The role of working memory in the probabilistic inference of future sensory events

Nathan Casdhollar1, Philipp Ruhnau2, Nathan Weisz2, & Uri Hasson1
1Center for Mind/Brain Sciences (CIMeC), University of Trento Trento, Italy
2Division of Physiological Psychology and Centre for Cognitive Neuroscience, University of Salzburg, Austria
Correspondence to: nate.cashdollar@gmail.com

Abstract—The ability to represent the emerging regularity of sensory information from the external environment has been thought to allow one to probabilistically infer future sensory occurrences, and thus optimize behavior. However the underlying neural implementation of this process is still not comprehensively understood. Through a convergence of behavioral and neurophysiological evidence, we establish that the probabilistic inference of future events is critically linked to people's ability to maintain the recent past in working memory. Magnetoencephalography recordings demonstrated that when visual stimuli occurring over an extended time series had a greater statistical regularity, individuals with higher working memory capacity displayed enhanced slow-wave neural oscillations in the theta band prior to, but not during stimulus appearance. This pre-stimulus neural activity was specifically linked to contexts where information could be anticipated and influenced the preferential sensory processing for this visual information after its appearance. A separate behavioral study demonstrated that this process intrinsically emerges during continuous perception and underpins a realistic advantage for efficient behavioral responses. In this way, working memory capacity optimizes the anticipation of higher-level semantic concepts that are expected to occur in the near future.

Keywords—Magnetoencephalography, Pre-stimulus, Theta-oscillations, Prediction, Statistical learning

I. INTRODUCTION

The human brain is particularly sensitive to the ongoing complexity of sensory information from the external environment (Pouget et al. 2013; Karuza et al. 2014). It has been thought that the ability to represent the statistical relationships of this information over time allows one to anticipate or predict future sensory inputs (Harrison et al. 2011; Bornstein and Daw 2012; Pouget et al. 2013), a process known as probabilistic inference. Previous behavioral research has established that individuals are exceptionally sensitive to stochastic regularities (Lewicki et al. 1992; Smithson 1997; Stephen and Dixon 2011; Emberson et al. 2015) that by their very nature reduce the uncertainty about future events, even if they occur non-deterministically within the environment.

Recent neurobiological research has also identified neural systems whose activity tracks the stochastic regularity of incoming sensory information, and that may implement the functionality of this process (Turk-Browne et al. 2009; Tobia, Iacovella, Davis, et al. 2012; Tobia, Iacovella and Hasson 2012; Karuza et al. 2014; Nastase et al. 2014; Emberson et al. 2015).

While the aforementioned work shows that there exists a functional capacity to recognize and capitalize on such forms of regularity, the underlying neural implementation of this predictive process as naturally induced by the statistical properties of the input is still largely unknown. Thus far, the canonical finding is that stimuli that are less expected to occur within a sequence of events evoke stronger neural responses after stimulus presentation e.g., within-sequence ‘prediction error’ to deviants (Huettel et al. 2005; Strange et al. 2005; Bubic et al. 2009; Vossel et al. 2009; Bubic et al. 2011; Karuza et al. 2014). This is thought to be based on the presence of patterns in the past triggering a match-mismatch evaluation after target presentation (Kumaran and Maguire 2009). Furthermore, statistical regularities between recently encountered stimuli can be maintained either via a simple consolidation-related process that chunks together frequently co-occurring stimulus categories (Perruchet and Pacton 2006) or via statistical learning mechanisms (Bornstein and Daw 2012) in order to create a ‘statistical model’ of previous information.

Our working assumption in the current study is that the regularities reflected within this statistical model can promote anticipatory or predictive neural activity that then further effect how stimuli are processed once they occur. It has been previously demonstrated that experiencing a deterministic regularity of ongoing sensory information can spontaneously promote a heightened processing of this predicted stimuli (Turk-Browne et al. 2010; Reddy et al. 2015). For instance, one study (Turk-Browne et al. 2010) where participants viewed a continuous stream of repeated images that contained sequential contingencies (i.e., some images were deterministically predictive of the next image), showed that participants...
spontaneously coded these relations, as seen in greater hippocampal activity for predictive images. A similar study using depth electrode recordings from the human temporal lobe also demonstrated anticipatory neural firing prior to stimulus presentation for learned stimulus associations within a deterministic series of images (Reddy et al. 2015). The relation between pre-stimulus neural state and post-stimulus processing has been repeatedly demonstrated in recent MEG, EEG and fMRI work (van Dijk et al. 2008; Mathewson et al. 2009; de Lange et al. 2013; Stokes et al. 2014). It has often been shown that differences in endogenous slow wave neural oscillations prior to stimulus presentation can influence the ability to detect near-threshold visual targets once they appear (van Dijk et al. 2008; Mathewson et al. 2009) and the expectation of a given stimulus category (via a deterministically informative cue) has been shown to differentially engage such pre-stimulus activity compared to when the category type could not be anticipated (Bollinger et al. 2010). Similarly, cues that are informative of a behaviorally-relevant target have also been shown to invoke specific patterns of post-cue activity (de Lange et al. 2013; Stokes et al. 2014). Because regular contexts are very likely to be associated with unique patterns of pre-stimulus activity as compared to random ones, comparing anticipatory activity in random series and in stochastically regular series provides an optimal opportunity for identifying the predictive mechanisms potentiated by the underlying maintenance of statistical models. Furthermore, if this anticipatory activity is related to the probabilistic inference about future events, then there should be a strong relation between a person’s degree of pre-stimulus neural activity and the extent to which they show differences in post-stimulus activity for expected vs. surprising stimuli.

A major goal of our current work was also to address the fundamental issue of whether probabilities are sufficient for triggering anticipatory activity or whether this depends on one’s working-memory capacity (WMC) that serves to transiently maintain this information over brief periods of time. Recent behavioral work has identified the crucial explanatory role of inter-individual differences in the ability to detect regularities within the external environment (Misyak et al. 2010; Frost et al. 2015; Siegelman and Frost 2015), however, neurobiological work has yet to examine this potential role. While initial investigations of the brain’s sensitivity to statistical structure of sensory inputs (e.g., Strange et al. 2005; Harrison et al. 2006; Bestmann et al. 2008; Mars et al. 2008) assumed that observers equally weigh all previous events (an ‘ideal Bayesian observer’ model’, Kersten and Yuille 2003; Geisler 2011), it is well established that both the ability to retain observations from the recent past (Luck and Vogel 2013; Ma et al. 2014), as well as the fidelity of these representations (Ma et al. 2014), are a capacity limited process subjected to decay. Indeed, some work (e.g., Harrison et al. 2011; Bornstein and Daw 2012) has shown that neural and behavioral responses to inputs are better modeled by formalisms in which recent observations are more strongly weighted than distant ones, e.g., by implementing a decay function as a proxy for forgetting (Harrison et al. 2011). Given the potential capacity limit for representing recent events we hypothesized that individuals with a higher WMC, allowing the maintenance of more information from the recent past, will also show stronger anticipatory signatures in regular contexts than in random ones.

To investigate whether anticipatory activity reflects an interaction between non-deterministic statistical regularities in the environment and an individuals’ WMC, we used Magnetoencephalography (MEG) to study neural activity when individuals observed series of unique visual images. These series were identical in all aspects apart from the extent to which their statistical structure (as quantified by a 1st order Markov process) licensed predictions about future events. Using this approach, we establish that i) statistically regular contexts evoke regularity-related pre-stimulus anticipatory signatures in the theta and alpha band that are not found for non-predictable contexts; ii) that these signatures are linked to differences in post-stimulus processing of predictable versus unpredictable visual stimuli; and iii) that both pre- and post-stimulus regularity-related processing was also strongly correlated with an individual’s WMC thus suggesting a strong relation between WMC and predictive capacity.

