Neural mechanisms underlying the influence of sequential predictions on scene gist recognition

by

Maverick Earl Smith

B.S., Mississippi State University, 2015
M.S., Kansas State University, 2019

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Psychological Sciences
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2021

Abstract

Rapid scene categorization is typically argued to be a purely feed-forward process. Yet, when navigating in our environment, we usually see predictable sequences of scene categories (e.g., offices followed by hallways, parking lots followed by sidewalks, etc.). Previous work showed that scenes are easier to categorize when they are shown in ecologically valid, predictable sequences compared to when they are shown in randomized sequences (Smith & Loschky, 2019). Given the number of stages involved in constructing a scene representation, we asked a novel research question: when in the time course of scene processing do sequential predictions begin to facilitate scene categorization? We addressed this question by measuring the temporal dynamics of scene categorization with electroencephalography. Participants saw scenes in either spatiotemporally coherent sequences (first-person viewpoint of navigating, from, say, an office to a classroom) or their randomized versions. Participants saw 10 scenes, presented in rapid serial visual presentation (RSVP), on each trial, while we recorded their visually event related potentials (vERPs). They categorized 1 of the 10 scenes from an 8 alternative forced choice (AFC) array of scene category labels. We first compared event related potentials evoked by scenes in coherent and randomized sequences. In a subsequent, more detailed analysis, we constructed scene category decoders based on the temporally resolved neural activity. Using confusion matrices, we tracked how well the pattern of errors from neural decoders explain the behavioral responses over time and compared this ability when scenes were shown in coherent or randomized sequences. We found reduced vERP amplitudes for targets in coherent sequences roughly 150 milliseconds after scene onset, when vERPs first index rapid scene categorization, and during the N400 component, suggesting a reduced semantic integration cost in coherent sequences. Critically, we also found that confusions made by neural decoders and human responses correlate more strongly in coherent sequences, beginning around 100 milliseconds. Taken together, these results suggest that predictions of upcoming scene categories influence even the earliest stages of scene processing, affecting both the extraction of visual properties and meaning.

Neural mechanisms underlying the influence of sequential predictions on scene gist recognition

by

Maverick Earl Smith

B.S., Mississippi State University, 2015
M.S., Kansas State University, 2019

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Psychological Sciences
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2021

Approved by:

Major Professor
Dr. Lester Loschky

Copyright

Abstract

Rapid scene categorization is typically argued to be a purely feed-forward process. Yet, when navigating in our environment, we usually see predictable sequences of scene categories (e.g., offices followed by hallways, parking lots followed by sidewalks, etc.). Previous work showed that scenes are easier to categorize when they are shown in ecologically valid, predictable sequences compared to when they are shown in randomized sequences (Smith & Loschky, 2019). Given the number of stages involved in constructing a scene representation, we asked a novel research question: when in the time course of scene processing do sequential predictions begin to facilitate scene categorization? We addressed this question by measuring the temporal dynamics of scene categorization with electroencephalography. Participants saw scenes in either spatiotemporally coherent sequences (first-person viewpoint of navigating, from, say, an office to a classroom) or their randomized versions. Participants saw 10 scenes, presented in rapid serial visual presentation (RSVP), on each trial, while we recorded their visually event related potentials (vERPs). They categorized 1 of the 10 scenes from an 8 alternative forced choice (AFC) array of scene category labels. We first compared event related potentials evoked by scenes in coherent and randomized sequences. In a subsequent, more detailed analysis, we constructed scene category decoders based on the temporally resolved neural activity. Using confusion matrices, we tracked how well the pattern of errors from neural decoders explain the behavioral responses over time and compared this ability when scenes were shown in coherent or randomized sequences. We found reduced vERP amplitudes for targets in coherent sequences roughly 150 milliseconds after scene onset, when vERPs first index rapid scene categorization, and during the N400 component, suggesting a reduced semantic integration cost in coherent sequences. Critically, we also found that confusions made by neural decoders and human responses correlate more strongly in coherent sequences, beginning around 100 milliseconds. Taken together, these results suggest that predictions of upcoming scene categories influence even the earliest stages of scene processing, affecting both the extraction of visual properties and meaning.

Table of Contents

List of Figures xi
List of Tables xx
Acknowledgements xxiii
Chapter 1 - Neural mechanisms underlying the influence of sequential predictions on scene gist recognition 1
Feed-forward gist processing 1
The Scene Perception and Event Comprehension Theory 6
Predictions may facilitate everyday scene perception 7
Mechanisms of scene facilitation 10
Early facilitation accounts 10
Matching process accounts 11
Post-identification accounts 14
Current Experiment 14
VERP Components of Interest 15
The P200. 16
The N400. 17
Hypotheses 18
Chapter 2 - Experiment 1 22
Method 23
Participants 23
Materials 24
Procedure 30
Results 32
Full study results overview 32
Behavioral Results 41
Discussion 46
Chapter 3 - Experiment 2 48
Method 50
Participants 50

EEG Data Acquisition and Preprocessing 51
Evoked Potential Recordings 53
Neural Decoding Procedure 56
Apparatus 58
Procedure 58
Results 60
Behavioral Results 60
vERPs to the Target Image 65
Frontal and Central Electrodes 67
50-149 ms window 70
150-249 ms window 71
250-449 ms window 72
Parietal/Occipital Electrode Sites. 73
50-149 ms window 75
150-249 ms window 76
vERPs to all of the images 77
Frontal and Central Electrodes 78
50-149 ms window 81
150-249 ms window 81
250-449 ms window 81
Parietal/Occipital Electrodes 82
50-149 ms window 83
150-249 ms window 84
Analysis of vERP divergence 84
Exploratory analyses of source localization 89
Changes in vERPs within a trial 97
Frontal and Central Electrodes 98
50-149 ms window 101
150-249 ms window 102
250-449 ms window 103
50-149 ms window 107

150-249 ms window. 108
Exploratory analyses of image predictability and image similarity 108
Neural Decoding of Image Categories 119
Simultaneity of visual representation and behavioral categorization 124
Discussion 130
Facilitation of the P200 131
The effect of image similarity and predictability on scene processing 134
Neural mechanisms underlying the effect of predictions on scene processing 135
Facilitation of the N400 137
Predictable scenes elicit clearer neural signals 140
Chapter 4 - Experiment 3 143
Method 144
Participants 144
EEG Data Acquisition and Preprocessing 145
Procedure 145
Results 146
Behavioral Results 147
vERPs to the Target Image 150
Frontal/Central Electrode Sites. 153
50-149 ms window. 156
150-249 ms window. 156
250-449 ms window. 157
Parietal/Occipital Electrode Sites. 158
50-149 ms window. 160
150-249 ms window. 160
vERPs to all of the images 161
Analysis of vERP divergence 162
Exploratory analyses of source localization 163
Changes in vERPs within a trial 169
Frontal and Central Electrodes. 170
50-149 ms window. 173

150-249 ms window..... 174
250-449 ms window..... 175
Parietal/Occipital Electrodes..... 175
50-149 ms window..... 178
150-249 ms window..... 179
Exploratory analyses of image predictability and image similarity..... 179
Neural Decoding of Image Categories..... 187
Simultaneity of visual representation and behavioral categorization..... 193
Discussion..... 197
Facilitation of vERPs..... 197
Neural mechanisms underlying the effect of predictions on scene processing..... 198
Decoding of brain signals..... 200
Chapter 5 - General Discussion..... 205
References..... 214
Appendix A - Analysis of the N300/N400..... 230
Experiment 2: Analyses including the N300..... 234
vERPs to the Target Image..... 234
N300..... 236
N400..... 237
vERPs to all of the images..... 239
N300..... 240
N400..... 241
Changes in vERPs within a trial..... 241
N300..... 243
N400..... 244
Experiment 3: Analyses including the N300..... 246
vERPs to the Target Image..... 246
N300..... 247
N400..... 248
vERPs to all of the images..... 249
N300..... 250

N400..... 250
Changes in vERPs within a trial 251
N300..... 253
N400..... 253
Appendix B - Analysis with N300 removed..... 255
vERPs to all of the images..... 255
Frontal/Central Electrodes. 255
50-149 ms window..... 256
150-249 ms window..... 257
250-449 ms window..... 257
Parietal/Occipital Electrodes..... 258
50-149 ms window..... 259
150-249 ms window..... 260
Analysis of vERP divergence 260

