 4-back was chosen because in the training study most participants were able to reach that level, 4-back allowed lures two positions away that were not the immediately prior stimulus (2-back), and because the previous 4-back data showed the same reaction time profile as was shown at higher levels of n. This pattern was not as consistent at lower levels of n, specifically 1- and 2-back.

Experiment

One hundred and forty-seven participants were randomly assigned into one of two counterbalanced conditions which determined if the participants performed sequences with lures first or second. Seventy-four participants were in the lure-first condition, seventy-three lure-second. Both conditions performed 16 sequences with lures and 16 sequences without lures. Each sequence was 25 letters long and contained five targets and either eight or zero lures. When the lures were present, there were two of each type in the sequence (2-, 3-, 5-, and 6-back lures). After completing the 4-back task, participants performed the block span and letter-number sequencing (LNS) tasks as measures of WM (Atkins et al., 2009).

Results

Note that all differences reported have a p value of .05 or less. Also, unless otherwise noted, within-participant analyses were used. As such, the figures showing results averaging over participants can be misleading. Finally, there were no significant differences due to condition assignment (lures first or lures second). Therefore, order is ignored in the reported analyses.

Accuracy The mean accuracy data by trial type is shown in Figure 3a. With or without lures, participants were most accurate with filler items and least accurate with target items. Performance on lures two away from the target (2- and 6-back) was worse than filler and better than performance on lures one away (3- and 5-back) from the target position. There was not a significant difference between lures the same distance away from the target. Comparing performance on sequences with lures against sequences without lures, there was not a significant difference in target performance, but participants were significantly better on filler items when there were lures.

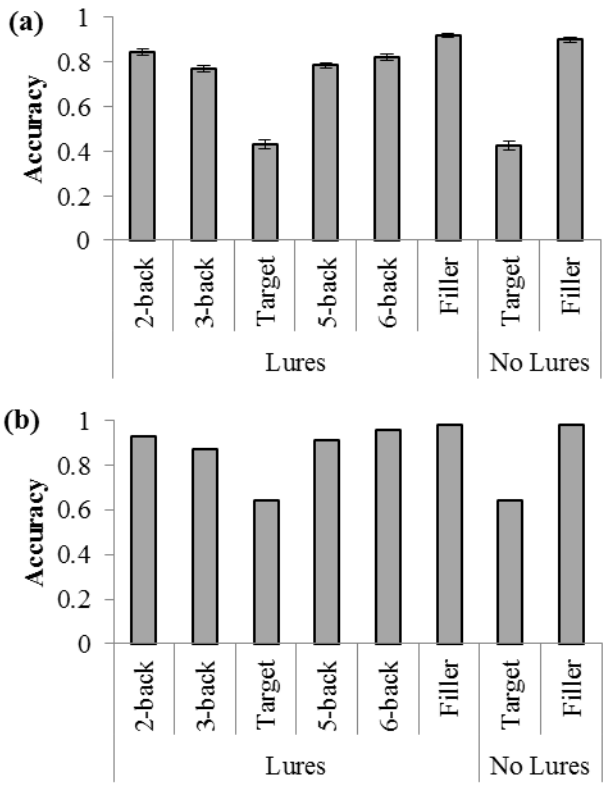

Figure 3. Participant (a) and Model (b) mean accuracy by trial type.

Serial Accuracy As shown in Figure 4a, participants showed primacy for target trials. Also, performance on lures two away from the target position were consistently better than performance on lures one away from the target position. Note that while target accuracy dropped below 50% in middle and later serial positions, this is not really chance performance, as participants would be expected to respond "target" only 20% of the time if they guessed "target" with the same probability as targets in the sequence.

Reaction Times As in the training data, participants were significantly faster to respond correctly to lure and filler items than they were to respond incorrectly, as shown in Figure 5a. In contrast, target RT was not significantly different for correct and incorrect responses. Also as in the training data, participants were quickest to respond to filler items correctly.