II. METHODS

A. Participants

Twenty participants took part in the MEG study. Data from two participants were excluded due to technical errors during data collection resulting in 18 participants with usable data (range = 19-35 y.o.a, M = 26.35, SD = 3.94; Female 13). All participants reported normal to corrected vision and were not using any medications known to affect cognitive functioning. Participants provided written informed consent and the study was approved by the University of Trento’s Ethical Review Board for human-based research. All participants were compensated at a rate of 12 euros per hour for their time (2 hours).

B. Change detection task (visual working memory capacity)

Prior to both Experiment 1 (MEG) and Experiment 2 (behavioural study), participants first performed a change detection task (Vogel et al. 2005; Fukuda and Vogel 2009) used to assess an individuals’ visual working memory capacity (Cowan 2001). The test-retest reliability of this task to assess working memory capacity is well documented (Kyllingsbaek and Bundesen 2009; Luck and Vogel 2013; Ma et al. 2014) and we confirmed high split-half reliability of this measure within our own collected data in both experiments (see Supplemental Results). This task consisted of stimulus arrays of either 2, 4, 6 or 8 colored squares presented briefly (250ms) on the computer screen. Participants were instructed to remember the stimulus array over a retention interval of 1000ms while fixating on a grey cross-hair. After which, a single colored square was presented in the same location as one of the previous items from the stimulus array and participants indicated whether the color of the square was the same or different as the original item in that location via a button press. On half of the trials, the color of the square was the same as the original item in that location and the other half of the trials presented a colored square that did not match the original item. Individual accuracy on each array size was transformed into a \( K \) estimate using a standardized formula (Cowan 2001) considered to be a robust measurement of individuals’ visual working memory capacity. In this formula, \( K = S(H-F) \), \( K \) is the memory capacity, \( S \) is the size of stimulus array, \( H \) is the observed hit rate, and \( F \) is the false alarm rate (e.g., Cowan 2001; Awh et al. 2007; Fukuda and Vogel 2009; Ma et al. 2014). Using this formula, we...
calculated the mean $K$ for each array size ($S$) for each individual, and these were averaged resulting in a single visual working memory capacity measure for each participant (following Cowan, 2001). Working memory capacity was the only covariate obtained for the participants in this study, given our a-priori hypothesis on the relation between WMC and statistically invoked predictive mechanisms.

C. Stimuli

Stimuli consisted of 2,825 unique gray-scale photographs from four distinct visual categories (animals, human faces, houses, and tools). All photographs were normalized to a mean grey value of 127 and a SD of 75, set at 300x300 pixels, matched for luminance, contrast and spatial variance using the SHINE package (Willenbockel et al. 2010) and presented upon a grey background (127 value). Furthermore, a subset of 150 composite images were created by randomly selecting one picture from each visual category and averaging them so that all aforementioned stimulus features were preserved, yet any distinct visual characteristics of the individual category features were indiscernible. In this way, the composite images maintained the essential low-level features of the discernible images, but did not contain any perceptual/semantic information. We presented these composite images during the inter-stimulus interval (rather than a simple mono-chrome screen or cross-hair fixation) to minimize any effects of major changes in contrast, luminance, and spatial frequency between displays.

D. Design and manipulation of statistical structure

The regular and ‘random’ series were generated via two separate Markov processes that differed in their level of statistical uncertainty (i.e. entropy; see Fig. 1). Markov entropy (ME) can be used to quantify the regularity of an input stream by calculating the mean level of transition constraints between each item within a series of information (i.e. given the current input ‘1’ what is the statistical probability that the following input would be ‘2’ based on all previous occurrences of the input ‘1’), and is derived by the following formula:

$$\text{Markov H} = \sum_i p(i) \sum_{j|i} p(j|i) \log P(j|i)$$

in which, $p(i)$ is the probability of an event, $i$, and, $p(j|i)$, is the probability of event, $j$, given the current event, $i$, and, $n$, indicates the number of different event types. For instance, the transition constraints of a continuous series of two repeating numbers that are ordered deterministically (i.e. ‘1,2,1,2,1,2...’) will have a first-order ME of 0, while a completely random series of two numbers will have a first-order ME of 1.0 (stationary distribution of 1 bit). Likewise, given a completely random series of stimuli consisting of four items, the stationary distribution’s ME would equal 2.0.

Using this approach, we constructed 2 types of transition matrices specifying the transition constraints between the four categories described above. These transition profiles allowed for two levels of ME between the four categories based on the transition matrices demonstrated in Fig. 1A,B. In the high-entropy (“Random”) condition, there were no transition constraints except that the same stimulus category could not be subsequently repeated (i.e. given the current input ‘1’ all subsequent possibilities are equally as probable (33.3%) and therefore the likelihood of one category occurring after the previous one was low) and so the ME of the stationary distribution was 1.57 bit. The low-entropy (more ordered)

Figure 1. Manipulation of statistical structure, task design and procedure. Markov entropy (ME) is a measure that quantifies the regularity of a continuous sequential input. This task implemented 2 types of transition matrices where the relational constraints between 4 items within a series followed two distinct levels of ME: (A) An example of an Ordered series (low-entropy) where given ‘1’ there is a 75% probability that the following input would be ‘2’ and a 25% probability that the following input would be ‘4’. (B) High-entropy (“Random”) series where no transition constraints exist apart from absence of repetitions. After these ‘4 Category’ series were created via a string of numbers (1,2,3,4), each number was assigned a stimulus category (i.e. Animals = 1, Human faces = 2, Houses = 3, and Tools = 4). These assignments were changed for each series to ensure that statistical associations would be re-learned in each series (see 2nd matrices with ‘A’ ‘F’ ‘H’ ‘T’ assignments). (C) A short example of an Ordered series (48 stimuli per series in total). Participants responded with a button press if an image appeared upside down (catch trials). The same picture was never presented twice during the experiment therefore only allowing for the possibility of learning the abstract semantically-related statistical associations between categories to occur.
condition was one where each category could transition to only two out of the four other categories in the series with a 75% probability for one transition and 25% for the other. Consequently, the ME of this matrix was 0.81 (Fig. 1A). To illustrate, in the low-entropy series, given the current input ‘1’ there was a 75% probability that the following input would be ‘2’ (predictable) and a 25% probability that the following input would be ‘4’ (surprising). This low-entropy series contains statistical regularities, (i.e., transition constraints), and therefore offers a basis for learning these associations of occurrence between visual categories, whereas the high-entropy series does not contain any statistical regularities and thus does not allow for any associative learning. We note that even the more regular series are essentially non-deterministic in that at no point is there certainty regarding the next category that could appear (that is, no cell in the transition matrix is 100% diagnostic about the next category that would appear). In addition, we note that all constraints were between adjacent stimuli, and we did not manipulate the strength of non-adjacent constraints.

From these two transition matrices (high- and low-entropy), we generated series of 48 items in length. For each individual series, one of four visual categories was assigned to a number within the generated series (i.e. Animals = 1, Human faces = 2, Houses = 3, and Tools = 4) resulting in visual series with two distinct categorical levels of ME (see example of low ME series Fig. 1C). We refer to these series assignments with low ME as the Ordered condition and to series with high ME as the Random condition. Importantly, our design assured that in both these conditions, the marginal probabilities of each category were identical in all cases and set to 25%. Thus, only transition probabilities differed Ordered and Random series.

For purposes of the current design, it was important that these 4-category transition constraints needed to be re-learned within each series in the case of the Ordered condition so as to rule out any longer-term transfer of statistical learning from one series to the next. For this reason, the assignment of categories to each series of generated numeric label (1, 2, 3, 4) was permuted across series until every possible number/category combination within these constraints was achieved for each entropy condition (16 series per entropy condition). We did this in order to make sure participants would be continually engaged in a learning processes within the regular series. Had we maintained the exact same transition mappings for the all the regular series in the experiment, participants could then rely on a simple recognition strategy of a single pattern across the entire experiment (this could be reduced to a relatively simple associative recognition process. Davachi and Wagner 2002; Bergmann et al. 2012) instead of learning the new sequential relationship between categories within each series.

