RUNNING HEAD: RECALL OUTPUT, MMSE AND HIPPOCAMPUS

Output order and variability in free recall are linked to cognitive ability and hippocampal volume in elderly individuals

Davide Bruno[1], Michel J. Grothe[2], Jay Nierenberg[3,4], John J. Sidtis[3,4], Stefan Teipel[2,5], Nunzio Pomara[3,4]

[1] Department of Psychology, Liverpool Hope University, Liverpool, UK

[2] German Center for Neurodegenerative Diseases (DZNE) – Rostock/Greifswald, Rostock, Germany

[3] Nathan Kline Institute for Psychiatric Research, Orangeburg, NY, USA

[4] Department of Psychiatry, School of Medicine, New York University, New York City, NY, USA

[5] Department of Psychosomatic Medicine, University of Rostock, Rostock, Germany

Corresponding Author:

Davide Bruno

Department of Psychology, Liverpool Hope University

Hope Park, Liverpool, L16 9JD

Phone: 0044 (0)151 2913832

Email: brunod@hope.ac.uk

Abstract

Adapted from the work of Kahana and colleagues (e.g., Kahana, 1996), we present two measures of order of recall in neuropsychological free recall tests. These are the position on the study list of the first recalled item, and the degree of variability in the order in which items are reported at test (i.e., the temporal distance across the first four recalled items). We tested two hypotheses in separate experiments: 1) whether these measures predicted generalized cognitive ability, and 2) whether they predicted gray matter hippocampal volume. To test hypothesis 1, we conducted ordinal regression analyses on data from a group of 452 participants, aged 60 or above. Memory performance was measured with Rey's AVLT and generalized cognitive ability was measured with the MMSE test. To test hypothesis 2, we conducted a linear regression analysis on data from a sample of 79 cognitively intact individuals aged 60 or over. Memory was measured with the BSRT and hippocampal volume was extracted from MRI images. Results of Experiment 1 showed that the position of the first item recalled and the degree of output order variability correlated with MMSE scores only in the delayed test, but not in the immediate test. In Experiment 2, the degree of variability in the recall sequence of the delayed trial correlated (negatively) with hippocampal size. These findings confirm the importance of delayed primacy as a marker of cognitive ability, and are consistent with the idea that the hippocampus is involved in coding the temporal context of learned episodes.

Keywords: Free recall; Output order; MMSE; Hippocampus.

Introduction

The input order of free recall has long been a subject of research in psychology (Murdock, 1962; Glanzer, 1972). Particularly, we know that items presented at the beginning of the list (primacy) and items presented at the end of the list (recency) are remembered better than middle items, and this is known as the serial position effect. Importantly, the serial position effect has also been examined in neuropsychological studies as a potential tool to predict future changes in performance in elderly participants, and especially to anticipate cognitive decline. Primacy performance, in particular, has been singled out as a very sensitive cognitive marker of prospective cognitive impairment. For example, Egli et al. (2014) have shown that poor performance on items presented at the primacy position is associated with a greater risk of conversion from mild cognitive impairment (MCI) to Alzheimer's disease (AD). Analogously, Bruno, Reiss, Petkova, Sidtis and Pomara (2013) have shown that poor primacy performance in delayed free recall predicts generalized cognitive decline in cognitively intact participants at baseline over a span of roughly five years, and does so better than performance in other regions of the serial position. Additionally, Bruno et al. (2015) have shown that delayed recall of primacy items correlated preferentially with the volume of hippocampal gray matter tissue in the brains of cognitively intact elderly individuals.

Bruno et al. (2013; 2015) have interpreted these findings as suggesting that recall performance of primacy items is a sensitive marker of cognitive decline and degeneration because it relies on hippocampal function, which in turn is affected early in neurodegenerative diseases, AD in particular (Raj et al., 2015). However, there remains a lack of understanding of the neurocognitive mechanisms underlying a preferential relationship between hippocampus and primacy, and ultimately of why delayed primacy performance is a strong predictor of future

cognitive decline. In this paper, we attempt to shed some light on these issues by examining the *output* order of free recall. Studies of the output order of recall (i.e., the order in which items are reported at test) have been far fewer than those of the input order. However, to paraphrase Howard and Kahana (1999), focusing only on the serial position tells merely half the story, since retrieval/output processes exert an influence that can be as important as that of encoding/input processes in the overall dynamic of free recall. Therefore, we posit that in order to understand the relationship between primacy and hippocampus, and primacy and decline, output order processes also need investigation.

Following on from the seminal work of Kahana (1996), a series of studies has investigated the patterns of recall output in different laboratory-based conditions (e.g., Kahana & Howard, 2005; Klein, Addis & Kahana, 2005; Sederberg, Howard & Kahana, 2008), across older and younger participants (e.g., Golomb, Peelle, Addis, Kahana & Wingfield, 2008; Kahana, Howard, Zaromb & Wingfield, 2002), and comparing individuals with anterograde amnesia to controls (Talmi, Caplan, Richards & Moscovitch, 2015). For instance, Kahana et al. (2002) showed that older, cognitively intact adults tend to begin recall largely in the same way as younger controls, but are less likely to use, and presumably benefit from, information on the temporal associations between presented items (i.e., less likely to retrieve contiguously presented items together at recall). Kahana et al. interpreted this finding as consistent with an age-related deficit in associative memory (Naveh-Benjamin, 2000).

We addressed the importance of the output order of recall to explain delayed primacy performance in two studies on separate cohorts. In the first study, we examined the relationship between indices of the output order of recall and generalized cognitive ability, as measured with the Mini-mental State Exam (MMSE; Folstein, Folstein & McHugh, 1975) test, to establish

whether output order indices are sensitive to general levels of cognitive ability. Additionally, we determined whether these indices are correlated with delayed primacy performance. If delayed primacy performance is sensitive to cognitive ability and decline, as previously demonstrated, and indices of the output order of recall are associated with both delayed primacy and general cognition, then the prerequisite for these indices to contribute to explaining the underlying mechanisms of primacy performance is set.

In the second study, we evaluated whether the indices of output order of recall were associated with hippocampal gray matter volume in a group of cognitively intact older individuals. The hippocampus and surrounding medial temporal lobe areas are affected early in AD (Raj et al., 2015), and thought to be involved in coding the temporal order information (Howard et al., 2005; Manning, Polyn, Baltuch, Litt & Kahana, 2011). Since delayed primacy performance was observed to associate preferentially with the hippocampus, we expect that relevant output order measures will also associate with hippocampal volume if related to primacy performance.