List of Figures

Figure 1. This is a trial schematic used in Smith & Loschky (2019). Panel a) is a simplified version of a trial. The sequence of scenes in i) are coherent. They begin in an office and end in a parking lot. The sequence of scenes in ii) are the same scenes in a randomized order. Panel b) illustrates a complete trial. After the scene was categorized, participants were shown a fixation dot, and then they pressed a button to view the remaining scenes from each sequence. A continuation of the images were shown in all cases when the target that participants were asked to categorize from 8 alternative options was not the 10th (i.e., final) image in a sequence. In that way, participants saw a complete 10-image sequence on every trial, but never knew which item in the sequence would be tested. 9

Figure 2. Example of a full 20 image base sequence for "Office to Parking Lot." There were 24 such base sequences, with 3 different exemplar base sequences for each, for a total of 72 such sequences (see Table 1 for full list). Each base sequence is from a first-person viewpoint, and is shown here in the spatiotemporally coherent order. Participants saw a subset of 10 images from such a base sequence on each trial. The images shown within each subset were randomly determined for each participant. 29

Figure 3. This is a trial schematic. Scenes were shown in either a a) coherent or b) randomized sequence. Scenes in the coherent sequence represented here, begin in an office and end in a parking lot. Instead of pausing the sequence when the target was presented (as in Figure 1), the target was absent, so the sequence paused after presentation of a 700 ms blank screen followed by a 100 ms noise mask. The mask thus served as a cue to participants to predict what the next scene would be. After participants selected their prediction from the 8-AFC response screen, they were shown a fixation dot, and they pressed a button to view the target they were asked to predict and the remaining scenes from each sequence. A continuation of the images were shown in all cases when the target that participants were asked to predict was not the 10th (i.e., final) scene. 30

Figure 4. Exp 1: Image predictability as a function of the spatiotemporal coherence of the image sequences. The proportion of times each image was accurately predicted is represented by the gray lines. Least square means generated from the estimated regression equation are

represented by the thick black line. The dashed line at 12.5% represents chance performance. 43
Figure 5 Exp 1: Image predictability as a function of the ordinal position (2-10) of the target on each trial, the spatiotemporal coherence of the image sequences, and the location the images were photographed. The proportion of instances when the target was correctly predicted at each ordinal position (2-10) is represented by the dots. The lines reflect the least square means calculated from the estimated regression equation. Error ribbons reflect 1 standard error to the estimated means. The dashed line at 12.5% represents chance performance. ... 46
Figure 6. EGI 64-channel sensor layout. The EGI 64 channel HydroCell Geodesic Sensor Net is displayed above. Each region of interest for the analysis is color coded. Red, orange, and yellow electrodes represent frontal electrodes, green electrodes represent central electrodes, and blue electrodes represent parietal/occipital electrodes. Electrodes indicated with an astericks correspond to locations of the 10-20 system..... 55
Figure 7. This is a trial schematic of Experiment 2. Scenes were shown in either a a) coherent or b) randomized sequence. Participants were asked to categorize 1 target on each trial in an 8 alternative forced choice task. The ordinal position of the target (2-10) was randomly chosen on each trial. The target scene and its ordinal position was the same in the coherent and randomized sequences. 58
Figure 8. Exp 2: Rapid scene gist categorization performance as a function of the spatiotemporal coherence of the image sequences. The proportion of times each participant accurately categorized the target images is represented by the thin gray lines. Least square means generated from the estimated regression equation are represented by the thick black line and dots. 62
Figure 9. Exp 2: Rapid scene gist categorization accuracy as a function of the ordinal position (2-10) of the target scene on each trial and the spatiotemporal coherence of the image sequences. The proportion of instances when the target image was correctly categorized is represented by dots in the figure. The lines reflect the least square means calculated from the estimated regression equation. 64
Figure 10. Exp 2: Grand average vERP waveforms time locked to the *target* image for the frontal and central regions. Responses to target scenes in the coherent sequences are represented in red and responses to scenes in the randomized sequences are in blue. The

difference between the waveforms of the coherent and randomized sequences are represented by the black line. Error ribbons correspond to 1 standard error to raw means.. 66
Figure 11. Exp 2: Grand average vERP waveforms time locked to the *target* image on a trial for the Parietal/Occipital regions. Responses to target images in the coherent condition are represented in red and responses to the images in the randomized condition are in blue. The difference between waveforms of the coherent and randomized sequences are represented by the black line. Error ribbons correspond to 1 standard error to raw means. 67
Figure 12. Exp 2: Least square means of amplitudes in response to the *target* at the frontal and central regions. Amplitudes are reported for the a) 50-149, b) 150-249, and c) 250-449 ms windows averaged across the location factor. Error bars correspond to 1 standard error around the estimated means. 68
Figure 13. Exp 2: Least square means of amplitudes in response to the *target* image at the parietal and occipital sites from a) 50-149 ms and b) 150-249 ms. Amplitudes in the coherent condition were not significantly different from amplitudes in the randomized condition in either the early component or in the analysis of the P200. See the text for details. 74
Figure 14. Exp 2: Scalp maps of the mean voltage time locked to the onset of scenes within the a) coherent, b) randomized sequences. The difference between the coherent and randomized conditions are represented in c). Scalp maps do not include behaviorally incorrect trials or responses to the first scene within a trial. Voltage ranged from -10 to +10 microvolts in the coherent and randomized sequences and -3 to +3 in the difference maps. 78
Figure 15. Exp 2: Amplitudes in response to all of the images shown in the coherent and randomized conditions excluding behaviorally incorrect trials and cases when the image was the first scene within a trial at the frontal and central regions. Amplitudes are averaged across the location factor. Responses to images presented in coherent sequences differed significantly from responses to images in randomized sequences in the 150-249 and 250-449 windows. 79
Figure 16. Exp 2: Amplitudes in response to all of the scenes shown in the coherent and randomized conditions at the Parietal/Occipital regions, excluding behaviorally incorrect trials and cases when the image was the first scene within a trial at the parietal/occipital

sites. Responses to images shown in the coherent and randomized conditions did not significantly differ in either the early component (50-149) or in the P200. 82
Figure 17. Exp 2: Grand average vERPs time locked to the onset of the scenes at time 0. Scene were presented in either coherent or randomized sequences. Average waveforms at a) Frontal b) Central, and c) Parietal/Occipital sites are on the top row. Bayes factors for each of the paired sample t tests within the epoch for d) Frontal e) Central, and f) Parietal/Occipital electrodes are provided in the bottom row. Green patches represent statistically significant comparisons. Red dashed lines in the Bayes factors plots represent a Bayes factor of 3 and purple lines represent a Bayes Factor of 1 and -1 respectively. 88
Figure 18. Exp 2: Grand average waveforms at *frontal* electrodes time locked to the onset of the scenes back projected from each of the 12 clusters of independent components. The cluster indices are represented in a) through l). Waveforms in response to scenes in the coherent sequences are in red, and waveforms from the randomized sequences are in blue. Green patches represent significant comparisons. 95
Figure 19. Exp 2: Grand average waveforms at *central* electrodes time locked to the onset of the scenes back projected from each of the 12 clusters of independent components. The cluster indices are represented in a) through l). Waveforms in response to scenes in the coherent sequences are represented in red, and waveforms in response to scenes shown in the randomized sequences are in blue. Green patches represent significant comparisons. 96
Figure 20. Exp 2: Clusters of sources of independent components for all subjects across trials and conditions for clusters a) 8 (Brodmann area 7: Precuneus) and b) 12 (Brodmann are 23: Posterior Cingulate Cortex). 97
Figure 21. Exp 2: Amplitudes at each ordinal position (1-10), excluding behaviorally incorrect trials. Responses to images in coherent sequences are in red, and responses to images in randomized sequences are in blue. 99
Figure 22. Exp 2: Amplitudes in response to each ordinal position (1-10) of a scene on a trial, excluding behaviorally incorrect trials. Amplitudes in response to scenes in coherent sequences are in red, and amplitudes in randomized sequences are in blue. 105
Figure 23. Exp 2: Scatterplots between the log of *image similarity* and voltage at LF) left, MF) middle, and RF) right *frontal* as well as LC) left and RC) right *central* regions..... 112