E. Experimental procedure (Experiment 1; MEG study)

The MEG experiment consisted of 10 recording blocks. Each block consisted of two Random and two Ordered series conditions (48 stimuli per series) and the presentation of conditions was randomized within each testing block (16 series per condition). Each series was presented as follows (see Fig. 1C): Prior to the start of each series, participants pressed a button to indicate they were prepared to view the series of pictures. After which, a black cross-hair was presented centrally on a grey screen for 3000ms, followed by a red cross-hair (1000ms) indicating the start of a series. Participants then continuously viewed 48 novel pictures (stimulus presentation of 500ms and an inter-stimulus-interval [ISI] = 1000ms) per series. No picture was presented twice during the experiment. This therefore only allowed for construction and evaluation of relatively abstract semantically-related statistical associations between categories and ruled out any lower-level associative memory effects that could hold between specific stimuli. During the ISI a single composite image was presented for 1000ms throughout an entire series and was changed prior to each series presentation by random assignment. - see Fig. 1C.

To ensure alertness during the study and reduce motor-related MEG artifacts, we implemented catch-trials on 5% of trials where participants were instructed to press a button when an image was presented upside-down (Fig. 1C). MEG recordings during catch-trials (and false alarms) were removed prior to analysis.

At the conclusion of each series, a new composite image was presented centrally on the screen and participants were then instructed, “Indicate if you find this image ‘Pleasant’ (right button) or ‘Un-Pleasant’ (left button)”. We included this judgment to briefly engage participants in an alternate task involving a subjective discrimination in order to disrupt any short-term retention of prior associative learning from one series to the next (in lieu of simple arithmetic or alphabetizing tasks often used for this purpose, which have an inherent ordering component involved in the completion of these tasks).

F. Trial selection (Experiments 1 and 2)

Trials within a series were identified for analysis based on their transition probability status (Fig. 1A,B). The Random series contained only random trials, with a transition probability of 33%. The Ordered series contained both ‘surprising’ trials, which were those with a 25% transition probability, and associatively ‘predictable’ trials, which were those with a 75% transition probability (Ord_25, Ord_75 henceforth; see Fig. 2A). Prior to analysis in both experiments, we excluded the first 8 trials from each series of 48 trials from analysis in order to avoid any sort of initial processes related to beginning a new series (and potential unlearning of the prior one) that could potentially add noise to our experimental data. We isolated all Ord_25 trials, and to equate the number of Ord_75 trials (which were by definition more frequent than the surprising ones), we selected only predictable trials within an Ordered series which were preceded by 4 prior Ord_75 trials. Therefore by the ≥5th presentation of an Ord_75 stimulus, participants had previously viewed at least one full cycle of the Markov transition matrix with no interfering surprising trial. This was done in order to minimize the recent occurrence of prediction errors in the analysis of Ord_75 trials, recovery from which could introduce noise into the trials we were interested in, thereby focusing our analyses on trials where participants would be in a state of an unimpeded process (i.e. ‘local streak’) of successful prediction. Furthermore, to pseudo-match for the number of trials extracted from the Random series to those extracted for Ord_25 and Ord_75 trials, we only extracted every 4th trial from the Random series (Rand trials; Fig. 2B). It was important to equate the number of trials extracted for each condition, as selecting unequal numbers could bias the signal-to-noise of specific conditions. For instance, had we selected all possible trials, there would be a much larger number of Rand and Ord_75 trials extracted per participant compared to
Ord_25 trials. All else being equal, this would make the contrast between the former two conditions more sensitive than contrasts against the Ord_25 condition. The issue of unequal-trials is of a particular concern for correlational analyses where a behavioural measure is correlated against the mean condition estimate calculated per participant – here, including different numbers of trials per condition could bias the precision across conditions and confound interpretation of the correlation values.

For both experiments, we report all of the results in this manner (Ord_25, Ord_75, and Rand), where Ord_25 and Ord_75 reflect associatively ‘surprising’ or associatively ‘predictable’ trials within the Ordered series respectively (Fig. 2A,B). Using this approach of pseudo-matching our trial selection, Rand trials formed 25% of all the trials from the Random series, Ord_25 trials formed 25% of all the trials from the Ordered series, and Ord_75 trials formed 24% of all the trials from the Ordered series (by trial selection of the ≥5th presentation this probability is 0.755 = 24%). Note that there were no statistical differences in the mean number of trials included in the MEG analysis after artifact rejection (all Ps > 0.11) and also, the mean number of trials for each participant did not significantly correlate individual K-scores (all Ps > 0.10), demonstrating that the number of trials included did not influence the MEG results.

G. Magnetoencephalography recording and pre-processing

MEG data were recorded in an electromagnetically shielded room (Vacuumschmelze, Hanau, Germany) using a 306-channel MEG (Vectorview, Elekta-Neuromag Oy, Helsinki, Finland) comprising 204 orthogonal planar gradiometers and 102 magnetometers combined in 102 locations above the participant’s head. Prior to the MEG recording session, cardinal points at the nasion and left and right preauricular points were digitized using a Polhemus FASTRAK 3D digitizer. During recording blocks, the position of the participant’s head was quasi-continuously measured using five head position indicator coils. The MEG acquisition threshold for head movements was less than 2 mm between recording blocks. Data were recorded at a 1000 Hz sampling rate and 0.01 Hz high pass filtering.

All MEG data were preprocessed and statistically analyzed using the Fieldtrip toolbox (Oostenveld et al. 2011). A discrete Fourier transform filter was applied to remove line noise (default values of 50, 100, 150 Hz) and data was epoched from -1.5 to 1.5s relative to stimulus onset and down-sampled to 250 Hz. All data were visually inspected to remove noisy trials and channels prior to an Independent Component Analysis (ICA; Bell and Sejnowski 1995). Components capturing ocular and cardiac artifacts were removed and the raw data reconstructed. After ICA, missing channels were interpolated using a nearest-neighbor approach.

G.1 Event Related Field pre-processing.

Epochs were band-pass filtered between 1-35 Hz and then averaged from -200 to 600ms relative to stimulus onset per trial. The 200ms prior to stimulus presentation was used as a pre-stimulus epoch for baseline correction. Statistical comparisons between trials types were conducted separately for Magnetometers (102 sensors) and Gradimeters (102 combined sensors).

G.2 Pre-stimulus time-frequency pre-processing.

For spectral analysis, epochs were high pass filtered (1 Hz) and no baseline normalization was applied because we had an a priori hypothesis concerning ongoing pre-stimulus differences between conditions, which were the focus of this current investigation. Trials were selected with a restriction allowing only for trials where individuals had experienced a succession of 4 standard (Ord_75) stimuli in Ordered series and a matched number of Rand trials from the Random series. The time–frequency distributions of pre-stimulus activity types were compared separately for the magnetometer and gradiometer sensors. Condition-related differences in oscillatory power were estimated using a multitaper FFT time–frequency transformation with frequency-dependent Hanning tapers (time window: $\Delta t = 4/f$ sliding in 50 ms steps). We calculated power from 2 to 30 Hz in steps of 2 Hz, separately for each series type (Ord_75 and Rand). The type-I error rate for the complete set of sensors (analysis was conducted separately for Magnetometers [102 sensors] and Gradimeters [102 combined sensors]) was controlled for multiple comparisons using cluster-extent family-wise error control (Maris and Oostenveld 2007; $p < .05$ on the cluster level) implemented in the Fieldtrip software. While we were primarily interested in the 1sec pre-stimulus ISI time-window, we also included the 0.5sec window in which the stimulus was on the screen in order to determine if any differences in oscillatory activity were specific to the timing of stimulus onset and off-set or were due to more continuous tonic differences between trial types that was unrelated to stimulus timing.