Typical output measures (e.g., Howard & Kahana, 1999) are obtained by presenting participants with multiple (different) study lists to learn, and then are calculated accounting for these unrelated trials (e.g., as a recall probability). However, this solution is not practical under all circumstances, since multiple study lists may be unavailable or undesirable (e.g., incidental memory tasks). This is also the case with standardized neuropsychological tests, where participants are generally required to learn multiple instances of the *same* study list. For this reason, we are employing adapted measures of output order in the present manuscript.

For the present research, we selected two measures, both of which require relatively simple computations and can be extracted by any common neuropsychological test of memory (e.g.,

AVLT). Our first aim was to pin point where recall was initiated on the serial position. Since primacy depends on retrieving items early on the list, we postulated that initiating recall from the primacy region should be associated with better retrieval of primacy items. Therefore, the first measure of output order is the study order position of the first item reported at test (i.e., first item reported; FIR). For example, if items TABLE, PINEAPPLE, MOUSE are studied in this order, and PINEAPPLE is recalled first, the recorded output order is 2 (i.e., FIR=2); if TABLE is recalled first, the output order is 1 (i.e., FIR=1); if MOUSE is recalled first, the output order is 3 (i.e., FIR=3), and so on. This measure is adapted from the probability of first recall (PFR; e.g., Kahana, 1996). FIR is a measure of which studied item is most memorable or distinctive to the participant, and can be used to index how recall is initiated. Typically, with lists of 10 items or more, participants tend to initiate recall in immediate tasks,with the last presented item, showing an output-order recency effect, whereas in delayed recall tasks, participants will usually initiate recall from the beginning of the list, showing an output-order primacy effect (e.g., Howard & Kahana, 1999), somehow analogously to the recency-primacy shift (Brown & Lewandosky, 2010). In addition, FIR should also impact on the standard recency and primacy effects, as retrieving items from each portion of the study list first increases the chances that more items from that portion will be recalled overall (see also output-order variability, below). Therefore, following Bruno et al. (2013; 2015), it would be expected that individuals at lower risk of cognitive decline would be more likely to initiate delayed recall tasks with primacy items. More precisely, we predict that lower FIR in delayed tasks should correlate with higher MMSE scores and more hippocampal gray matter volume.

Our second aim was to measure whether, once recall was initiated, participants were then likely to continue recall by reporting items that were contextually contiguous on the study list

(e.g., report items learned around the same time), perhaps by using temporal context information as a retrieval cue. As noted (Kahana et al., 2002), older cognitively intact participants struggle to do that more than younger controls, thus suggesting that contiguous recall is sensitive to age and, possibly, cognitive impairment. Moreover, if participants initiate recall from the beginning of the study list, but do not report temporally related items, then primacy recall will not be high despite beginning with a primacy item. Therefore, the second measure of output order is based on Howard and Kahana's (1999) conditional response probability, which measures the relationship and temporal distance between subsequent instances of recall. Using the previous example, a participant studies TABLE, PINEAPPLE and MOUSE, in this order; PINEAPPLE is recalled first (2), and then MOUSE (3) is recalled. The temporal distance between PINEAPPLE and MOUSE is one unit (3-2=1) and the items are recalled consecutively in a forward trajectory. In contrast, recalling TABLE (1) and then MOUSE, while maintaining a forward trajectory, leads to a greater gap between studied positions (3-1=2), as the items were not learned consecutively. Focusing on the distance between studied positions in the output order and, again, relying on a single trial, it is possible to obtain a measure of absolute (i.e., independent on trajectory) output order variability by calculating the absolute distance between study order positions. For practicality, we will call this measure output variability (OV, or OV-X with X being the number of items this is calculated on). The number of retrieved items for OV was set to four in the present paper (i.e., OV-4; OV was calculated over the first four recalled items) to ensure a balance between variability in the output order of recall and sample size.[1] OV is a test of associative memory in that it indices a person's ability to form temporal connections between

[1] For reference, OV-4 allowed us to retain 64% of the total 987 participants, not counting inclusion criteria (see Participants below); in contrast, OV-5 would have left us with 57% of the participants, and OV-6 with only 50% of the participants.

items: the lower the variability, the better this ability is. As shown by Sederberg, Miller, Howard and Kahana (2010), lower variability tends to correlate with better performance in episodic memory tests. Therefore, it is expected that lower OV values should also correlate with greater generalized cognitive ability and higher gray matter hippocampal volumes.

Experiment 1

Methods

Participants. A total of 987 volunteers from the Memory Evaluation Research Initiative (MERI) program (Reichert, Sidtis & Pomara, 2015) were available for Experiment 1. These participants were enrolled at the Nathan Kline Institute in Orangeburg, NY. From this total pool, we extracted 617 participants who matched our inclusion criteria: over 60 years of age; no prior diagnosis of dementia; MMSE score of 24 or higher (one of three criteria used in the ADNI 2 study to define a participant as cognitively normal – of note, selecting participants above this cut-off score does not exclude that some may be affected by memory age-related impairment or incipient neuropathology); and recalling at least one item at both the immediate and delayed recall trials. Finally, owing to the fact that not all participants recalled four or more items at both the immediate and delayed trial, the final sample included a total of 452 valid cases (see Table 1 for demographic information), of whom 279 were females (62%). Previous diagnoses of medical conditions associated with increasing age were also reported in a portion of the sample. The individuals were otherwise asymptomatic and medically stable at the time of testing. Seventy-three participants (16.2%) reported being diabetic, and 222 participants (49.1%) had hypertension. Subjects received no compensation for this study. The MERI program has received ethical approval by the institutional review board of the Nathan Kline Institute.

Procedure. The study procedure is the same as previously reported in Bruno et al. (2013). In

summary, after providing informed consent, participants underwent a vitals examination, a blood draw, and a general medical intake questionnaire, including the MMSE test. Memory performance was tested subsequently with the Rey Auditory Verbal Learning Test (AVLT), which was part of a wider neuropsychological battery lasting ~ 2 hours. In the AVLT, participants are read a list of 15 unrelated words and then are asked to free recall the word items immediately (immediate recall trial; trial 1). Subsequently, this process is repeated four more times with the same words. After 20-25 minutes, the participants are then again tested for their memory of the item list with free recall instructions (delayed recall trial). Three to six (depending on whether the test was carried out before or after June 2014, respectively) alternative versions of the study lists were available, and one was assigned randomly to each participant.