Figure 24. Exp 2: Scatterplots between the log of *image similarity* and voltage at LO) left, MO) middle, and RO) right *parietal/occipital* regions. 113
Figure 25. Exp 2: Scatterplots between *image predictability* and voltage at LF) left, MF) middle, and RF) right *frontal* as well as LC) left and RC) right *central* regions. 114
Figure 26. Exp 2: Scatterplots between *image predictability* and voltage at LO) left, MO) middle, and RO) right *parietal/occipital* regions. 115
Figure 27. Exp 2: ***Decoding accuracy*** as a function of time in the epoch for the a) Frontal, b) Central, and c) Parietal/Occipital regions. Bayes Factors for each of the paired sample t tests within the epoch for d) Frontal e) Central, and f) Parietal/Occipital electrodes are provided in the bottom row. Green patches represent clusters of statistically significant comparisons. Red dashed lines in the Bayes Factors plots represent a Bayes Factor of 3 and purple lines represent a Bayes Factor of 1 and -1 respectively. Error ribbons correspond to 95% confidence intervals around the means. 121
Figure 28. Exp 2: Decoding accuracy after the onset of the images as a function of the spatiotemporal coherence of the image sequences. Decoding accuracy for individual participants are represented by the lines. Least square means generated from the estimated regression equation are represented by the thick black line and dots. The dashed line at .125 represents chance level performance. 123
Figure 29. Exp 2: Confusion matrices for coherent and randomized image sequences for on- (top row) and off-campus (bottom row) images. Confusions in coherent sequences across participants are represented in a) and c). Confusions in randomized sequences across participants are represented in b) and d). Rows represent the target image category, and columns represent the average responses made for each response category. Thus, responses on the main diagonal are correct responses. Images belonging to indoor categories were often confused with other indoor categories, and images belonging to outdoor categories were often confused with other outdoor categories. 127
Figure 30. Exp 2: *Unique variance* in the behavioral confusion matrices explained by confusions made by the neural decoders over time. Error bars represent between subject 95% confidence intervals at each time point. 130
Figure 31. Exp 3: Rapid scene gist categorization performance as a function of the spatiotemporal coherence of the image sequences. The proportion of times each participant

accurately categorized the target scenes is represented by the lines. Least square means generated from the estimated regression equation are represented by the thick black line and dots. 148
Figure 32. Exp 3: Rapid scene gist categorization accuracy as a function of the ordinal position (2-10) of the target scene on each trial, the spatiotemporal coherence of the image sequences, and the location the images were photographed. The proportion of instances when the target image was correctly categorized is represented by dots in the figure. The lines reflect the least square means calculated from the estimated regression equation..... 150
Figure 33. Exp 3: Grand average vERP waveforms time locked to the *target* image for the Frontal/Central electrodes. Responses to target images in the coherent condition are in red and responses to the images in the randomized condition are in blue. The difference between the coherent and randomized lines are represented by the black line in the figure. 152
Figure 34. Exp 3: Grand average vERP waveforms time locked to the *target* image for the Parietal/Occipital electrodes. Responses to target images in the coherent condition are in red and responses to the images in the randomized condition are in blue. 153
Figure 35. Exp 3: Least square means of amplitudes in response to the *target* image at the frontal and central sites. Amplitudes are reported for the a) 50-149, b) 150-249, and c) 250-449 windows. 154
Figure 36. Exp 3: Least square means of amplitudes in response to the *target* scene at the parietal/occipital regions. Amplitudes are reported for the a) 50-149, and b) 150-249 windows. Error bars correspond to 1 standard error around the estimated means. 159
Figure 37. Exp 3: Scalp maps of the mean voltage time locked to the onset of scenes within the a) coherent, b) randomized sequences. The difference between the coherent and randomized conditions are represented in c). Scalp maps do not include behaviorally incorrect trials or responses to the first scene within a trial. Voltage ranged from -10 to +10 microvolts in the coherent and randomized sequences and -5 to +5 in the difference maps. 162
Figure 38. Exp 3: Grand average waveforms at frontal regions time locked to the onset of the scenes back projected from each of the 12 clusters of independent components. Waveforms in response to scenes shown in the coherent sequences are represented in red and

waveforms in response to images shown in the randomized sequences are represented in blue. Green patches represent clusters of statistically significant comparisons. 166
Figure 39. Exp 3: Grand average waveforms at central electrodes time locked to the onset of the images back projected from each of the 12 clusters of independent components. Waveforms in response to scenes shown in the coherent sequences are represented in red and waveforms in response to images shown in the randomized sequences are represented in blue. Green patches represent clusters of significant comparisons..... 167
Figure 40. Exp 3: Clusters of sources of independent components for all subjects across trials and conditions for a) Cluster 2 (Brodmann area 23), b) Cluster 4 (Brodmann are 31), c) Cluster 6 (Brodmann area 6), and d) Cluster 9 (Brodmann area 37). 168
Figure 41. Exp 3: Frontal/Central electrode amplitudes in response to each ordinal position (1-10) of the scenes on a trial, excluding behaviorally incorrect trials. Responses to scenes in coherent sequences are in red, and responses to scenes in randomized sequences are in blue. 171
Figure 42. Exp 3: Parietal/Occipital electrodes average amplitudes time locked to the onset of scenes shown at each ordinal position. Behaviorally incorrect trials were removed from this analysis. Responses to images in coherent sequences are in red, and responses to images in randomized sequences are in blue..... 176
Figure 43. Exp 3: Scatterplots between *image similarity* and voltage at LF) left, MF) middle, and RF) right *frontal* as well as LC) left and RC) right *central* regions. 180
Figure 44. Exp 3: Scatterplots between the log of ***image similarity*** and voltage at LO) left, MO) middle, and RO) right ***parietal/occipital*** regions..... 181
Figure 45. Exp 3: Scatterplots between *image predictability* and voltage at LF) left, MF) middle, and RF) right *frontal* as well as LC) left and RC) right *central* regions..... 182
Figure 46. Exp 3: Scatterplots between *image predictability* and voltage at LO) left, MO) middle, and RO) right *parietal/occipital* regions..... 183
Figure 47. Exp 3: ***Decoding accuracy*** as a function of time in the epoch for the a) Frontal, b) Central, and c) Parietal/Occipital regions. Bayes Factors for each of the paired sample t tests within the epoch for d) Frontal e) Central, and f) Parietal/Occipital electrodes are provided in the bottom row. Green patches represent clusters of statistically significant comparisons. Red dashed lines in the Bayes Factors plots represent a Bayes Factor of 3 and purple lines

represent a Bayes Factor of 1 and -1 respectively. Error ribbons correspond to 95% confidence intervals around the means. 189
Figure 48. Exp 3: Decoding accuracy after the onset of the images as a function of the spatiotemporal coherence of the sequences. Decoding accuracy for individual participants are represented by the light gray lines, and least square means generated from the estimated regression equation are represented by the thick black line and dots. The dashed line at 12% represents chance level performance. 191
Figure 49. Exp 3: Confusion matrices for coherent and randomized image sequences for on- (top row) and off-campus (bottom row) images. Confusions in coherent sequences across participants are represented in a) and c). Confusions in randomized sequences across participants are represented in b) and d). Rows represent the target image category, and columns represent the average responses made for each response category. Thus, responses on the main diagonal are correct responses. Images belonging to indoor categories were often confused with other indoor categories, and images belonging to outdoor categories were often confused with other outdoor categories. 194
Figure 50. Exp 3: *Unique variance* in the behavioral confusion matrices explained by confusions made by the neural decoders over time. Error bars represent between subject 95% confidence intervals at each time point. 196
Figure 51. Exp 2: Least square means of amplitudes in response to the target at the frontal and central regions. Amplitudes are reported for the a) N300 and d) N400 components. Error bars correspond to 1 standard error around the estimated means. 235
Figure 52. Exp 2: Least square means of amplitudes in response to the all of the scenes excluding behaviorally incorrect trials at the frontal and central regions. Amplitudes are reported for the a) N300 and d) N400 components. Error bars correspond to 1 standard error around the estimated means. 239
Figure 53. Exp 2: Amplitudes at each ordinal position (1-10), excluding behaviorally incorrect trials. Responses to images in coherent sequences are in red, and responses to images in randomized sequences are in blue. 242
Figure 54. Exp 3: Least square means of amplitudes in response to the target scene at the frontal and central regions. Amplitudes are reported for the a) N300 and d) N400 components. Error bars correspond to 1 standard error around the estimated means. 246