H. MEG Event Related Fields analysis

To investigate differences in the event related neural response to trial types within Ordered series (Ord_75, Ord_25 trials), we calculated ERF differences between trial-types using an a priori time window of interest from 200 – 400ms after stimulus onset based on well-established event-related components common for novelty and ‘odd-ball’ detection (Polich and Comerchero 2003; Gonsalves et al. 2005; Cycowicz and Friedman 2007; for review see, Polich 2007). We then conducted paired-sample t-tests ($n=18$, two-tailed, thresholded at $P < 0.05$, FDR corrected for multiple comparisons) within this averaged time-window of interest to evaluate differences between trial types across magnetometer and gradiometer sensors separately. Pair-wise differences were only considered significant for clusters of four or more neighboring sensors (same threshold as time-frequency analysis below).

I. MEG Time-Frequency analysis

1.1 Time-Frequency differences between Order and Random trials

The power differences (2-30 Hz) during ISIs in the Ordered and Random series were compared using a cluster-based nonparametric, permutation-based statistic that controls for type-I errors with respect to multiple comparisons (Nichols and Holmes 2002; Maris and Oostenveld 2007). First, Student’s-T statistics for the Ord_75 vs. Rand contrast were calculated. The cluster-finding algorithm identified clusters of neighboring sensors (minimum cluster of ≥4 neighboring sensors) and frequency bins where the t statistics for the contrast exceeded a significance level of $p < 0.05$. The cluster-level test statistic was a cluster-mass measure defined as the sum of the t statistics of the sensors in a cluster. In a nonparametric statistical test, cluster-level test statistic was determined based
on construction of a null distribution. The null distribution was obtained by randomly permuting the data between the two trial types within every participant. By creating a reference distribution from 500 random sets of permutations, the cluster-level p value was estimated as the 95% percentile of the randomization null distribution. In summary, this permutation procedure identifies clusters in the data where the contrasts on the single sensor exceeds the p < .05 level, and the number of adjacent sensors showing this effects exceeds that likely to be found by chance.

1.2 Pre-stimulus Time-Frequency correlation with WMC

The mean power difference between both trial types (Ord_75 – Rand) was calculated for each participant and averaged over the entire 1.5sec time-window, in the same manner as the group analysis above (ISI: -1.0 to 0 sec and Stimulus presentation: 0 to 0.5 sec). Note this covers the entire total time period of a series, split trial-by-trial, and allows investigating the impact of the experimental design both on the pre-stimulus period as well as for the period during which the stimulus was present on the screen. We then conducted a sensor-wise Pearson correlation test to assess if individual differences in K-score correlated with (Ord_75 – Rand) power differences during the ISI, for each of the frequencies and time bins of interest. Using this approach, we imposed no a priori assumptions of sensor location or specific frequency band between 2-30 Hz. Control for family wise error was implemented via cluster-level correction as described above.

1.3 Pre-stimulus Time-Frequency correlation with post-stimulus ERF

For each participant (n=18) we identified the sensor showing the largest power difference (within the theta and alpha bands separately) between the Ord 75 and Rand trial types, averaged only within the pre-stimulus time-window (ISI period: -1.0 to 0 seconds prior to stimulus onset). Next we determined if these individual-level pre-stimulus power differences correlated with individual ERF differences between conditions (Ord_75, Ord_25, and Rand) within two a priori time windows of interest based on well-established event-related components (Luck et al. 2000; Polich 2007). In order to isolate early attentional processing we selected one time-window from 0 – 200ms after stimulus presentation (Heinze et al. 1994; Valdes-Sosa et al. 1998; Luck et al. 2000). We also examined a later time-window from 200 – 400ms after stimulus onset, commonly found for novelty and ‘odd-ball’ detection ERF components (Polich and Comerchero 2003; Gonsalves et al. 2005; Cycowicz and Friedman 2007). We then conducted a sensor-wise permutation-based (described above) Pearson correlation test to evaluate these relations (based on our previous ERF and Time-frequency results this analysis was only conducted on magnetometer sensors).

It is important to note, that we imposed no a priori constraints on sensor location for either the maximal pre-stimulus power difference calculation or for significant clusters reflecting a correlation with post-stimulus onset ERF difference between trial types. Therefore, this analysis was independent of any prior one to maintain statistical independence. Significant clusters reflecting a correlation with ERF differences between trial types were identified by groupings of four or more neighboring channels (identical to all prior analyses).

J. Behavioral Experiment 2

Participants. Twenty healthy participants took part in the study (19-31 years, mean= 23.95, SD 3.62; Female 12). This number of participants is within the range of previous investigations using the same change detection task as a covariate for a secondary task in order to assess individual behavioural variance within a population (Luck and Vogel 2013; Ma et al. 2014). All participants reported normal or corrected vision and were not using any medications known to affect cognitive functioning. Participants provided written informed consent and the study was approved by the University of Trento’s Ethical Review Board. All participants were compensated at a rate of 10 Euro per hour for their time.

Design and procedure. In this study, participants underwent a slightly modified variant of the MEG design where they were asked to make a ‘Living / Non-Living’ judgment for each stimulus presented (Supplementary Material: Methods and Figure S1).

III. RESULTS

During MEG recordings, we presented participants with continuous series of unique visual stimuli drawn from four distinct categories (Animals, Houses, Faces, Tools) where the probabilistic likelihood of one category type transitioning to the next was systematically varied (Fig. 1). In the Random condition, there were no transition constraints between categories apart from the fact that a category could not appear twice in a row (Fig. 1B), and the probability of each category (the marginal frequency) was set at 33.3% (Rand trials). In the Ordered condition, the appearance of a particular category could be predicted with a 75% probability (Ord_75 trials) and deviations (Ord_25 trials; Fig. 1A) occurred on 25% of trials.

A. Working memory capacity and behavior during MEG study

As expected, participants’ accuracy during the change detection task (used to assess working memory capacity, see Methods) decreased with the size of the target array to be encoded (Pair-wise t-tests, two-tailed, mean ± S.D., 2-items: 95.8 ± 3.64% vs. 4-items: 81.5 ± 10.9%; t(17) = 7.23, P = 0.0001; 4-items vs. 6-items: 73.5 ± 10.3%; t(17) = 3.77, P = 0.002; and 6-items vs. 8-items: 67.2 ± 9.77%; t(17) = 2.79, P = 0.013). When calculating participants’ individual visual working memory capacity (K-score), this resulted in a mean K of 2.70 and median K of 2.35 (± 1.02), which is similar to previous reports using this measure (see review in, Luck and Vogel 2013). The split-half reliability of this behavioral test (Spearman-Brown corrected) was r=0.83 (Supplemental Results). During the MEG study, participants were accurate in identifying catch trials (80.6 ± 8.4%), which is a similar level of performance with previous reports of stimulus detection for inverted pictures (Scapinello and Yamey 1970; Diamond and Carey 1986). Furthermore, there were no differences in participant’s performance between conditions (Pair-wise t-tests, two-tailed, mean ± S.D., Order: 80.4 ± 11.1% vs. Random: 80.9 ± 9.2%, p = 0.88) demonstrating that comparable levels of attention were maintained for both conditions. Although participants displayed a very low mean false alarm rate of 2.4% (i.e. incorrectly responding to a trial when the stimulus within a series that was not upside-down), we also calculated a corrected hit-rate (Hits – False Alarms). Pair-wise t-tests confirmed there was no differences in response bias between
In a cluster of frontal sensors, [Ord_75 – Rand] pre-stimulus power differences in the alpha band (M = -0.034 ± 0.072 fT) were also positively correlated with individual’s WMC estimates (collapsing across sensors in the significant cluster r(18) = 0.61, P = 0.008). Notably for the alpha band, the correlation held throughout both the inter-stimulus-interval and during stimulus presentation (Fig. 3D,E,F) and is consistent with previous investigations of pre-stimulus oscillatory activity (Alpha: Addante et al. 2011; Jutras et al. 2013), we then filtered for results specifically within theta (4-8 Hz) and alpha (8-12 Hz) separately to better isolate and to interpret these processes. To summarize our analysis approach, the first part of the analysis (multi-plot results presented in Fig. S2) is completely independent of any prior analysis, whereas the follow up ‘drill down’ into the alpha andtheta band is based on a follow up descriptive procedure whose purpose is to meaningfully describe the core features of these data patterns (Fig. 3). This includes generating scatter plots which are necessary for understanding the direction of the correlation and the range of values it subsumes (Fig. 3C,F).