Study Design and Analysis. In order to identify output order targets for analysis, we first ran bivariate correlations between the output order indices and delayed primacy performance. These indices were the immediate recall trial FIR index (iFIR), the delayed recall trial FIR index (dFIR), the immediate recall trial OV index (iOV-4) and the delayed recall trial OV index (dOV-4). Subsequently, we set out to determine whether the relevant output order indices predicted the MMSE score. The outcome variable was the MMSE score; and the predictors were age, sex, reported years of formal education, and the selected output order indices. Due to the heavily skewed shape of the distribution of MMSE scores, an ordinal regression analysis was employed.

Table 1 here

Results

Figure 1a and 1b report the shape of iFIR and dFIR. A visual inspection of these plots suggests that both measures largely mirror typical serial position shapes. In immediate recall (iFIR), both a primacy and a recency effect are observed; in contrast, in delayed recall, the recency effect is

absent while the primacy effect is preserved. The median for iFIR was 10 (range: 1-15) and the median for dFIR was 2 (range: 1-15). Means (and SDs) of the output order variability measures, which produced shapes closely resembling a normal distributions, were 13.65 (6.61; range: 3-38) for iOV-4 and 11.83 (6.40; range: 3-36) for dOV-4.

Delayed primacy performance. Unsurprisingly, dFIR was negatively correlated with delayed primacy performance (r=-.381, p<.001) and so was dOV-4 (r=-.295, p<.001), which, in turn, were also correlated with each other (r=.271, p<.001). In contrast, neither iFIR (r=.020, p=.681) nor iOV-4 (r=-.021, p=.662) was significantly associated with delayed primacy performance. On this basis, dFIR and dOV-4 were selected for further analyses. Due to possible multicollinearity issues, these analyses were conducted separately on dFIR and dOV-4, and so the α level was set to 0.025 following Sidak correction. Bivariate correlations between dFIR and age (r=0.145, p=.002), and dFRI and level of education (r=-0.102, p=.030) were both significant. In contrast, bivariate correlations between dOV-4 and age (r=0.032, p=.495), and dOV-4 and level of education (r=-0.090, p=.055) did not reach significance.

FIR. The ordinal analysis with dFIR fit well (p<.001) and had a pseudo R^2 of 0.117. Age (p<.001, Wald t = 21.285, coefficient estimate = -.061) and Education (p<.001, Wald t = 11.531, coefficient estimate = .112) were significant predictors, suggesting that the MMSE score decreases with age, and increases with more formal education. In addition, dFIR was also significant (p=.002, Wald t = 9.246, coefficient estimate = -.062), suggesting that participants who initiate recall with primacy items in the delayed trials tend to have higher MMSE scores.

OV-4. The analysis with dOV-4 provided similar results. The fit was good (p<.001) and a pseudo R^2 of 0.110 was provided. Age (p<.001, Wald t = 24.573, coefficient estimate = -.065) and Education (p=.001, Wald t = 11.993, coefficient estimate = .114) were again significant

predictors. Moreover, dOV-4 also was a significant predictor (p=.014, Wald t = 5.639, coefficient estimate = -.034), indicating that more variability in the delayed trial is negatively associated with the MMSE score.

Experiment 2

The results of Experiment 1 suggest a relationship between dFIR and dOV-4, and both delayed primacy performance and the MMSE score. These results suggest that delayed measures of output order may be helpful in understanding the mechanisms underlying the relationship between delayed primacy performance and generalized cognitive ability/decline. In Experiment 2, we set out to establish whether these indices also correlated with hippocampal gray matter volume, which was found previously to predict delayed primacy performance in cognitively intact elderly individuals (Bruno et al., 2015).

Methods

Participants. The sample for Experiment 2 was made up of two cohorts, described already in Bruno et al. (2015). These studies received ethical approval by the institutional review boards of the Nathan Kline Institute and the New York University School of Medicine. In one cohort, participants were recruited via the MERI program at the Nathan Kline Institute as part of a study on late-life depression (Pomara et al., 2012). All participants provided formal consent prior to testing and were paid up to $450.00 for their participation in the study. A total of 133 participants were recruited originally for the study. For the purpose of the current analysis, we excluded all subjects who presented evidence of confluent deep or periventricular white matter hyperintensities in the MRI; had an MMSE score below 28; or were diagnosed with major depressive disorder, leaving us with a total of 54 participants. For the second cohort, a total of 76 participants were recruited originally for a study on Lorazepam and APOE alleles on cognition.

Participants provided informed consent prior to testing and were paid $200 for their participation; none showed signs of cognitive impairment, significant neurological or medical illnesses, or were currently using psychotropic medications. Medical conditions associated with increased aging, such as hypertension or diabetes, were reported as previously diagnosed in a subset of participants, who were otherwise asymptomatic and medically stable at the time of testing. All also had an MMSE score of 28 or higher and a Clinical Dementia Rating of 0. Included in this analysis are the 28 participants who received an MRI scan of the head. Table 2 reports the population demographics split by cohort (see also Bruno et al., 2015). Due to two participants not retrieving at least four or more items in the delayed trial, and one participant not disclosing years of education, the total number of valid cases for the analysis was eventually 79.

Table 2 here

MRI Acquisition. The MRI acquisition was performed on a 1.5 T Siemens Vision system (Erlangen, Germany) at the Nathan Kline Institute. For more information on the specifications, please refer to published methodologies in Bruno et al. (2015).

MRI preprocessing and analysis. MRI data processing followed procedures described previously (Bruno et al., 2015) and illustrated in Figure 1. Briefly, MPRAGE images were segmented into gray matter, white matter, and cerebrospinal fluid partitions and high-dimensionally registered to Montreal Neurological Institute (MNI) standard space, using a segmentation routine without reliance on tissue priors and the diffeomorphic DARTEL warping algorithm (Ashburner, 2007), respectively, both implemented in the VBM8-toolbox. Warping parameters were applied to individual gray matter maps and voxel values were modulated to account for the volumetric differences introduced by the high-dimensional warps, such that the total amount of gray matter volume present before warping was preserved.

Individual gray matter volumes of the hippocampus were extracted automatically from the warped gray matter segments by summing up the modulated voxel values within a predefined hippocampus mask in template space. This mask was obtained by manual delineation of the hippocampus in the MNI standard space template used for high-dimensional image normalization in the VBM8 toolbox. Tracing of the hippocampus outlines followed recently developed international consensus criteria for manual hippocampus segmentation on MRI (Frisoni et al., 2015; http://www.hippocampal-protocol.net/SOPs/index.php) and was performed by a certified tracer (MJG) using MultiTracer 1.0 software (http://www.loni.usc.edu/Software/MultiTracer). Figure 2 illustrates the hippocampal regions of interest (ROI). The total intracranial volume (TIV) was used in the statistical model to account for differences in head size (see below), and was calculated as the sum of the total segmented gray matter, white matter and cerebrospinal fluid volumes in native space.