There was not a significant difference between 6-back and 2-back lures for incorrect responses, but there was for correct responses. This pattern of results was predicted by the model. However, there were also some inconsistencies with the previous data. Inconsistent with both the training data and the model's predictions, the present experiment found incorrect filler responses were faster, not slower, than the incorrect responses to lures, on average. Also, participants were quicker to respond to target items than predicted by the model. Both correct and incorrect target responses were significantly faster than the average lure responses.

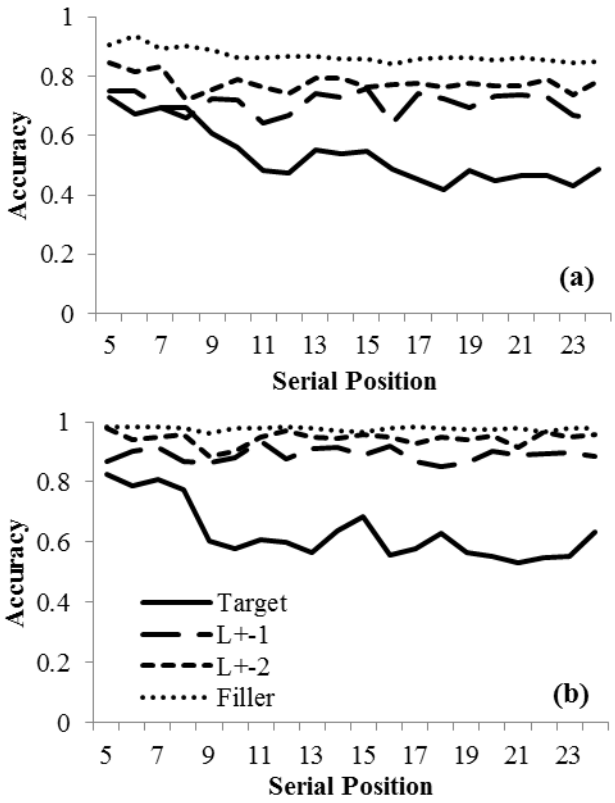

Figure 4. Participant (a) and Model (b) serial accuracy results by trial type.

Working Memory There was a weak but significant correlation of both LNS and block span with target performance (r's from 0.188 to 0.283). Lure and filler accuracy were not correlated with these WM measures (r's < 0.135). This result is consistent with previous assessment versions of the n-back Oberauer (2005) but not previous training data (Harbison et al., 2011) which found the relationship with lure but not target performance.

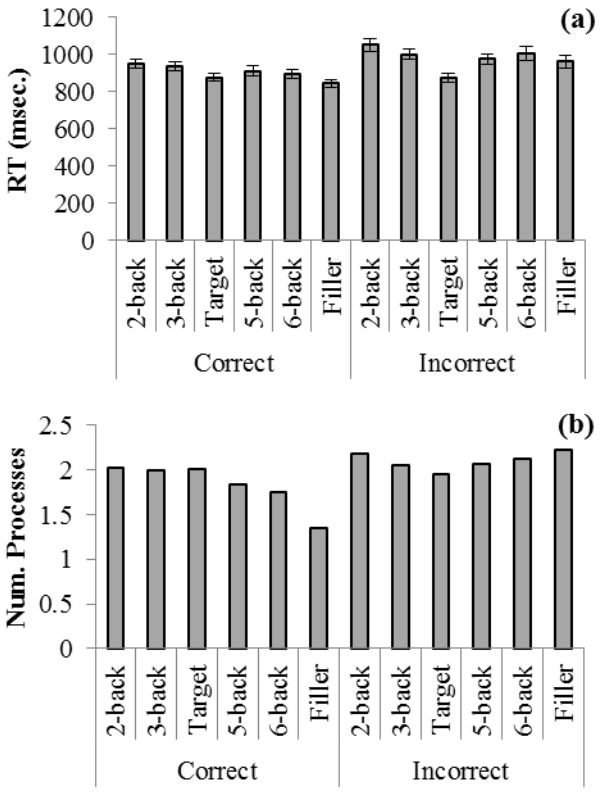

Figure 5. Participant (a) and Model (b) reaction time results by trial type and accuracy.