In a cluster of central sensors, [Ord_75 – Rand] pre-stimulus power differences in the theta band (M = 0.045 ± 0.12 fT) were positively correlated with individuals’ WMC estimates (collapsing across sensors in the significant cluster r(18) = 0.59, P = 0.009; Fig. 3A,C), This theta band pattern held throughout the pre-stimulus period but terminated just prior to stimulus onset (Fig. 3B) indicating this relationship was temporally synchronized with the stimulus timing of the presented visual series. We compared Time-frequency power signatures during pre-stimulus intervals, for the Ordered and Random series (range 2 – 30 Hz, separately for the magnetometer and gradiometer sensors). An initial analysis that was independent of the working-memory capacity measure revealed no statistically significant differences in pre-stimulus activity in these series (see Materials and Methods). However, power differences during the pre-stimulus intervals showed topologically wide-spread and statistically significant correlations with participants’ working memory capacity, within clusters including primarily frontal and central magnetometer sensors (see Fig. S2). While this result speaks very strongly to the importance of examining WMC in relation to statistical contexts, it contains an enormous amount of data, with significant effects found in different frequency bands and brain regions. Specifically, when examined over the entire pre-stimulus time window, the frequency distribution of this significant cluster encompassed the entire frequency range between 2 – 10 Hz. Because this range includes both alpha- and theta-bands, which are often associated with different cognitive functions, and in accordance with previous investigations of pre-stimulus oscillatory activity (Alpha: van Dijk et al. 2008; Mathewson et al. 2009, Theta: Addante et al. 2011; Jutras et al. 2013), we then filtered for results specifically within theta (4-8 Hz) and alpha (8-12 Hz) separately to better isolate and to interpret these processes. To summarize our analysis approach, the first part of the analysis (multi-plot results presented in Fig. S2) is completely independent of any prior analysis, whereas the follow up ‘drill down’ into the alpha and theta band is based on a follow up descriptive procedure whose purpose is to meaningfully describe the core features of these data patterns (Fig. 3). This includes generating scatter plots which are necessary for understanding the direction of the correlation and the range of values it subsumes (Fig. 3C,F).

In a cluster of central sensors, [Ord_75 – Rand] pre-stimulus power differences in the theta band (M = 0.045 ± 0.12 fT) were positively correlated with individuals’ WMC estimates (collapsing across sensors in the significant cluster r(18) = 0.59, P = 0.009; Fig. 3A,C). This theta band pattern held throughout the pre-stimulus period but terminated just prior to stimulus onset (Fig. 3B) indicating this relationship was temporally synchronized with the stimulus timing of the presented visual series. We computed the orthogonal pairwise contrasts between ERFs in the three trial types (Ord_75, Ord_25, and Rand) during a 200-400ms post stimulus onset window in order to target the commonly reported novelty and ‘odd-ball’ components (Polich and Comerchero 2003; Gonsalves et al. 2005; Cycowicz and Friedman 2000; for review see, Polich 2007; see Fig. 2A,B for trial selection procedure). This analysis revealed a pattern consistent with statistical learning of the transition structure, seen in a greater amplitude for Ord_25 trials (7.35 ± 1.82 fT) relative to Ord_75 trials (-0.57 ± 2.15 fT) (Fig. 2C) on the magnetometer sensors (average over cluster: t(17) = 2.92, P = 0.01). The contrast between the Ord_25 and Rand conditions (Rand = 5.02 ± 2.13 fT) was not statistically significant (t(17) = 0.49, P = 0.63). No differences were found for combined gradiometer sensors.

C. Pre-stimulus Time-Frequency effects in relation to working memory capacity

We compared Time-frequency power signatures during pre-stimulus intervals, for the Ordered and Random series (range 2 – 30 Hz, separately for the magnetometer and gradiometer sensors). An initial analysis that was independent of the working-memory capacity measure revealed no statistically significant differences in pre-stimulus activity in these series (see Materials and Methods). However, power differences during the pre-stimulus intervals showed topologically wide-spread and statistically significant correlations with participants’ working memory capacity, within clusters including primarily frontal and central magnetometer sensors (see Fig. S2). While this result speaks very strongly to the importance of examining WMC in relation to statistical contexts, it contains an enormous amount of data, with significant effects found in different frequency bands and brain regions. Specifically, when examined over the entire pre-stimulus time window, the frequency distribution of this significant cluster encompassed the entire frequency range between 2 – 10 Hz. Because this range includes both alpha- and theta-bands, which are often associated with different cognitive functions, and in accordance with previous investigations of pre-stimulus oscillatory activity (Alpha: van Dijk et al. 2008; Mathewson et al. 2009, Theta: Addante et al. 2011; Jutras et al. 2013), we then filtered for results specifically within theta (4-8 Hz) and alpha (8-12 Hz) separately to better isolate and to interpret these processes. To summarize our analysis approach, the first part of the analysis (multi-plot results presented in Fig. S2) is completely independent of any prior analysis, whereas the follow up ‘drill down’ into the alpha and theta band is based on a follow up descriptive procedure whose purpose is to meaningfully describe the core features of these data patterns (Fig. 3). This includes generating scatter plots which are necessary for understanding the direction of the correlation and the range of values it subsumes (Fig. 3C,F).

In a cluster of central sensors, [Ord_75 – Rand] pre-stimulus power differences in the theta band (M = 0.045 ± 0.12 fT) were positively correlated with individuals’ WMC estimates (collapsing across sensors in the significant cluster r(18) = 0.59, P = 0.009; Fig. 3A,C). This theta band pattern held throughout the pre-stimulus period but terminated just prior to stimulus onset (Fig. 3B) indicating this relationship was temporally synchronized with the stimulus timing of the presented visual series.

In a cluster of frontal sensors, [Ord_75 – Rand] power differences in the alpha band (M = -0.034 ± 0.072 fT) were also positively correlated with individual’s WMC (average of all sensors over-time within the significant cluster r(18) = 0.61, P = 0.008). Notably for the alpha band, the correlation held throughout both the inter-stimulus-interval and during stimulus presentation (Fig. 3D,E,F) and is consistent with previous investigations of pre-stimulus oscillatory activity (Alpha: Addante et al. 2011; Jutras et al. 2013), we then filtered for results specifically within theta (4-8 Hz) and alpha (8-12 Hz) separately to better isolate and to interpret these processes. To summarize our analysis approach, the first part of the analysis (multi-plot results presented in Fig. S2) is completely independent of any prior analysis, whereas the follow up ‘drill down’ into the alpha and theta band is based on a follow up descriptive procedure whose purpose is to meaningfully describe the core features of these data patterns (Fig. 3). This includes generating scatter plots which are necessary for understanding the direction of the correlation and the range of values it subsumes (Fig. 3C,F).

In a cluster of central sensors, [Ord_75 – Rand] pre-stimulus power differences in the theta band (M = 0.045 ± 0.12 fT) were positively correlated with individuals’ WMC estimates (collapsing across sensors in the significant cluster r(18) = 0.59, P = 0.009; Fig. 3A,C). This theta band pattern held throughout the pre-stimulus period but terminated just prior to stimulus onset (Fig. 3B) indicating this relationship was temporally synchronized with the stimulus timing of the presented visual series.
gradiometers. Taken together, these analyses indicate that WMC is associated with anticipatory patterns of neural activity that are reflected in greater pre-stimulus power during Ordered as compared to Random series.