Figure 2 here

Procedure. The first group of participants was tested at the Nathan Kline Institute and at the New York University Langone Medical Center, over three visits on three successive weeks. On the first one, participants provided informed consent, and were then administered a general medical intake questionnaire; subsequently to this, their vital signs were examined and the MMSE score was collected. On a second visit, participants received an MRI scan of the head. On a third and final visit, participants underwent a comprehensive neuropsychological assessment, including the Buschke Selective Reminding Test (BSRT; Buschke & Fuld, 1974) to assess memory performance. The BSRT requires participants to be read a list of 16 unrelated nouns. After presentation, participants are asked to recall as many words as possible and to indicate when no more words can be recalled. Two trials are relevant for the current analysis, as in Experiment 1:

in the first trial (immediate recall), participants are asked to free recall as many words as possible immediately after presentation; in the delayed trial, participants are asked to free recall the original study list after a 20-25 minutes gap. The second group was examined at the Nathan Kline Institute. All relevant tests were conducted in a single session, after obtaining vital signs. These tests included a comprehensive neuropsychological assessment; memory was again assessed with the BSRT.

Study Design and Analysis. To test whether hippocampal size was associated to delayed recall measures of output order (dFIR & dOV-4; uncorrelated, r=.002, p=.985), a multiple linear regression analysis was carried out. The outcome variable was hippocampal gray matter volume, which was reasonably normally distributed in our data (*skewness/error* = -0.59; *kurtosis/error* = 0.62). The analysis was carried out with a three model structure. In Model 1, Age, Cohort, Sex, Years of Education, and TIV were entered as predictors; in Model 2, dFIR was entered as a predictor; and in Model 3, dOV-4 was finally added as a predictor. As dFIR and dOV-4 did not correlate, they were considered safe to be included in the same analysis.

Results

The mean of dFIR was 4.89 (SD = 4.96, range: 1-16) and the mean of dOV-4 was 14.57 (SD = 6.32, range: 4-31). All models fit the data satisfactorily (p's < .001). Adding dFIR (Model 2) did not significantly contribute to variance explained [$F(1,72)<.001$, p=.993, $\Delta R^2<.001$], whereas adding dOV-4 (Model 3) did result in a significant effect [$F(1,71)=4.458$, p=.038, $\Delta R^2=.032$]. The relationship between dOV-4 and hippocampal size was, as predicted, negative [β=-.186, partial R=-.243] (see Figure 3). Among the remaining predictors, the only relevant association to report is that with Age [β=-.285, partial R=-.352, p=.002].

Figure 3 here

Discussion

In two experiments, we set out to examine the relationship between measures of the output order of recall and indices of cognitive and brain health. Our goal was to improve our understanding of the mechanisms underlying delayed primacy performance, which has been shown to be a strong predictor of cognitive decline in elderly individuals (e.g., Bruno et al., 2013). In Experiment 1, we showed that measures of output order based upon delayed free recall performance – dFIR, which measures how recall is initiated, and dOV-4, reflecting the influence of temporal order information on recall – were associated with generalized cognitive ability (as measured by the MMSE score) in individuals aged 60 or higher. Lower values of dFIR, indicating that delayed recall begins by reporting primacy items, were found to associate with higher MMSE scores. Moreover, lower values of dOV-4, showing that individuals tend to retrieve items from contiguous study order position, thus suggesting usage of temporal order information, were also found to correlate with higher MMSE scores. These two findings are consistent with previous research highlighting the importance of delayed primacy as a predictor of cognitive decline (Bruno et al., 2013). Participants with better cognitive ability a) tend to begin recall from primacy positions (i.e., early in the study list), and b) tend to retrieve items from contiguous study positions; it follows then that these same individuals will also retrieve more primacy items. To note also is the emphasis on the *delayed* task, as opposed to the immediate recall trial, which is thought to tap into the individual's ability to consolidate information (Bruno et al., 2015; Gomar, Bobes-Bascaran, Conejero-Goldberg, Davies & Goldberg, 2011).

The finding that individuals who tend to retrieve items from contiguous study order positions, thus suggesting effective usage of temporal order information, were found generally to

present higher MMSE scores is consistent with the idea that the hippocampus, an area that has been found to be highly susceptible to neurodegeneration (Raj et al., 2015), is involved in encoding and maintaining the temporal context information (Davachi & DuBrow, 2015; Howard et al., 2005; Hsieh, Gruber, Jenkins & Ranganath, 2014). This latter claim was specifically tested and confirmed in Experiment 2, where we showed that hippocampal gray matter volume in two groups of cognitively intact elderly participants was negatively associated with dOV-4, suggesting that greater hippocampal size is linked to frequent recall of items from nearby study positions.

Manning et al. (2011) had noted previously that the neural signature for temporal context was strongly represented in the medial-temporal lobe, which includes the hippocampus, consistently with our findings. However, Manning et al. did not find this signature to correlate with primacy performance. Somewhat analogously, in our results, although we did find dOV-4 to correlate with hippocampal volume, the measure of recall initiation more likely to associate with whether items are retrieved from the primacy region of the serial position, i.e., dFIR, was not correlated with hippocampal volume. Therefore, it is possible to speculate, based on these findings, that delayed primacy performance may rely upon at least two distinct mechanisms, each relying on a separate neuronal structure. First, in order for primacy items to be retrieved successfully, participants begin recall most often by reporting early-list items first. This mechanism does not appear to depend heavily upon the hippocampus, and we suggest it may be related to attention and frontal activity: participants who are quick to focus on a new set of items will pay special attention to early-list stimuli; in contrast, participants who are slower, and perhaps carry some degree of attentional deficit, will start processing the incoming stimuli later, thus causing a shift in focus towards mid-list items. Second, once an item is retrieved, the study

temporal context can be activated, thus cueing retrieval of contiguous items; this process appears to rely on the hippocampus. Future research is needed to consolidate this proposal particularly by helping identify the neuronal correlates of dFIR.