Figure 55. Exp 3: Least square means of amplitudes in response to all of the scenes, excluding behaviorally incorrect trials, at the frontal and central regions..... 249
Figure 56. Exp 3: Amplitudes at each ordinal position (1-10), excluding behaviorally incorrect trials. Responses to images in coherent sequences are in red, and responses to images in randomized sequences are in blue..... 251
Figure 57. Exp 3: Least square means of amplitudes in response to all of the scenes, excluding behaviorally incorrect trials, at the frontal and central regions. Amplitudes are reported for the a) 50-149, b) 150-249, and c) 250-449 windows..... 255
Figure 58. Exp 3: Least square means of amplitudes in response to all of the scenes, excluding behaviorally incorrect trials, at the parietal/occipital regions. Amplitudes are reported for the a) 50-149, b) 150-249, windows..... 258
Figure 59. Exp 3: Grand average vERPs time locked to the onset of the scenes at time 0. Scenes were presented in either coherent or randomized sequences. Average waveforms at a) Frontal b) Central, and c) Parietal/Occipital sites are on the top row. Bayes factors for each of the paired sample t-tests within the epoch for d) Frontal e) Central, and f) Parietal/Occipital electrodes are provided in the bottom row. Green patches represent clusters of statistically significant comparisons. Red dashed lines in the Bayes factors plots d) through f) represent a Bayes Factor of 3 and purple lines represent a Bayes factor of 1 and -1, respectively. 263

List of Tables

Table 1. *Table of base sequences. Each base sequence begins with a starting location and terminates in a destination location. Observers appear to be navigating through the transitional categories when navigating between starting and destination locations. There were 3 versions of each base sequence showing the same starting and destination categories, but from different locations.* 27

Table 2. *Summary of the results of Experiments 1, 2, and 3.*..... 34

Table 3 Exp 2: *Summary of the results for the frontal and central regions for each of the three windows (starting from 50 ms to 449 ms). Amplitudes were time locked to the onset of the target scene.* 69

Table 4. Exp 2: *Summary of the results for the parietal/occipital regions. Amplitudes were time locked to the target scene.*..... 74

Table 5 Exp 2: *Summary of the results for the frontal/central electrodes. Amplitudes were time locked to the onset of the scenes in the experiment. Observations from the first image within each sequence were removed from the analyses as well as behaviorally incorrect responses to the target.* 80

Table 6. Exp 2: *Summary of the results for the parietal/occipital electrodes. Amplitudes were time locked to the onset of the scenes in the experiment. Observations from the first scene within each sequence and behaviorally incorrect responses to the target were removed from the analysis.*..... 83

Table 7 Exp 2: *Montreal Neurological Institute (MNI) coordinates and labels of the centroids of independent component clusters. Clusters are in no particular order.* 91

Table 8. Exp 2: *Summary of the results for the frontal/central electrodes. Amplitudes were time locked to the onset of the scenes in the experiment in the 1st through the 10th position.* 100

Table 9. Exp 2: *Summary of the results for the parietal/occipital electrodes. Amplitudes were time locked to the onset of the images in the experiment in the 1st – 10th position.*..... 106

Table 10. Exp 2: *Partial correlation coefficients between the log of image similarity and mean amplitude, controlling for image predictability. Partial correlation coefficients between image predictability and the mean amplitudes within each of the time windows (50-149), (150-249), and (250-449) controlling for the effect of log of image similarity.* 116

Table 11. Exp 3: *Summary of the results for the frontal/central electrodes. Amplitudes were time locked to the target image.* 155
Table 12. Exp 3: *Summary of the results for the parietal/occipital electrodes. Amplitudes were time locked to the target image.* 159
Table 13. Exp 3: *Montreal Neurological Institute (MNI) coordinates and labels of the centroids of independent component clusters. Cluster are in no particular order.* 164
Table 14. Exp 3: *Summary of the results for the frontal/central electrodes. Amplitudes were time locked to the onset of the scenes in the experiment in the 1st through the 10th positions.* 172
Table 15. Exp 3: *Summary of the results for the parietal/occipital electrodes. Amplitudes were time locked to the onset of the images in the experiment in the 1st – 10th position.* 177
Table 16. Exp 3: *Partial correlation coefficients between the log of image similarity and mean amplitude, controlling for image predictability. Partial correlation coefficients between image predictability and the mean amplitudes within each of the time windows (50-149), (150-249), and (250-449) controlling for the effect of log of image similarity.* 184
Table 17. *Electrodes within each region of interest for each of the 3 vERP components of interest. Authors who demonstrated differential activity between expected and unexpected stimuli at each electrode location are provided in the far-right column.* 231
Table 18. Exp 2: *Summary of the results for the frontal and central regions. Amplitudes were time locked to the onset of the target scene.* 235
Table 19. Exp 2: *Summary of the results for the frontal/central electrodes. Amplitudes were time locked to the onset of the scenes in the experiment. Observations from the first image within each sequence were removed from the analyses as well as behaviorally incorrect responses to the target.* 240
Table 20. Exp 2: *Summary of the results for the frontal/central electrodes. Amplitudes were time locked to the onset of the scenes in the experiment in the 1st through the 10th position.* 243
Table 21. Exp 3: *Summary of the results for the frontal and central regions. Amplitudes were time locked to the onset of the target scene.* 247
Table 22. Exp 3: *Summary of the results for the frontal and central electrodes. Amplitudes were time locked to the onset of all of the scenes within a trial excluding observations from the first image within each trial and behaviorally incorrect responses to the target.* 249

Table 23. Exp 3: *Summary of the results for the frontal/central electrodes. Amplitudes were time locked to the onset of the scenes in the experiment in the 1st through the 10th position.* 252
Table 24. Exp 3: *Summary of the results for the frontal and central electrodes. Amplitudes were time locked to the onset of the scenes. Observations from the first image within each sequence were removed from the analyses as well as behaviorally incorrect responses to the target.* 256
Table 25. Exp 3: *Summary of the results for the parietal/occipital electrodes. Amplitudes were time locked to the onset of the scenes. Observations from the first scene within each sequence were removed from the analyses as well as behaviorally incorrect responses to the target.* 259

Acknowledgements

I would like to give a special thanks to my advisor Dr. Lester Loschky for his continued guidance, support, and dedication since I arrived at Kansas State University. His vast amount of knowledge and attention to detail changed my approach to research, writing, and teaching. Many thanks also go to committee members Dr. Heather Bailey (Kansas State University), who was a second mentor to me through graduate school, Dr. Michelle Greene (Bates College), and Dr. Matthew Wisniewski (Kansas State University) for their help through the course of completing this project. A special extension of my gratitude goes to Dr. Alexandria Zakrzewski, the Electroencephalography Core director, and Destiny Bell for their help with the EEG system and data analyses. I was very fortunate to have the opportunity to work with them, and I will forever be grateful for all of their help. Additionally, I would like to thank the many current and former graduate and undergraduate research collaborators who gave me feedback on the project and helped run participants and analyses. They also provided emotional support and encouragement throughout graduate school. The undergraduate collaborators include: Thomas Hinkel, Katie Tran, Yuhang (Doris) Ma, Kahler Fontaine, Ashley Faiola, and Cashel Fitzgibbons, and the graduate student collaborators include: Dr. Ryan Ringer, Dr. Jared Peterson, Dr. John Hutson, Dr. Kimberly Newberry, Kari Payne, and Prasanth Chandran, and Taylor Simonson.

Research reported in this dissertation was supported by the Cognitive and Neurobiological Approaches to Plasticity (CNAP) Center of Biomedical Research Excellence (COBRE) of the National Institutes of Health under grant number P20GM113109. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health.

Chapter 1 - Neural mechanisms underlying the influence of sequential predictions on scene gist recognition

Imagine you are visiting a friend at their house for the very first time. You probably have a prediction as to what scene you will see once you are greeted by your friend after you ring their doorbell, even though you have never been inside their home before this visit. Perhaps the scene you expect to see is a hallway or a living room because those are the scene categories that are typically found near the entryway of houses. Such scene categories are much more likely than a bedroom or bathroom to be near the front door of someone's home; however, a bedroom or bathroom are both certainly more likely than an office or a forest to be on the other side of your friend's front door. You rarely, if ever, experience a scene category (e.g., an office, a forest, a hallway, etc.) you do not predict to see as you navigate your environment. Scene categories do not appear randomly and unexpectedly from one moment to the next even when you visit novel places. How do predictions made prior to viewing a scene influence your ability to *see* the upcoming scene?