**D. Pre-stimulus Time-Frequency and post-stimulus ERFs**

The analyses above suggest that individuals with higher WMC show stronger pre-stimulus preparatory activity in statistically regular contexts, consistent with recent work suggesting a role for WMC in capitalizing on statistical information, as outlined in the Introduction. However, the question still remains of how this activity relates to neural processing during stimulus appearance. Several prior MEG studies have linked pre-stimulus oscillations to post-stimulus ERFs (van Dijk et al. 2008; Lange et al. 2012; Wutz et al. 2014). We conducted a similar analysis to determine whether individuals who more strongly differentiated Ordered from Random series during the pre-stimulus interval show stronger post-stimulus ERF differences between stimuli drawn from Ordered vs. Random series.

For each participant we identified the sensor showing the maximal power difference (within the theta and alpha bands separately) between the Ord_75 and Rand trial types, during the inter-stimulus interval. Next we assessed if these difference magnitudes correlated with inter-individual differences in ERFs for the different trial types (Ord_75, Ord_25, and Rand). We examined differences in ERFs in two time windows after stimulus onset: 0-200ms, an epoch that has been linked to early attentional processes (Heinze et al. 1994; Valdes-Sosa et al. 1998; Luck et al. 2000) and 200-400ms, an epoch linked to oddball-detection effects (Polich and Comerchero 2003; Gonsalves et al. 2005; Cycowicz and Friedman 2007; Polich 2007). In summary, this sensor-wise analysis identified clusters where ERF differences were explained by differential pre-stimulus activity between Ordered and Random series. (Note that by identifying the sensor with the strongest pre-stimulus effect separately for each participant, we imposed no a-priori assumption on sensor location, thus maintaining independence from the prior analyses.)

We found that pre-stimulus Ord_75 - Rand differences in the theta band positively correlated with the (Ord_75 – Rand) ERF during the 0-200ms window, within a cluster of occipital sensors (** indicates P < 0.01). Specifically, (A) pre-stimulus power differences in the theta band (4-8 Hz) between Ord_75 and Rand trials positively correlated with individuals’ K-scores within a cluster of central sensors (C) and continued throughout the entire pre-stimulus period but terminated with the onset of the stimulus (B). Similarly, within a cluster of frontal sensors (D) pre-stimulus power differences in the alpha band (8-12Hz) between Ord_75 and Rand trials also positively correlated with individual’s K-scores (F), yet this pattern continued throughout stimulus presentation (E).

Figure 3: Pre-stimulus oscillations and working memory capacity. The mean power differences between Ord_75 and Rand trials (ISI: -1.0 to 0 sec and stimulus presentation: 0 to 0.5 sec) positively correlated with participants’ individual working memory capacity (K-score) within clusters including frontal and central magnetometer sensors (** indicates P < 0.01).

Specifically, (A) pre-stimulus power differences in the theta band (4-8 Hz) between Ord_75 and Rand trials positively correlated with individuals’ K-scores within a cluster of central sensors (C) and continued throughout the entire pre-stimulus period but terminated with the onset of the stimulus (B). Similarly, within a cluster of frontal sensors (D) pre-stimulus power differences in the alpha band (8-12Hz) between Ord_75 and Rand trials also positively correlated with individual’s K-scores (F), yet this pattern continued throughout stimulus presentation (E).

We found that pre-stimulus Ord_75 - Rand differences in the theta band positively correlated with the (Ord_75 – Rand) ERF during the 0-200ms window, within a cluster of occipital sensors (** indicates P < 0.01). The average correlation over this sensor cluster was also statistically significant, r(18) = 0.53, P = 0.024. This indicates that enhanced early post-stimulus processing of a predictable stimulus (Ord_75 - Rand) is found for individuals for which Ordered series were associated with greater pre-stimulus theta band power as compared to Random series. For the theta band, we found no other statistically significant correlation between pre-stimulus power and post-stimulus ERF differences in any other comparison, within the early or late time-windows.

We conducted a similar analysis for the alpha band power. Pre-stimulus alpha differences between Ord_75 and Rand trials positively correlated with the (Ord_25 - Ord_75) ERF during the 200-400ms window, within a cluster of parieto-occipital sensors (** indicates P < 0.01). The average correlation in this sensor cluster was statistically significant, r(17) = 0.57, P = 0.017. This indicates that a stronger ‘odd-ball’ ERF (Ord_25 - Ord_75) in this time window was found for individuals for which Ordered series invoked greater pre-stimulus alpha band power as compared to Random series. (In Supplemental Results (section 1.2.2) we report a similar sensor-wise analysis investigating clusters where post stimulus ERF differences could be explained by differential time-frequency activity between Ordered and Random series during the processing of the previous stimulus, which returned a null result, all P > 0.05 after cluster-correction).
ERFs for the three trial types (Ord_75, Ord_25, and Rand) during the -200ms pre-stimulus time window (baseline period) and found no significant differences between any trial types even without cluster-correction (all p’s >0.05). Next, we repeated the same analysis reported above (Fig. 4) of maximal time-frequency power differences (within the theta and alpha bands separately) between the Ord_75 and Rand trial types, during the inter-stimulus interval and found no significant correlations of these difference magnitudes with inter-individual differences in ERFs for the different trial types (Ord_75, Ord_25, and Rand) during the -200ms pre-stimulus time window used as a baseline period (all p’s >0.05). Therefore, all reported ERP waveform differences are not due to baseline shifts during the inter-stimulus interval.

E. Moderation analysis of pre-stimulus and post-stimulus processing with WMC

The relationship presented in the prior section between pre-stimulus power differences for Ordered and Random series, and post-stimulus onset ERFs (‘PreTheta/PostERF’ and ‘PreAlpha/PostERF’; Fig. 4), was established independently from any relationship with WMC. Our final analysis examined whether the relation between (regularity-related) pre-stimulus power differences and post-stimulus ERF patterns is itself moderated by WMC.

To examine this issue we conducted a moderation analysis (implemented via regression models; Baron and Kenny 1986) to determine the potential role of WMC, operationalized via K-scores, in moderating pre-stimulus to post-stimulus relationships (Supplementary Methods). Here we report this analysis for the theta band only, as the data for the alpha band did not satisfy the typical requirements for a moderation analysis (the two non-direct pathways were not associated with a significant correlation or one approaching significance). The moderation analysis showed that the relationship between the PreTheta and PostERF variables was moderated by WMC (Fig. S3). Specifically, after fixing for WMC, the statistically significant correlation between PreTheta and PostERF (τ(18) = 0.53, p = 0.024) was no longer significant. We note however, that we did not identify a significant difference in the magnitude of this direct pathway (a post-hoc Sobel test was non-significant, p>0.05).

F. Behavioral evaluation (Experiment 2)

The MEG findings corroborated our hypothesis that statistically regular series promote a specific anticipatory pre-stimulus processing that scales with working memory capacity, and that these anticipatory processes impact post-stimulus processing. To evaluate whether WMC is related to the behavioral ability to use this statistical information we conducted an additional behavioral study (Experiment 2: N = 20) with the same design of the MEG study, but that required the participant to indicate by button press if each picture presented was a ‘living’ or ‘non-living’ item (Supplementary Methods). This study also included a WMC assessment, exactly as detailed for Experiment 1. The split-half reliability of this behavioral test (Spearman-Brown corrected) was r=0.77 (Supplemental Results).