As noted, the measures of output order that we adopted in the present manuscript were inspired by the work of Kahana and colleagues (e.g., Howard & Kahana, 1999), but were not identical to the original indices. This is because our memory data were based on single-list recall tests, rather than upon learning multiple, unique word lists. Although this solution may not be ideal for isolating order effects since performance on a single list is more likely to result in noisy data, we believe our proposed measures are indeed largely comparable to the original ones. Our proposed measures represent, in our view, a practical solution for experimental or clinical studies in which only one item list is learned by the participants (e.g., incidental memory tasks; studies employing neuropsychological test batteries). Consequently, these indices could be useful as part of the neuropsychological assessment of elderly individuals. The dFIR and dOV-4 measures can be calculated relatively easily from the delayed recall trial of different standardized memory tests, such as the AVLT (Exp. 1), the BSRT (Exp. 2), and other list learning tests (e.g., HVLT); lend themselves to comparisons across different memory tests (e.g., AVLT vs. BSRT[2]); and are shown in our present manuscript to be sensitive to changes in cognitive ability across individuals. However, one limitation of dFIR and dOV-4 is that they require retrieval of at least one or four items, respectively. Therefore, participants who do not recall a sufficient amount of items are ruled out from benefiting from these measures. Having said that, a complete or

[2] In the BSRT, items are presented again (i.e., reminded) to the participant in trials 2-7 only when they have not been previously recalled. Therefore, unlike with the AVLT (Exp. 1) for example, only *some* words are presented again to the participant. This is not ideal for our purposes since it introduces inconsistency across participants with regards to the sequence in which items were learned, as well as the overall number of repetitions. Although this does not affect trial 1 performance, it may have an effect on delayed performance. However, the fact that we have previously shown delayed serial position effects, with this test, on hippocampal gray matter volume (Bruno et al., 2015) suggests to us that this limitation only has a limited impact on the results.

substantial lack of recall is likely to be a sign of cognitive impairment and therefore represent in and of itself a sufficient indication that a person may be suffering from some degree of neurodegeneration, hence rendering the need for subtle cognitive testing largely moot.

Another potential limitation of our study is the use of the MMSE in Experiment 1. The MMSE is a test that includes three tasks (i.e., Recall, Registration and Repetition), out of eight in total, that are meant to measure memory function. Therefore, it is possible that by attempting to predict the MMSE score with *other* memory measures, we are merely correlating a function to itself. In order to test this claim, we generated a composite measure of global cognition by combining a series of available cognitive outcomes, for which we obtained standardized scores within our sample; these measures were: the MMSE score; the (scaled) score of the digit symbol substitution test (DSST); the FAS test (animals); the total recall score from the AVLT; and the Trail Making score (B-A). All values were added to each other except the Trail Making score that was subtracted. The composite score data distributed fairly normally and, thus, we carried out two separate linear regression analyses following Experiment 1: Model 1 included age, sex and level of education; Model 2 included either dFIR or dOV-4. The α level was again set to 0.025, and due to missing data in the FAS test, the DSST and Trail Making, the total sample of this analysis was 276 participants. Results confirmed the original findings. In the dFIR test, both models fit the data well (*p's*<0.001); all the predictors in Model 1 were significant, suggesting that the composite score was effective in tracking global cognitive ability (e.g., it decreased significantly with age, and was positively related to level of education); and most critically, dFIR was negatively correlated with the composite score, β = -.136, p = .0.013. Similarly, in the dOV-4 test, the models fit well (*p's*<0.001); the control variables were significant; and d-OV4 was negatively correlated with the composite score, β = -.183, p = .0.001.

Finally, a remaining issue is how delayed recall output order measures fared in comparison to (total) delayed recall performance in the prediction of Experiments 1 and 2's outcomes. In Experiment 1, dFIR and dOV-4 were both highly correlated with the delayed recall score (p's < .001) and so we conducted a separate ordinal regression with delayed recall in place of either output order measure for comparison. Unsurprisingly, delayed recall performance predicted the MMSE (p<.001, Wald t = 27.552, coefficient estimate = .167) and DSST (p<.001, Wald t = 30.662, coefficient estimate = .159) scores very well, in fact better numerically than both dFIR and dOV-4. In Experiment 2, dFIR and dOV-4 were not significantly correlated with delayed recall (lowest p value = .675) and thus we re-ran the same regression analysis including delayed recall as a predictor in Model 1. The results showed that dOV-4 remained a significant predictor of hippocampal volume [β=-.174, partial R=-.236, p=.046] while controlling for delayed recall [β=.196, partial R=.241, p=.041]. Although these additional findings do not show output order measures to be better than traditional neuropsychological measures of memory at predicting generalized cognitive ability, which indeed was not the goal of the present study, they do show that output order variability provides an independent contribution to the prediction of hippocampal volume over and above that given by standard delayed recall performance.

To conclude, in the present paper we set out to determine whether output order of recall was linked to generalized cognitive ability in elderly individuals, and demonstrated that measures of recall initiation and output order variability predict the MMSE score, when taken from a test of delayed memory performance. Second, we showed that dOV-4, indexing the ability to use temporal order information in a delayed memory test, but not dFIR, used as a measure of recall initiation, is associated with hippocampal gray matter volume in cognitively intact elderly participants. Collectively, we believe these findings support the suggestion that

delayed primacy performance, which has been shown to predict cognitive decline from a cognitively healthy baseline, may rely upon at least two separate mechanisms, each supported by a distinct neuronal structure. Future research should explore this claim and attempt to clarify which neuronal substrates may be associated with measures of recall initiation.

Funding and Conflict of Interest

These studies were funded in part by NIMH grants (R01 MH-080405 & R01 MH-056994) to NP. The MERI program was funded in part by Rockland County. No conflicts of interest to declare.

Acknowledgments

We would like to acknowledge the help provided by Chelsea Reichert. The results of these studies were presented at the 2015 meeting of the Alzheimer's Association International Conference, Washington, D.C., USA.

References

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm.*Neuroimage*, *38*(1), 95-113.

Brown, G. D., & Lewandowsky, S. (2010). 4 Forgetting in memory models.*Current Issues in Memory*, 49.

Bruno, D., Grothe, M. J., Nierenberg, J., Zetterberg, H., Blennow, K., Teipel, S. J., & Pomara, N. (2015). A study on the specificity of the association between hippocampal volume and delayed primacy performance in cognitively intact elderly individuals. *Neuropsychologia*, *69*, 1-8.

Bruno, D., Reiss, P. T., Petkova, E., Sidtis, J. J., & Pomara, N. (2013). Decreased Recall of Primacy Words Predicts Cognitive Decline. *Archives of clinical neuropsychology*, *28*(2), 95-103.

Buschke, H., & Fuld, P. A. (1974). Evaluating storage, retention, and retrieval in disordered memory and learning. *Neurology*, *24*(11), 1019-1019.

Davachi, L., & DuBrow, S. (2015). How the hippocampus preserves order: the role of prediction and context. *Trends in cognitive sciences*, *19*(2), 92-99.