Feed-forward gist processing

Despite the surprise you might experience if you saw an office cubicle inside your friend's home where you would expect to see their living room, research has found that people can accurately identify the *gist* of a scene presented for as little as 12 milliseconds followed by a perceptual mask (Bacon-Mace et al., 2005; Greene & Oliva, 2009a; Loschky et al., 2008; Potter et al., 2014). We define the theoretical construct of *scene gist* as a viewer's holistic semantic representation of a scene that can be acquired within a single eye fixation (Larson et al., 2014). Scene gist is an important construct in theories of scene perception because the gist of a scene influences where people look (Henderson & Hayes, 2017; Torralba et al., 2006), how objects are

perceived (Biederman et al., 1982; Davenport & Potter, 2004; Palmer, 1975a), and long-term memory for a scene's contents (Brewer & Treyens, 1981; Pezdek et al., 1989). Scene gist recognition has been operationalized in numerous ways, but it is usually in terms of the ability to classify a briefly flashed scene at the basic level (Schyns & Oliva, 1994; Tversky & Hemenway, 1983). However, the theoretical construct of gist implies more than how it is measured. Throughout this document, we will use the term *scene gist* when discussing the theoretical construct, and we will refer to *rapid scene categorization* when referring to how we operationalized the construct.

As mentioned previously, prior work has demonstrated that the ability to categorize scenes at their basic level is significantly above chance with masked viewing durations as fast as 12 ms, and scene categorization performance is near perfect at 100 ms stimulus onset asynchrony (SOA) - the time between the scene onset and the perceptual mask onset, which serves to make the scene harder to categorize (Bacon-Mace et al., 2005; Greene & Oliva, 2009a; Loschky, Simons, et al., 2007; Potter et al., 2014).

After visual information hits the retina, it is processed in a series of stages of increasing complexity through a hierarchy of brain regions, with cells at each successive stage processing inputs from increasingly larger regions of space (Marr, 1982; Riesenhuber & Poggio, 1999; Serre et al., 2007). The time course of scene gist perception has been investigated using a variety of different methods. Typically, in studies of rapid scene categorization, pictures of scenes are flashed very briefly in a rapid serial visual presentation (RSVP), ranging in length from a minimum of two, to a maximum of 20-30 images. Participants are either asked to categorize the scene from different options, where accuracy is the primary dependent measure (Biederman et al., 1974; Loschky, Sethi, et al., 2007; Potter, 1976), or respond by releasing a button or by

making an eye movement whenever a flashed scene belongs to a pre-specified target category (Go-no-go: e.g., release a button if the flashed scene is an office and withhold from responding otherwise), where reaction time is the primary dependent measure (Kirchner & Thorpe, 2006; Rousselet et al., 2002; Rousselet et al., 2005; Thorpe et al., 1996). Median reaction times to scenes presented in RSVP are around 400 milliseconds (approximately half a second), and accurate responses can be as fast as 260 milliseconds (Fabre-Thorpe et al., 2001; Thorpe et al., 1996). This implies that enough visual information has been processed in 260 milliseconds to begin to discriminate one scene category from another. Specifically, reaction time reflects the time needed to process the visual features of a scene, identify its category, and plan and execute a motor response. Thus, reaction times place an upper bound on the speed of visual processing.

Neurophysiological measures coupled with machine learning techniques have constrained this upper bound even further. Scene representations can successfully be classified using multivariate decoding methods from electroencephalographic (EEG) and magnetoencephalographic (MEG) signals as early as 100 ms (Bankson et al., 2018; Greene & Hansen, 2020; Lowe et al., 2018; Ramkumar et al., 2016), and these representations are similar to those that emerge in the layers of feedforward deep neural networks (Cichy et al., 2017; Greene & Hansen, 2018). Importantly, enough visual information is processed in 150 ms to begin to activate high-level category selective representations that permit a categorization response (Fabre-Thorpe et al., 2001; Goffaux et al., 2005; Ramkumar et al., 2016; Thorpe et al., 1996). For instance, representations between the pattern of errors made by neural decoders and behavioral errors correlate maximally between 100-250 milliseconds after stimulus onset (Ramkumar et al., 2016). In addition, the latency of visually evoked event related potentials (vERPs) between rapidly presented target and distractor stimuli presented in RSVP begin to

significantly diverge approximately 150 ms after scene onset (Johnson & Olshausen, 2002; Johnson & Olshausen, 2005; Thorpe et al., 1996; VanRullen & Thorpe, 2001c). This divergence is larger over frontal than occipital regions (Thorpe et al., 1996). Importantly, this divergence appears to either reflect categorization of a stimulus or post sensory decision processes rather than physical differences between categories of stimuli that are correlated with higher-level categorical decisions (Johnson & Olshausen, 2005; VanRullen & Thorpe, 2001c). In addition, there is no difference in the onset of this differential activity even when participants receive extensive training on a set of target and non-target scenes (Fabre-Thorpe et al., 2001). Potentials in response to familiar target and nontarget scenes diverge around 150 milliseconds, just as vERPs do to novel scenes. This work suggests that the visual system processes novel and familiar stimuli at a speed that cannot be compressed with experience.

Given the number of synapses between the retina and inferior temporal cortex and the latency of synaptic transmission (Thorpe, 2002; Thorpe & Fabre-Thorpe, 2001; VanRullen & Thorpe, 2001b), the above findings suggest that scene processing may primarily be supported by feed-forward mechanisms (Riesenhuber & Poggio, 1999; Serre et al., 2007; VanRullen, 2007). These feed-forward mechanisms may be so optimized that predictions for what one will see from moment-to-moment may have no impact on rapid scene categorization.[1]

In contrast, the presence of anatomical feedback connections in the brain suggests a much more complex picture, such that feed-back projections from higher-level regions may be necessary to support perception (e.g., Bullier, 2001). The visual system contains many feedback

[1] Even though very rapid and unidirectional mechanisms are often sufficient to perform complex categorization tasks, there is no reason to believe that this would signal the end of visual processing per se. Processing of a briefly flashed scene will continue after the scene is identified as more information is extracted from the input.

connections (Angelucci et al., 2002; Salin & Bullier, 1995). For example, the majority of the connections between the lateral geniculate nucleus (LGN) and V1 as well as those between V1 and V2 are bidirectional. Even at the earliest stages in processing, it appears that V2 feedback connections help shape responses in V1 (Shmuel et al., 2005). Furthermore, different research methodologies, spanning MEG, EEG, and TMS, have provided evidence that recurrent connections may influence processing in the early (Boehler et al., 2008; Camprodon et al., 2010; Foxe & Simpson, 2002) and late visual cortex (Kar & DiCarlo, 2021). Thus, the conclusion that feed-back connections may have a limited role in rapid scene gist categorization under such brief time frames is not fully supported by what we know from the anatomy of the ventral visual pathway; though it has been proposed that the ratio of feedforward and feedback connections in the visual system may serve to carry attentional signals that modulate responses in lower levels (Macknik & Martinez-Conde, 2007).[2]

In addition, the standard way of presenting scenes from trial to trial in psychophysical experiments is to randomize their order to ensure that stimuli cannot be predicted on the basis of past experiences, which lacks ecological validity.[3] Visual processing may rely predominantly on feed-forward mechanisms in the laboratory where upcoming scene categories from one fixation or from one trial to the next is unknown. Alternatively, scene gist perception performed as

---

[2] While feed-forward models do not deny the role of feed-back connectivity as processing continues over time, they propose that high-level category selective representations that enable at the very least, a crude form of recognition, emerge within a single feedforward sweep of neural activity.

[3] Researchers have presented scenes to participants in a randomized order because of the research questions that they have asked until recently (i.e., what information underlies scene gist recognition? what is the time course of rapid scene recognition?, etc.). Order effects are removed by randomizing the order of scene presentation; however, ecological validity is sacrificed by randomizing scene order. In addition, in the real world, we do not experience random sequences of scenes of different categories. We only see highly interrelated sequences of scenes, as constrained by the structure of the environment.

people navigate their environment may rely extensively on interactive mechanisms when accuracy is at a premium and constraints on response speed are not emphasized. It is possible that activity in higher order areas may generate predictions for what will be seen from moment-to-moment, and that predictions, made prior to any input from the current eye fixation, may feedback and facilitate visual processing. We ask a novel question: How long does it take the visual system to recognize a scene when the scene category is predictable compared to when the same scene is embedded within a randomized sequence of scenes?