Pair-wise t-tests demonstrated that participants’ accuracy (correctly classifying pictures as ‘living’ or ‘non-living’) was significantly better on Ord_75 trials (91.64 ± 4.23%) when compared to Rand trials (90.69 ± 4.26%; t(18) = -2.40, P =
0.028), demonstrating a behavioral facilitation of visual target identification when categorical information was associatively predictable. In addition, participants showed significantly better accuracy for Ord_75 trials when compared to trials that were associatively ‘surprising’ within an Ordered series (Ord_25: 90.00 ± 3.92%; t(18) = -3.06, P = 0.007). This pattern of results is important as it indicates that the behavioral facilitation of responses for Ord_75 trials is not a generalized effect of those trials appearing in low-entropy series per se, but instead a differentiation of performance based on the predictability each specific trial type within a series. In accordance with this notion, there was no difference in accuracy between Rand trials and Ord_25 trials (t(18) = 1.17, P = 0.259). Subsequent pair-wise t-test comparisons of participant’s speed of response (RT: reaction times) did not significantly differ between any of the trial types (all P > 0.05), yet there was a trend for slower responses on Ord_75 trials (453.14 ± 13.9ms) when compared to Rand trials (448.17 ± 13.7ms; t(18) = 1.97, P = 0.065) suggesting a beneficial increase of RT might be related to the higher accuracy in Ord_75 trials. To summarize, these patterns strongly suggest that in Ordered series, the participants as a group were sensitive to the statistics of the series and used this information to anticipate the more predictable category.

We were however also interested in how WMC could be related to the benefit provided by statistically regular series. We derived a dependent measure of behavioral efficiency (reaction time/accuracy, known as Inverse Efficiency Score [IES] – see Supplemental Results) for each trial type, where a lower score reflects more efficient behavior (Townsend 1978, 1983). Using this approach, we found that participants with higher WMC more strongly benefited from trial predictability. Consistent with what could be expected from the MEG results, participants’ WMC estimates were negatively correlated with the (Ord_75 – Rand) differences in IES, r(19) = -0.534, P = 0.019. That is, participants with higher WMC showed more efficient response behavior. There was no correlation between WMC and the (Ord_25 – Rand) difference in IES, r(19) = -0.018, P = 0.942. To summarize, this study showed that participants with greater WMC made better use of prior statistical information, resulting in more efficient processing of predictable trials.

IV. DISCUSSION

Current neurobiological models attribute a fundamental role of people's ability to assess the uncertainty of sensory events from the environment as a foundation for perception and cognition (Friston 2009). This statistical knowledge about the relative probabilities of sensory occurrences can be used to optimize sensory processing either by mechanisms engaged after stimulus presentation (Bar et al. 2006) and/or via the construction of anticipatory predictions prior to stimulus appearance. People’s sensitivity to the statistical structure of temporally extended events (Lewicki et al. 1992; Smithson 1997; Stephen and Dixon 2011; Karuza et al. 2014; Emberson et al. 2015) would suggest that such probabilities could be used for constructing anticipatory predictions about forthcoming events, a process known as probabilistic inference.

However to date, the neurobiological mechanisms that allow using statistical information to optimize processing have not been sufficiently delineated. Most importantly, the potential role of working memory ability has been virtually absent in discussions of the neurobiology of statistical learning more generally, and prediction specifically (but see Misjak et al. 2010; Frost et al. 2015; Siegelman and Frost 2015, for examinations in the behavioral literature). One reason may be that initial neurobiological models emphasized the idea that individuals aggregate statistical knowledge over large temporal constants, which was taken to reflect an ‘ideal Bayesian observer’ mechanism (Strange et al. 2005; Harrison et al. 2006). However, subsequent modeling of these experimental data were highly suggestive of a role for working-memory capacity (WMC), showing that probabilistic inference is more accurately potentiated from a very limited number of recent events (Harrison et al. 2011). Here we find that individuals display a specific sensitivity to the non-deterministic associative regularities of occurrence between semantic categories and most importantly, we establish that the probabilistic inference of future events is critically influenced by differences in people’s ability to maintain the recent past in short-term working memory. Specifically, visual WMC was related to differences in slow-wave neural oscillations prior to the stimulus appearing, primarily in the Theta frequency band (Fig. 3A,B,C) and this enhancement of pre-stimulus activity was further linked to preferential early attentional processing after stimulus presentation in statistical regular contexts (Fig. 4A). This therefore suggest that one’s ability to represent statistical relationships within the recent past helps anticipate events in the near future (via preparatory slow-wave neural oscillations) and thus facilitate early processing of future sensory inputs.

A. Signatures of statistical learning

Our results highlight how the statistical structure of sensory information from the environment is related to anticipatory pre-stimulus activity, the role of working memory in this anticipatory process, and the relation of pre-stimulus activity to post-stimulus processing. In both our experiments, we demonstrate that this process occurs spontaneously, since the predictability of future events was manipulated orthogonally to the task demands and afforded participants no explicit benefit nor did it impose any explicit memory requirements. Our MEG results demonstrated a component (Fig. 2C) highly similar to the commonly reported novelty and ‘odd-ball’ components (Polich and Comerchero 2003; Gonsalves et al. 2005; Cycowicz and Friedman 2007; Polich 2007). These findings are completely consistent with prior work showing spontaneous learning in image streams that contain deterministic association between image pairs (Turk-Browne et al. 2010; Reddy et al. 2015) and with studies that documented statistical learning of the more and less probable transitions (Bornstein and Daw 2012; Tobia, Iacovella and Hasson 2012). In the behavioral study we also found facilitated responses to stimuli that satisfied the more likely transition probability and this enhancement of pre-stimulus activity was further linked to preferential early attentional processing after stimulus presentation in statistical regular contexts (Fig. 4A). This therefore suggest that one’s ability to represent statistical relationships within the recent past helps anticipate events in the near future (via preparatory slow-wave neural oscillations) and thus facilitate early processing of future sensory inputs.

B. The role of working memory

Importantly, we found a convergence of evidence for the role of working memory in probabilistic inference. WMC was related to differences in neural processing during both the pre- and post-stimulus onset periods. Because our MEG experiment was designed to contrast pre-stimulus activity epochs between statistically-regular and random series, we could directly link anticipatory processes to the statistical structure of event sequences (such a contrast partials out any general anticipatory
mechanisms). We found that during the pre-stimulus period, individuals with higher WMC showed greater differentiation in the theta frequency band, which held throughout the pre-stimulus period, but terminated just prior to the onset of the next stimulus (Fig. 3B). During pre-stimulus periods, higher WMC also correlated with greater differentiation in the alpha band, but this pattern held throughout the pre-stimulus period, as well as during stimulus presentation (Fig. 3E).

We also independently identified posterior MEG sensors where differential pre-stimulus theta power correlated with the difference in response to predictable vs. non-predictable stimuli just after stimulus presentation (Fig. 4A). This relationship was also found to be moderated by participant’s working memory (Fig. S3), indicating that the relation between statistically related anticipatory activity, and statistically influenced stimulus processing is linked to one’s working memory abilities. Including a WMC mediator was important for determining whether the direct link between pre- and post-stimulus activity is itself significant. In absence of the mediator variable, one would have concluded that pre-stimulus activity directly influences post-stimulus activity to predictability.

Taken together our findings suggest that one’s ability to represent the recent past helps anticipate events in the near future (via preparatory slow-wave neural oscillations) and thus facilitate early processing of future sensory inputs that could benefit behavior. In accordance with this notion, experiment 2 demonstrated that individuals with greater WMC also displayed more efficient behavioral responses to predictable than to non-predictable stimuli (see Results). In fact WMC was found to play such an important moderating role, that when inter-individual differences in WMC were not taken into account, differences between predictable and non-predictable stimuli were greatly minimized in the behavioral data, and were virtually absent in the pre-stimulus MEG findings. This result not only shows the relation between statistical learning processes and working memory capacity, but also suggests that future work may benefit from considering this factor of inter-individual differences in statistical learning - see also (Frost et al. 2015).