Egli, S. C., Beck, I. R., Berres, M., Foldi, N. S., Monsch, A. U., & Sollberger, M. (2014). Serial position effects are sensitive predictors of conversion from MCI to Alzheimer's disease dementia. *Alzheimer's & Dementia*.

Frisoni, G. B., Jack, C. R., Bocchetta, M., Bauer, C., Frederiksen, K. S., Liu, Y., ... & Duchesne, S. (2015). The EADC-ADNI Harmonized Protocol for manual hippocampal segmentation on magnetic resonance: Evidence of validity. *Alzheimer's & Dementia*, *11*(2), 111-125.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). “Mini-mental state”. A practical

method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12*, 189–198.

Glanzer, M. (1972). Storage mechanisms in recall. In G. H. Bower (Ed.), *The psychology of learning and motivation*. New York: Academic Press. Pp. 129-153.

Gomar, J. J, Bobes-Bascaran, M. T., Conejero-Goldberg, C., Davies, P., & Goldberg, T. E. (2011). Utility of combinations of biomarkers, cognitive markers, and risk factors to predict conversion from mild cognitive impairment to Alzheimer disease in patients in the Alzheimer's disease neuroimaging initiative. *Archives of General Psychiatry, 68*, 961-969.

Golomb, J. D., Peelle, J. E., Addis, K. M., Kahana, M. J., & Wingfield, A. (2008). Effects of adult aging on utilization of temporal and semantic associations during free and serial recall. *Memory & Cognition*, *36*(5), 947-956.

Howard, M. W., Fotedar, M. S., Datey, A. V., & Hasselmo, M. E. (2005). The temporal context model in spatial navigation and relational learning: toward a common explanation of medial temporal lobe function across domains. *Psychological review*, *112*(1), 75.

Howard, M. W. and Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **25**, 923–941.

Hsieh, L. T., Gruber, M. J., Jenkins, L. J., & Ranganath, C. (2014). Hippocampal activity patterns carry information about objects in temporal context. *Neuron*, *81*(5), 1165-1178.

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, **24**(1), 103–109.

Kahana, M. J. and Howard, M. W. (2005). Spacing and lag effects in free recall of pure lists. *Psychonomic Bulletin & Review*, **12**, 159–164.

Kahana, M. J., Howard, M. W., Zaromb, F., & Wingfield, A. (2002). Age dissociates recency and lag recency effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 530.

Klein, K. A., Addis, K. M., and Kahana, M. J. (2005). A comparative analysis of serial and free recall. *Memory & Cognition*, *33*, 833–839.

Manning, J. R., Polyn, S. M., Baltuch, G., Litt, B., and Kahana, M. J. (2011). Oscillatory patterns in temporal lobe reveal context reinstatement during memory search. *Proceedings of the National Academy of Sciences*, *108*, 12893–12897.

Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology, 64*, 482-488.

Naveh-Benjamin, M. (2000). Adult-age differences in memory performance: Tests of an associative deficit hypothesis. Journal of Experimental Psychology: Learning, Memory, and Cognition, 26, 1170–1187.

Pomara, N., Bruno, D., Sarreal, A., Hernando, R., Nierenberg, J. J., Petkova, E., Sidtis, J. J., Mehta, P. D., Wisniewski, T.M., Pratico, D., Zetterberg, H. & Blennow, K. (2012). Lower CSF amyloid beta peptides and higher F2-isoprostanes in cognitively intact elderly individuals with major depressive disorder.*The American Journal of Psychiatry*, *169*, 523-530.

Raj, A., LoCastro, E., Kuceyeski, A., Tosun, D., Relkin, N., Weiner, M., & Alzheimer's Disease Neuroimaging Initiative (ADNI. (2015). Network diffusion model of progression predicts longitudinal patterns of atrophy and metabolism in Alzheimer's disease. *Cell reports*, *10*(3), 359-369.

Reichert, Sidtis & Pomara (*In Press*). The Memory Education and Research Initiative: A Model

for Community Based Clinical Research. In D. Bruno (Ed.), *The Preservation of Memory: Theory and Practice for Clinical and Non-Clinical Populations*. Psychology Press: Hove, UK.

Sederberg, P. B., Howard, M. W., and Kahana, M. J. (2008). A context-based theory of recency and contiguity in free recall. *Psychological Review*, *115*(4), 893–912.

Sederberg, P. B., Miller, J. F., Howard, M. W., & Kahana, M. J. (2010). The temporal contiguity effect predicts episodic memory performance. *Memory & Cognition*, *38*(6), 689-699.

Talmi, D., Caplan, J. B., Richards, B., & Moscovitch, M. (2015). Long-Term Recency in Anterograde Amnesia. *PloS one*, *10*(6), e0124084.

Table 1. *Study demographics for 452 participants included in Experiment 1. Age in years (mean and standard deviation, and range); MMSE score (with standard deviation and range); Years of Education (with standard deviation and range); AVLT total recall (with standard deviation and range); immediate recall (with standard deviation and range); and delayed recall (with standard deviation and range).*

	Means	SDs and ranges
Age	71.15	6.98; 60-91
MMSE	29.13	1.08; 24-30
Education (years)	15.78	2.76; 4-25
Total recall	45.47	9.60; 12-72
Immediate recall	5.84	1.57; 3-13
Delayed recall	8.75	3.16; 2-15

Table 2. *Study demographics by cohort: Number of subjects (i.e., N); Age in years (mean and standard deviation); MMSE score (mean and standard deviation); Gender (proportion of females); Diabetes (proportion of reported cases); Hypertension (proportion of reported cases); Years of Education (mean and standard deviation); Total Recall score (mean and standard deviation); and Delayed Recall score (mean and standard deviation).* T-tests *and* χ^2 *tests were used to test for significant differences across groups;* p *values are reported on the far right column.*

	Cohort 1	Cohort 2	*p* value
N	52	27	
Age	67.71 (5.98)	64.63 (3.74)	0.006
MMSE	29.67 (0.51)	29.41 (0.80)	0.125
Gender (females)	31 (60%)	17 (63%)	0.773
Diabetes (yes)	5 (10%)	1 (4%)	0.347
Hypertension (yes)	16 (31%)	4 (15%)	0.122
Education (years)	16.40 (2.54)	15.85 (2.41)	0.354
Total recall	66.50 (13.69)	55.19 (15.84)	0.001
Delayed recall	9.02 (2.86)	7.48 (3.03)	0.029

Figure 1a. Histogram of iFIR frequency: number of individuals initiating recall from each position (1-15) of the study list.