The Scene Perception and Event Comprehension Theory

The motivation for this research question comes from the Scene Perception and Event Comprehension Theory (SPECT) (Loschky et al., 2018; Loschky et al., 2020). SPECT is an integrative framework that has been proposed to explain how people comprehend visual narratives. SPECT bridges theories of scene perception, event cognition, and narrative comprehension unlike any other model in vision science. Importantly, SPECT distinguishes *front-end* from *back-end* processes. Front-end processes are those that occur during single eye fixations, which include among other things, scene gist perception. Back-end processes are those that occur across multiple fixations in working and long-term memory. Back-end processes support the construction of the current event model in working memory (Zacks et al., 2007). Event models reflect our understanding of "what is happening now". Importantly, SPECT proposes that both front- and back-end processes support the creation of the current event model (Loschky et al., 2020). Specifically, front-end information extraction of the current spatiotemporal context (i.e., the gist of a scene) guides moment-to-moment construction of the current event model in working memory. Importantly, SPECT hypothesizes that back-end

processes, involved in constructing the current event model, feedback to front-end processes, facilitating information extraction of a scene's gist (Smith & Loschky, 2019).

Predictions may facilitate everyday scene perception

In everyday life, the scene categories we experience are not random. The knowledge we have accrued from a lifetime of experiences enables us to generate predictions about what scene categories we will experience from moment-to-moment, which reduces uncertainty (Bar, 2004; McLean et al., 2021; Smith & Loschky, 2019). Importantly, predictions facilitate categorization (Biederman et al., 1982; Kveraga et al., 2007; Lauer et al., 2018; Smith & Loschky, 2019; Summerfield & De Lange, 2014). For instance, objects within scenes that violate semantic predictions in terms of their size, location, or semantic consistency with their background or surrounding objects are detected more slowly and are harder to categorize than objects consistent with their background (Biederman et al., 1982; Boyce & Pollatsek, 1992; Davenport, 2007; Davenport & Potter, 2004; Lauer et al., 2018; Munneke et al., 2013). Analogous effects are observed in the scene categorization literature. For instance, scenes that violate schemas are harder to identify (Caddigan et al., 2017; Davenport, 2007; Greene et al., 2015) and they are harder to discriminate from random noise (Greene et al., 2015).

Importantly, generating predictions for a particular scene category by either pre-cueing it with a word (Evans et al., 2011; Kumar et al., 2021; Potter, 1976; Potter et al., 2014; Reinitz et al., 1989) or by embedding a scene within a sequence of spatiotemporally coherent scene images (McLean et al., 2021; Smith & Loschky, 2019) facilitates rapid scene categorization as well as the ability to discriminate an intact from a phase-randomized scene. An example trial from Smith and Loschky (2019) is shown in Figure 1. Smith and Loschky (2019) examined the role of sequential predictions on rapid scene categorization performance by showing scenes to

participants as first-person viewpoint image sequences. Scene sequences appeared as if an observer were traveling from one location in their environment to another (e.g., navigating from an office to a parking lot; navigating from a courtyard to a classroom, etc.). Importantly, scenes appeared in their coherent spatiotemporal order (e.g., office, hallway, stairwell, sidewalk, parking lot), consistent with how Smith and Loschky (2019) filmed the sequences to create the scenes, or scenes were shown in a randomized order (e.g., office, sidewalk, stairwell, parking lot, hallway). According to SPECT, observers should construct a richer event model when the sequence is coherent than when it is randomized (Loschky et al., 2020). Consistent with the hypothesis that back-end processes feedback to influence front-end processes, Smith and Loschky (2019) found that both prediction and categorization accuracy were greater when scenes were shown in their spatiotemporally coherent sequence (Smith & Loschky, 2019). These results have now been replicated by independent researchers using a modified methodology (McLean et al., 2021)[4].

Importantly, Smith and Loschky (2019) also found that the predictability of a scene contributed uniquely to recognition accuracy above and beyond the influence of low-level feature overlap between sequential pairs of scene images, as evident from a partial correlation analysis. Thus, in opposition to purely feed-forward accounts, which assume minimal top-down modulation of gist processing (Serre et al., 2007; VanRullen, 2007), the results of Smith & Loschky (2019) and McLean et al. (2021) suggest that predictions informed from the current

[4] McLean et al. (2021) showed participants spatiotemporally connected images that led to either expected or unexpected destinations. For instance, they showed participants a series of scenes in RSVP that appeared as if an observer were walking into a house from outside. Primes were followed by either the expected scene (e.g., the inside of the house) or an unexpected scene (e.g., the inside of a parking garage). Rapid scene categorization was better for expected scenes.

event model facilitate scene gist perception (Loschky et al., 2020). The goal of the current series of experiments is to explore *how* sequential predictions facilitate rapid scene categorization by examining *when* top-down processes influence it. Three general classes of models, motivated from the object perception literature may account for how predictions facilitate scene perception.

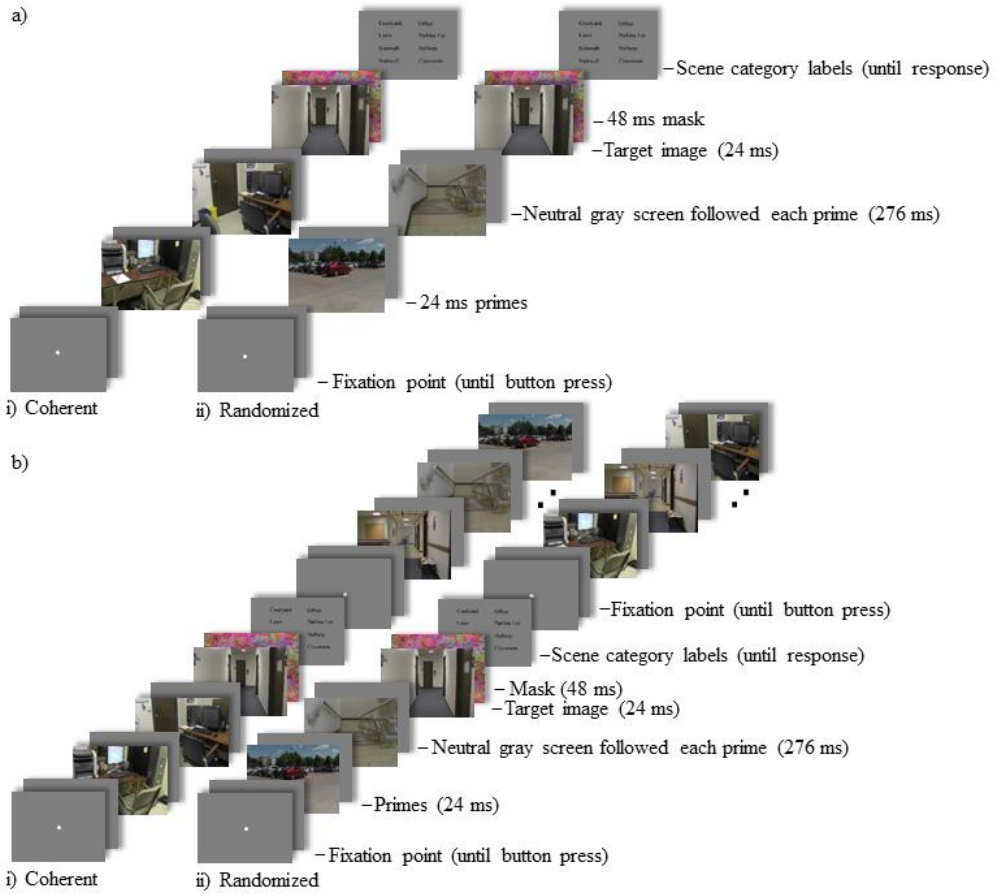


Figure 1. This is a trial schematic used in Smith & Loschky (2019). Panel a) is a simplified version of a trial. The sequence of scenes in i) are coherent. They begin in an office and end in a parking lot. The sequence of scenes in ii) are the same scenes in a randomized order. Panel b) illustrates a complete trial. After the scene was categorized, participants were shown a fixation dot, and then they pressed a button to view the remaining scenes from each sequence. A

continuation of the images were shown in all cases when the target that participants were asked to categorize from 8 alternative options was not the 10th (i.e., final) image in a sequence. In that way, participants saw a complete 10-image sequence on every trial, but never knew which item in the sequence would be tested.