Discussion

The results from the 4-back task are largely consistent with the results of the adaptive, training version of the n-back task where difficulty is adjusted based on participant performance. In the present experiment three of the four results were replicated: target accuracy showed primacy, incorrect responses took longer than correct responses for lure and filler stimuli, and correct responses to filler items were faster than responses to any other trial type. In these ways the results are consistent with both the training data and the n-back model that was based on the training data.

In addition, the new data supported two novel predictions made by the n-back model. First, lure accuracy fit the predicted pattern, with lures one away from target position being more difficult than lures two away from the target position, while lures the same distance away were performed with approximately the same accuracy. Second, reaction times were predicted by the model to be longer for correct responses to 2-back than 6-back lures despite equivalent accuracy and equivalent RT for incorrect responses. It should be noted that the model was constructed using training results with lures only in positions one away from the target position, and yet was able to accurately

predict performance on lures two away from the target position both in terms of accuracy and reaction time.

The results were not without discrepancies. Participants in the present experiment were faster at responding both correctly and incorrectly to targets than either found in the training data or predicted by the model. Also, incorrect filler responses were not the slowest overall responses in the present data. Instead, incorrect lure responses were the slowest. Both of these results, plus the fact that participants showed a different RT profile in the training version of the n-back task at n levels of 1 and 2 from the general trend found at n levels 3 and above indicate the model is at best incomplete. One natural extension of the model which could account for at least some of these results is to include the area of direct access in addition to the activated subset of long-term memory currently implemented (Cowan, 1988). Items in the area of direct access would be able to forgo the recollection process, as they would be immediately available.

The present study provides additional support for the account of n-back performance as driven by recognition processes. Both target (Oberauer, 2005; the present study) and lure (Harbison et al, 2011) performance have been found to correlate with other WM assessments. Both of these trial types rely on recollection according to the present model. In contrast, filler trial performance can be accounted for by familiarity alone and has not been found to be related to other measures of WM. This could account for the inconsistent and/or weak relationship between overall n-back performance and other measures of WM as a large portion, often more than half, of the stimuli in a given n-back sequence are filler trials.

The purpose of this study is to improve the understanding of the cognitive mechanisms behind performance on the n-back task. As with other working memory tasks which correlate with many higher level cognitive processes, it is important to determine what is being measured by the n-back assessment and what might be improved by training versions of this task (Shipstead, Redick, & Engle, 2012). The results suggest that the relationship between n-back performance and other measures of working memory are dependent on a specific process, recollection, or the ability to comply with the demands of the task and inhibit responses based on familiarity alone in order to use recollection as the basis for response.

Acknowledgments

This research was supported in part by the University of Maryland Center for Advanced Study of Language with funding from the Department of Defense.

References

Atkins, S. M., Harbison, J. I., Bunting, M. F., Tuebner-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 Meeting of the Psychonomic Society.

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information processing system. *Psychological Bulletin, 194,* 163-191.

Gray, J. R., Chabris, C. F., & Braver, T. S. (2002). Neural mechanisms of general fluid intelligence. *Nature Neuroscience, 6,* 316-322.

Harbison. J. I., Atkins, S. M., & Dougherty, M. R. (2011). N-back training task performance: Analysis and model. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp.120-125). Austin, TX: Cognitive Science Society.

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.

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.

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

Miyake, A.**,** Friedman, N.P., Emerson, M.J., Witzki, A.H., Howerter, A., Wager, T. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41,* 49–100.

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.

Shipstead, Z., Redick, T. S., & Engle, R. W. (2012, March 12). Is working memory training effective? *Psychological Bulletin*. Advanced online publication. doi:10.1037/a0027473.

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