Figure 1b. Histogram of dFIR frequency: number of individuals initiating recall from each position (1-15) of the study list.

Figure 2. Representative coronal sections illustrating the hippocampus ROI along its anterior (top) to posterior (bottom) extension.

Figure 3. Scatterplot of unstandardized residuals of output order measures (dFIR and dOV-4; X-axis) and gray matter hippocampal volumes (Y-axis).

Figure 1a

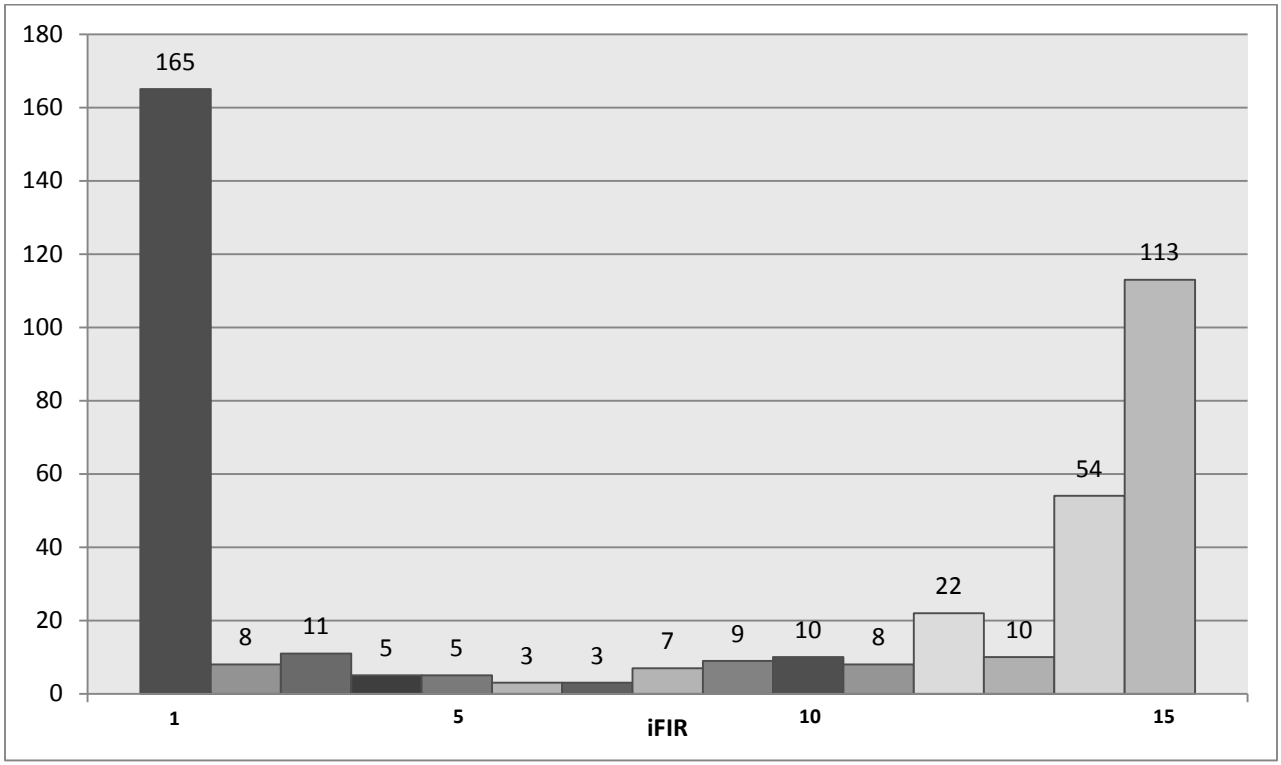

180
160
140
120
100
80
60
40
20
0
165
8
11
5
5
3
3
7
9
10
8
22
10
54
113
1
5
10
15
iFIR

Figure 1b

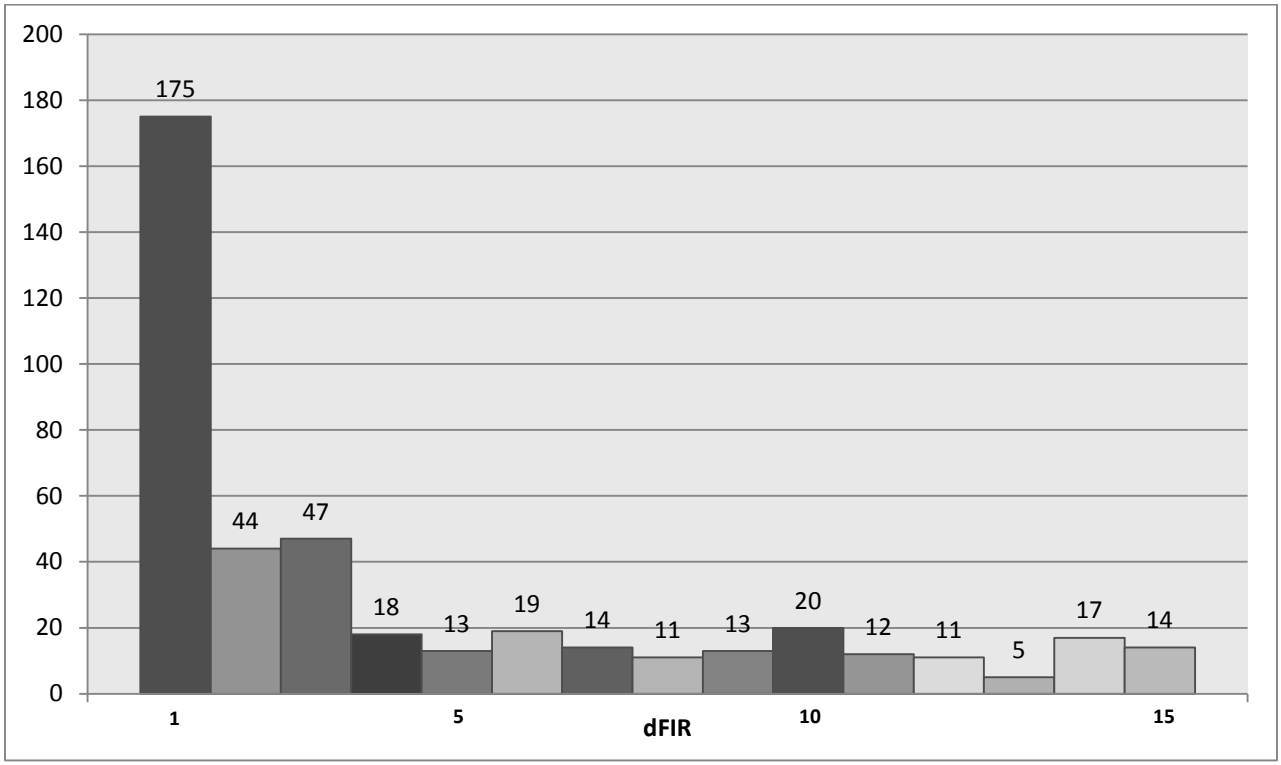

200
180
160
140
120
100
80
60
40
20
0
175
44
47
18
13
19
14
11
13
20
12
11
5
17
14
1
5
dFIR
10
15