The WMC-related findings that we have identified, and the account we outlined, do however raise the question of how young children are able to achieve statistical learning (Saffran et al. 1999; Emberson et al. 2015) given their lower working memory capacity. In addressing this issue, we consider three themes that should be evaluated conjointly: i) what is the temporal integration windows over which statistical information appears to be integrated in adults, ii) is there indication that children’s WMC is sufficient to support integration on this scale, and iii) what is the degree of robustness of children’s statistical learning, and could these differences be related to WMC. Concerning integration time-windows in adults, as mentioned above, it is becoming increasingly clear that adults do not integrate statistical information in a way that equally weighs all prior instances in these types of paradigms. For instance, a study by Harrison et al. (2011) modeled the surprise response (prediction error) to stimuli and found that the best model assumed sensitivity to only the last 4 items. They concluded, “no matter how many samples are presented, observers have a threshold on the effective number of past observations that guide their behavior. This provides evidence that observers discount distant information when making inference about statistical regularities in their environment”. Similarly, Bornstein and Daw (2012) implemented a 4 x 4 transition matrix similar to ours (but with transitions holding between specific exemplars rather than categories). They found that the predictive power of stimulus A on stimulus B depended on how much time had passed since the last AB combination had been presented, with a very sharp decay function (over the last 4 trials), and noted that this pattern minimizes the efficacy of any model that does not incorporate a forgetting process. Finally, in our own work (Tobia, Iacobella and Hasson 2012) we implemented a paradigm where the Markov entropy changed gradually over short time scales. We identified brain regions sensitive to these entropy levels in the very recent past (within the prior 10 sec), but also other areas that tracked the trajectory of changes in regularity over greater time-windows (direction of gradual increase/decrease in regularity over longer time periods). To conclude this point, in stochastic contexts (i.e. non-deterministic), there is evidence that even adults retain information over a relatively recent window.

We note this might not be the case for deterministic contexts, i.e., ones where a stimulus is completely predictive of another stimulus (Turk-Browne et al. 2010), and which is an operationalization of statistical learning often used in children’s studies (where ‘words’ reflect a substring of stimuli that are completely predictive, e.g. Saffran et al. 1996). Can children’s working memory support such temporal windows? Studies using a WMC capacity assessment similar the one we used here, suggest that young children could most likely approach the adult WMC of 3-4 items in memory within the first year of life (Rose et al. 2001; Ross-Sheehy et al. 2003), and paradigms using object occlusion point to the ability to maintain 2-3 items in memory even at the age of 7 months (Moher et al. 2012). Yet comparisons of WMC in adults to that of the different stages of development in children can be sometimes quite tenuous due to differences in task design used to assess this capacity within different age groups may not always be directly comparable and should be taken with some degree of caution (Simmering 2012). To summarize, if behavior in stochastic contexts depends mainly on the very recent past, it might be that children could use their WMC to support this process, in a similar way to that of adults. It is important to note that while children’s ability is considered robust, commonly used measures to assess these capacities in children (e.g., preferential looking or preferential listening; Saffran et al. 1999; Kirkham et al. 2007) make it difficult to say just how robust this learning actually is in within and across individuals, especially when comparing these findings to those in adults. Finally, recent work on the relation between attention and statistical learning suggests that some statistical learning may take place “under the radar” with little cognitive control. For instance, automatic learning of statistical features of an unattended stream can take place when 2 information streams are presented in parallel (Musz et al., 2015). This may allow children, who have less developed control mechanisms to achieve statistical learning.

**C. Neural oscillations in the Theta- and Alpha- frequency bands**

Recent work has shown that neural activity occurring prior to the presentation of a visual stimulus can strongly impact subsequent stimulus processing. Particularly relevant to our study, pre-stimulus increases in theta power have been shown
to facilitate memory encoding (Jutras et al. 2013) and delayed retrieval (Addante et al. 2011) of visual items committed to memory. Neural modulations within the theta band have also been associated with the active maintenance of complex visual stimuli, similar to those used here, during the retention intervals of visual working memory tasks (Cashdollar et al. 2009; Cashdollar et al. 2013). Furthermore, theta activity has been known to support the maintenance of temporally separate events (Hsieh et al. 2011; Roberts et al. 2013) and has been proposed as a likely neural mechanism supporting the process of pattern completion (Hasselmo et al. 1995) to instantiate the next event when only fragmentary information is available.

On the basis of these findings we suggest that in the current study, theta band activity during pre-stimulus periods may not indicate solely the construction of a prediction, but perhaps fulfills a dual role within statistically regular sequences – that of consolidation and prediction. Consider the schematic series A, B, C that stands for a subsection of a statistically regular series. On the one hand, theta band activity may be related to forming retrospective associative links between the current and prior event, i.e., after being presented events A, B, the link of AB is reinforced, and the strength of this link reflects the transition strengths between the categories. Such an operation would be consistent with evidence showing individuals are sensitive to the probability of past events given the present (reverse transition probability, (Perruchet and Desaulty 2008; Pelucchi et al. 2009). On the other hand, activity in the theta band may also be related to constructing a prospective association, i.e., given the established link BC, the presentation of B will increase the accessibility of C. Therefore, the pre-stimulus theta activity found here may be related to both backward and forward oriented operations. The forward-looking operation may be described from the perspective of prediction making (minimizing uncertainty or free energy; Friston 2009). Both the forward and backward-linking operations are compatible with the perspective of automatic pattern completion as described by “chunking” approaches (Perruchet and Pacton 2006; Kumaran and Maguire 2009). Indeed, our own work (Tremblay et al. 2013) suggests that statistically regular auditory series are related to perceptual grouping. In this study, participants were presented with regular or random series that always contained 4 items. Yet, after hearing these series, participants reported hearing fewer distinct items in the regular series suggesting an associative binding occurred during the perception of these series (but this effect was absent in random series).

Pre-stimulus alpha band differences between statistically-regular and random series were also related to WMC, but these differences also encompassed the entire epoch in which the stimuli were on the screen which suggests a general, tonic modulation of attention between series. It has been previously shown that modulations of pre-stimulus alpha oscillations enhance the detection of low-level visual features (van Dijk et al. 2008; Mathewson et al. 2009; Spaak et al. 2014). Here we demonstrate a similar effect for a higher-level cognitive process, where pre-stimulus alpha differences between statistically-regular and random series enhanced the post-stimulus sensitivity to surprising (Ord_25) stimuli as compared to predictable (Ord_75) stimuli (Fig. 4B). However, this correlation relationship was not moderated by WMC (Supplemental Methods) suggesting that WMC is not uniquely responsible for this relationship as measured here.

It has recently been suggested that visual WMC, as assessed by a change detection task, is related to the ability to probabilistically infer future sensory information (Ma et al. 2014). For instance, in cases where sensory features are probabilistically varied in such tasks (i.e., likelihood of cue to target match), observers have been shown to retain the corresponding probabilities over time in order to optimize their behavioral precision (Najemnik and Geisler 2005; Ma et al. 2011). Furthermore, it has also been found that associatively binding items within a static visual display into patterns of occurrence will affect the precision of subsequent change detection estimates (Brady and Tenenbaum 2013). In comparison, these studies manipulated low-level visual features within a static visual search display (Hollingworth et al. 2008; Carlisle et al. 2011), or target probabilities in the context of the change detection task itself (Najemnik and Geisler 2005; Ma et al. 2011), and did not address the process of probabilistic inference in a more naturalistic setting where events occur over an extended time series and where the specific visual features of subsequent items could not be known with certainty. Our work suggests that such WMC driven predictions occur spontaneously during perception since the predictability of yet-unseen visual categories was manipulated orthogonally to the task demands and afforded participants no explicit benefit. Overall, this is an intuitively satisfying notion that one’s ability to retain the recent past allows them to predict the near future, and our results are consistent with work showing that when modeling the statistical features of a sensory input, individuals tend to be influenced primarily by events in the very recent past (Harrison et al. 2011).

In summary, our study provides a convergence of evidence establishing the vital role of working memory in probabilistic inference of future events. This capacity to represent the recent past specifically facilitates processing within statistically regular contexts in that it boosts anticipatory slow-wave neural oscillations prior to the appearance of visual information and preferential sensory processing for these anticipated events once they are presented. This process, which is moderated by one's working memory ability, appears to intrinsically occur during continuous perception and poses a realistic advantage for human behavior by allowing the anticipation of higher-level semantic concepts that are expected to occur in the near future.

REFERENCES