Mechanisms of scene facilitation

Early facilitation accounts

Predictions for an upcoming scene may aid in the detection of perceptual features (orientation, spatial frequency, color, texture, size, etc) very early in perceptual analysis (Aitken et al., 2020; Biederman et al., 1982; Palmer, 1975b). Smith and Loschky (2019) found behavioral support for this hypothesis. To test if prediction-based facilitation was simply due to being able to guess the category of the target, Smith and Loschky (2019) had participants performed a two alternative forced choice task where they indicated if the target on each trial was intact or phase-scrambled (Greene et al., 2015). Smith and Loschky (2019) found that sensitivity was greater for targets shown in coherent sequences, which suggests that the facilitation was perceptual/sensory in nature, and clearly ***not*** due to guessing at the point of making a response.

Facilitation from the event model may even occur as early as V1 (Aitken et al., 2020; Edwards et al., 2017; Muckli et al., 2015; Muckli & Petro, 2013; Petro et al., 2014). For instance, Muckli et al. (2015) presented scenes to participants either in full, or with parts of the scene occluded. Taking advantage of the retinotopic layout of V1, they found that neural decoders could successfully decode functional information from a scene in V1 even when it was not receiving direct feed-forward stimulation (i.e., when parts of the scene were occluded). This suggests that information at higher, more abstract levels, feedback to V1 even when V1 is not

directly stimulated from the external world. Furthermore, Muckli and Petro (2013) proposed that predicted representations of the occluded image are more lowpass than the actual stimulus. This is because higher-level cells that provide feedback to V1 have larger receptive fields than those in V1. Thus, perceptual predictions informed by the current event model may "fill-in" missing or inferred scene information early in the time course of perceptual analysis.

Sequential predictions may alternatively facilitate the perception of scene layout, which is thought to be processed in the parahippocampal place area (Epstein & Kanwisher, 1998; Epstein & Baker, 2019). Prior work found that reaction time to identify which of two objects within a scene are closer in pictorial space is faster for scene layouts that are primed by an identical scene compared to scenes that are either not primed at all, or are primed by a different scene category (Sanocki, 2003; Sanocki & Epstein, 1997). Such facilitation may be due to shared low-level information between primes and the target image held in sensory memory (Shafer-Skelton & Brady, 2019), the sudden onset of the critical objects in the target scene (Germeys & d'Ydewalle, 2001), or the maintenance of scene layout information in working memory (Sanocki, 2003). Consistent with the latter explanation for scene layout priming, prior work has also demonstrated scene layout priming between scenes with different lighting directions or two different views of the same scene (Castelhano & Pollatsek, 2010; Sanocki, 2003). Thus, predictions may facilitate perception of scene layout, possibly in the parahippocampal place area.

Matching process accounts

According to a second class of theories, predictions informed by the current event model, may reduce the computational load of the matching process after or in parallel to the construction of a scene's *structural description* - a high-level *visual representation* of the shape and structure of a scene. Namely, predictions may facilitate the process of matching the scene's structural

representation to a representation in long-term memory. According to matching process accounts, semantic predictions facilitate identification by limiting the number of candidate scene category representations to match the visual input (Bar, 2004; Bar & Ullman, 1996; Friedman, 1979; Leroy et al., 2020; Mudrik et al., 2010; Palmer, 1975a; Schendan, 2019; Trapp & Bar, 2015). According to this class of theories, early perceptual processes analyze the input and transform it into a structural description (Humphreys et al., 1999; Kosslyn et al., 1994; Palmer, 1975b). This early perceptual stage may be impervious to top-down influences (Firestone & Scholl, 2016). Subsequent processes match the structural description to representations stored in semantic memory, producing entry-level categorization. This matching process may be subserved by inferior temporal regions important for rapid scene categorization (Epstein & Baker, 2019; Park et al., 2011; Walther et al., 2009), and frontal regions, which, after training, also contain category selective neurons (Freedman et al., 2001, 2003). If one conceptualizes this matching process as a selection process where the visual system scans multiple categorical representations to find a good match for the visual input, then predictions could facilitate scene perception by reducing the size of this search space, perhaps by pre-activating candidate scene category representations of to-be-presented scenes. For instance, if one's current event model enables the prediction that a hallway will lie on the other side of an office door, then a prediction could prioritize those semantic representations for comparison over representations for sidewalks, parking lots, forests, or bathrooms, which are all less likely to appear. Only the initial guess needs to be considered as a likely interpretation of the input. This is not to say that all possible interpretations from memory need to be considered when matching a structural description to a representation in semantic memory. The feed-forward sensory signal itself likely constrains the options to be considered for further analysis. The matching account contrasts with

early facilitation accounts. Early facilitation accounts propose that feed-back mechanisms act directly on the integration of local features that form scenes, and thus, the construction of the scene's structural description (Biederman et al., 1982; Palmer, 1975b).

From a computational perspective, scene identification terminates with a successful match, which prior research suggests between 150 to 250 milliseconds after scene onset (e.g., VanRullen & Thorpe, 2001c). There is debate as to whether the matching process is primarily perceptual, or cognitive, in nature. The fact that Sensation & Perception textbooks usually devote an entire chapter to recognition indicates that researchers in the field have long considered the matching process to be perceptual. However, this assumption has been criticized (Block, 2008; Burge, 2014). Block (2008) argued that everything before the matching process is perceptual, since those processes involve operations on iconic representations, but that the matching process involves categorization (i.e., "this [x] is [p]"), so it is cognitive. A separate but related argument by Firestone and Scholl (2016) is that the matching process undoubtedly involves memory; therefore, matching should be considered a cognitive rather than a perceptual phenomenon. Nevertheless, a more recent counter-argument by Mandelbaum (2018) is that matching, though categorical, is nevertheless perceptual, because the rapidity of the entire set of feed-forward processes, from stimulus onset to matching, is too fast to involve feedback from higher cognitive areas. Cermeño-Aínsa (2021) argues further that perception must include the matching stage, because the time course of ventral stream brain activity, decoded based on participants' categorization responses, completely overlaps with that decoded based on mid-level visual features (Ramkumar et al., 2016). We consider matching to be part of the perceptual process. As such, the processes involved ***after*** finding a match are typically referred to as *post-*

perceptual (Hollingworth & Henderson, 1998, 1999) though visual processing of a scene does not actually terminate after recognition (Malcolm et al., 2016).

Post-identification accounts

A third possibility is that sequential predictions facilitate even later processes such as those involved in semantic knowledge activation (Ganis & Kutas, 2003), semantic integration (Hagoort et al., 2009; Sitnikova et al., 2008), or processes involved in response selection (Hollingworth & Henderson, 1998, 1999). According to this account, feed-forward visual analysis is sufficient to discriminate one scene category from another. The improved versus impaired performance observed when categorizing scenes presented in coherent and randomized sequences (Smith & Loschky, 2019) may stem from processing how the already identified current scene is consistent versus inconsistent with information that came before it. For example, if you enter your friends' house expecting to see a living room, but instead you walk into your friends' office, then the ability to assign the office to its appropriate category may be impaired because of the difficulty in *integrating* the already identified office information with the previous information of your friend's sidewalk and doorstep rather than by impairing your ability to *see* or *categorize* to the scene.

Current Experiment

It is difficult to infer the time course of scene processing and how back-end processes involved in constructing the current event model influence front-end processes using behavioral measures alone. Behavioral responses reflect the downstream effects of the experimental manipulation from the earliest perceptual stages to the motor response. Behavioral basic level scene categorization could reflect facilitation at any stage in the visual processing hierarchy. Therefore, we choose to examine the time course of predictions' effects on scene perception

using a noninvasive brain activity measure with very high temporal precision. Here, using electroencephalography (EEG), we compared visually evoked event-related potentials (vERPs) to infer the stage in perceptual processing when predictions affect scene perception. Visually evoked potentials enable continuous monitoring of neural activity elicited by experimental manipulations, and have provided crucial information regarding the time course of scene, language, and object processing (Greene & Hansen, 2018; Harel et al., 2016; Harel et al., 2020; Kutas & Federmeier, 2011; Thorpe et al., 1996). To assess when in the time course of scene processing predictions begin to facilitate scene gist perception, we compared vERPs elicited by each target scene ($target_{i,j}$) when it was shown in either a spatiotemporally coherent ($target_{i,c}$) or randomized ($target_{i,r}$) sequence. In such an analysis, the timing of the vERP differences between the two conditions provides an estimate of the time when neural representations involved in predicting an upcoming scene begin to interact with identification processes. It is generally agreed upon that the first 150 milliseconds of cortical processing in response to a scene reflects brain activity that is driven by stimulus features, and activity after 150 milliseconds reflects categorization of the visual input or task specific activity (Johnson & Olshausen, 2002; Johnson & Olshausen, 2005; VanRullen & Thorpe, 2001c). As such, early differences support early facilitation accounts, and late differences support matching or post-identification accounts.