Figure 2

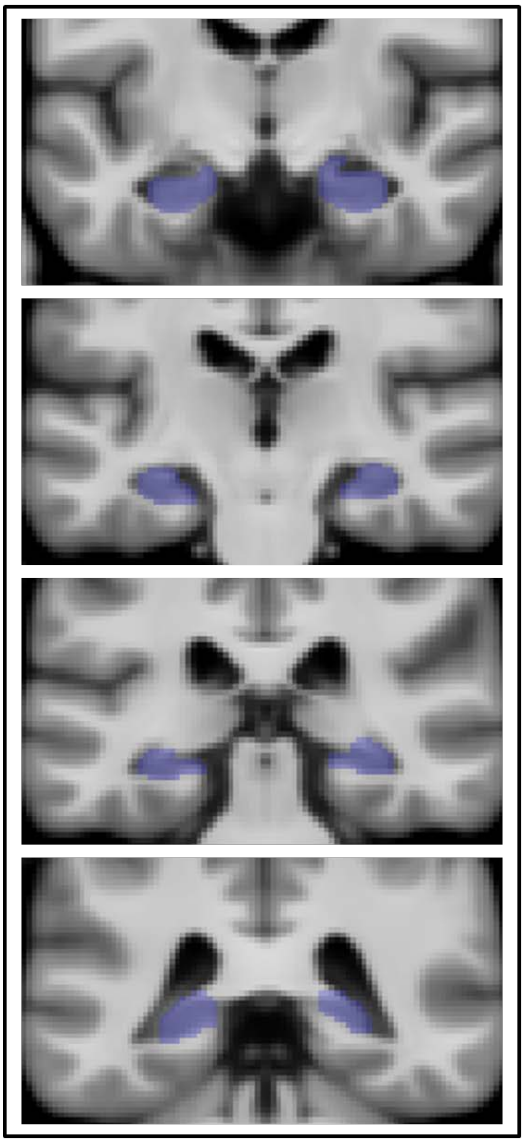
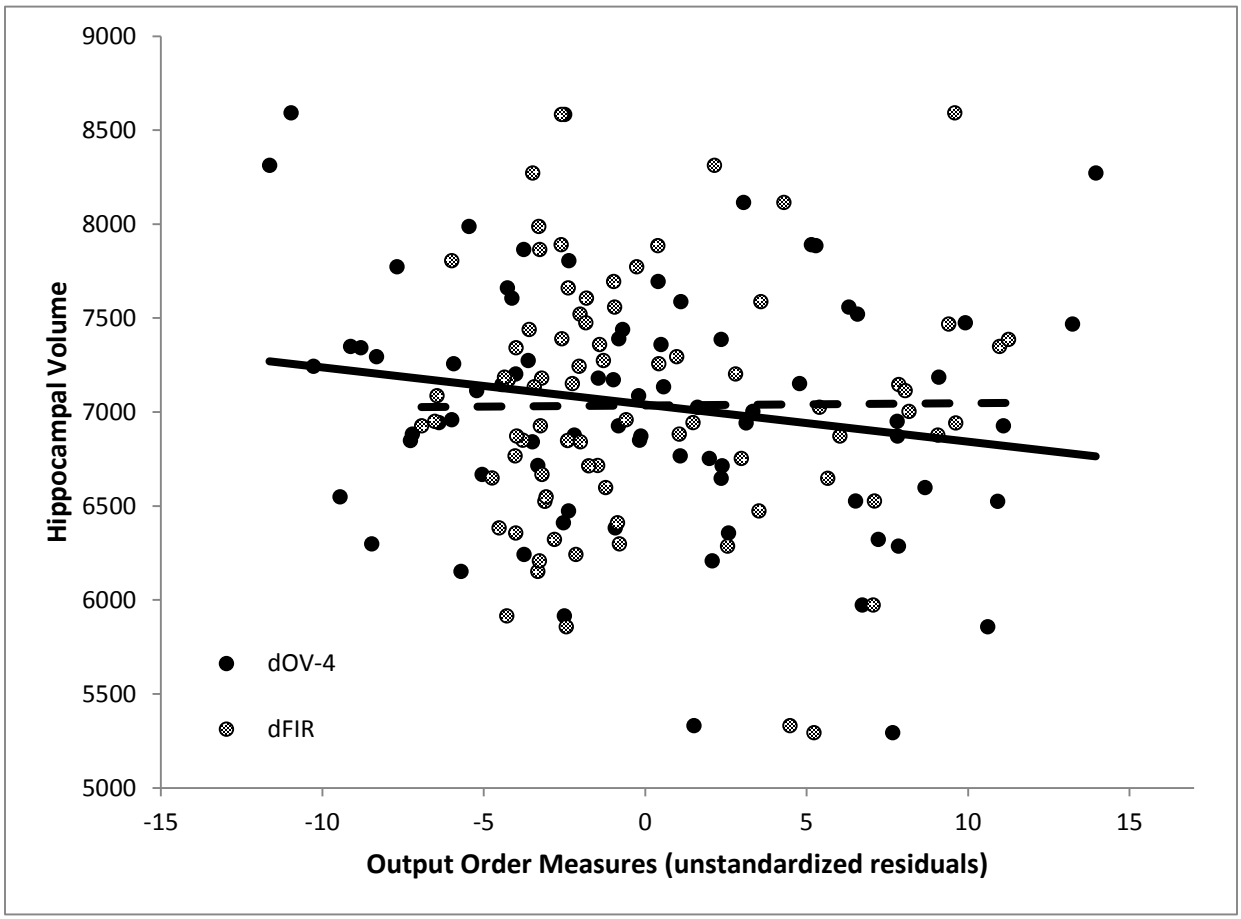


Figure 3


9000
8500
8000
7500
7000
6500
6000
5500
5000
Hippocampal Volume
dOV-4
dFIR
-15
-10
-5
0
5
10
15
Output Order Measures (unstandardized residuals)