VERP Components of Interest

We examined two different vERP components (P200 and N400) to determine *when* in the time course of scene processing the current event model facilitates scene perception. The P200 is sensitive to scene category-specific information (Harel et al., 2016) and there is evidence to suggest it is more positive when we experience an unexpected scene category after several leading images to a destination (McLean et al., 2021). The N400 is associated with semantic

access and semantic integration processes (Hagoort et al., 2009; Kutas & Federmeier, 2011). It is more negative when an observer encounters a scene that is inconsistent with a pre-cue (Kumar et al., 2021), an unexpected scene in a picture story (Cohn & Kutas, 2015; Sitnikova et al., 2008; West & Holcomb, 2002), or when an observer encounters an unexpected object within a scene (Mudrik et al., 2010; Truman & Mudrik, 2018; Võ & Wolfe, 2013).

The P200.

The P200 is a positive neural component arising 150 to 249 ms post-stimulus, peaking around 220 milliseconds after scene onset. The P200 reflects scene-specific processing, as it is the first vERP component that responds more strongly to scenes than to other categories of objects (e.g., faces) (Harel et al., 2016). Importantly, the P200 varies with a scene's openness, relative distance, and naturalness (Hansen et al., 2018; Harel et al., 2016; Harel et al., 2020; Lowe et al., 2018). It likely reflects Gestalt perceptual grouping processes (Schendan & Kutas, 2007; Schendan & Lucia, 2010) as it becomes smaller with better grouping (Halgren et al., 2003; Han et al., 2005). Unlike later components, there is evidence to suggest the P200 is sensitive to global scene properties such as a scene's openness or naturalness, regardless of whether participants are instructed to attend those aspects of a scene or not (Hansen et al., 2018). Thus, the P200 may not be influenced by top-down factors, such as an observer's goals. Alternatively, McLean et al. (2021) demonstrated that the P200 is sensitive to scene predictability. McLean et al. (2021) found that unexpected scenes were categorized more poorly than expected scenes and they elicited a more positive P200 than scenes that were expected given a series of primes that lead up to the target. According to purely feed-forward accounts of visual processing (e.g., Serre et al., 2007; VanRullen, 2007; VanRullen & Thorpe, 2001b), scene recognition is alleged to occur during this time window (see also Ramkumar et al., 2016 who found that correlations

between the performance of neural decoders and human behavior peak between 150 and 250 milliseconds). Thus, the P200 could be a component that indexes recognition processes in the front-end (i.e., Loschky et al., 2020).

The N400.

The N400 component has previously been used to investigate semantic processing. The N400 is a negative going waveform appearing roughly 250-500 milliseconds after the onset of a meaningful stimulus over central scalp regions (Kutas & Federmeier, 2000, 2011); though it is more frontally distributed when observer's view pictures (Ganis et al., 1996). The amplitude of the N400 is more negative when one encounters a semantic violation or a violation of a prediction. For instance, the word *socks* in the following sentence elicits a larger N400 than a word that makes sense within the sentence: "He spread his warm bread with *socks*" (Kutas & Hillyard, 1980). Importantly, the N400 is also more negative when participants experience an unpredictable panel in a comic (Coderre et al., 2020; Cohn & Kutas, 2015), an unexpected action in a short video clip (an iron instead of a knife was used to cut bread) (Sitnikova et al., 2008), a violation of world knowledge (Hagoort et al., 2004), or an unexpected scene within a sequence of lead-up scenes to a destination location (McLean et al., 2021).

The N400 is a late, possibly a post-perceptual, component within the processing stream. The N400 is associated with the integration of meaning from linguistic and nonlinguistic sources. It is more negative when someone attempts to integrate semantic information accessed from the current word, comic strip, object, or scene with meaningful information from the preceding context (Demiral et al., 2012; Hagoort, 2007; Hagoort et al., 2009; Kutas & Hillyard, 1983; Sereno et al., 1998). Thus, it appears to reflect processing in an amodal semantic system (see Kutas & Federmeier, 2011 for a review). It may index the ease of mapping information onto the

event model in working memory (Cohn et al., 2012; Gernsbacher, 1990; Loschky et al., 2020)[5]. Afterall, its time course is later than the 150-249 milliseconds required to extract the gist of the scene, and it seems to reflect processing that occurs over multiple items in working memory.

Semantic violations are additionally associated with an earlier negative component, known as the N300 (Hamm et al., 2002; Holcomb & Mcpherson, 1994; Mudrik et al., 2010; Schendan & Kutas, 2002; Smith & Federmeier, 2020; Võ & Wolfe, 2013). However, questions remain as to whether the N300 and N400 reflect the same or distinct psychological processes (Draschkow et al., 2018; Truman & Mudrik, 2018; Willems et al., 2008). Both are sensitive to similar manipulations and they have similar scalp distributions. Unfortunately, our current design cannot verify if the N300 and N400 are separate neural components. Furthermore, we did not find major differences between the two components when we analyzed them separately (See the Appendix) consistent with prior work (Draschkow et al., 2018; Federmeier & Kutas, 2001; Kumar et al., 2021; Lauer et al., 2018; Mudrik et al., 2010; Võ & Wolfe, 2013; Willems et al., 2008). As such, we will report results, and focus conclusions on the N400 though details about the N300 are provided in the Appendix.

Hypotheses

Experiment 1 is a behavioral experiment, and it was conducted to demonstrate that scenes shown in the coherent sequences were more predictable than scenes shown in randomized

[5] *Matching* the structural description of a scene to a representation in semantic memory, and *mapping* processes are distinct. Matching is typically considered a perceptual phenomenon occurring within single eye fixations (Kumar, Federmier, & Beck, 2021, Schendan & Kutas, 2003; Schendan & Maher, 2008; Smith & Federmieier, 2020; Bar, 2004) whereas mapping information onto the developing event model is considered to be a conceptual phenomenon, possibly occurring after or in parallel to recognition across multiple fixations (Gernsbacher, 1990; Loschky et al., 2020).

sequences. This was previously shown in Smith and Loschky (2019), but with a smaller subset of stimuli. Thus, it was both useful to replicate the finding from the original study, and as importantly, to test the hypothesis using the current expanded stimulus set. Experiment 2 added the measurement of vERPs to determine *when* in the time course of scene gist processing the current event model facilitates information extraction in the front-end. Experiment 3 replicated Experiment 2 with a slightly modified paradigm in which we flashed images at a faster rate and masked the target. We made the changes from Experiment 2 to 3 to decrease participants' rapid scene categorization performance so that we could potentially extract more information from their behavioral confusion matrices (i.e., more off-diagonal cell entries), which we then correlated with the responses of neural decoders.

By examining the aforementioned vERP components, we can determine if top-down predictions, informed by the contents of the current event model facilitate scene perception, or only post-perceptual processes involved in integrating information into the event model. If differential vERP waveforms between coherent and randomized sequences of scenes appear early (i.e., 0-149 ms),[6] then this would be consistent with early facilitation accounts (Biederman et al., 1982; Muckli et al., 2015; Palmer, 1975b). Early differences could also arise from visual similarity between successive scenes that share visual features (Shafer-Skelton & Brady, 2019; Sperber et al., 1979). We will explore this alternative explanation for our results in an exploratory analysis below. If differences arise later (i.e., 150-249), this could support matching accounts of facilitation (Bar, 2004; Bar & Ullman, 1996; Friedman, 1979; Mudrik et al., 2010;

[6] This component could correspond to either an N1 or P1. To our knowledge, none of the work that has investigated object-scene consistency effects or the effects of consistent and inconsistent scenes (Mclean et al., 2021) has demonstrated differential activity between expected compared to unexpected stimuli in this early component.

Trapp & Bar, 2015). Alternatively, if the differences have an even later time course (250-449 ms), then violations of predictions may reflect computations during post-perceptual processing of the scene, potentially reflecting difficulties in semantic integration of the current scene within the event model (Cohn & Kutas, 2015; Demiral et al., 2012; Hagoort et al., 2009; Kutas & Federmeier, 2000). Differences arising within an intermediate window between the P200 and N400 (250-349) could reflect facilitation in the N300, which has also previously been argued to reflect categorization (Hamm et al., 2002; Schendan, 2019; Smith & Federmeier, 2020). While these are alternative hypotheses, they are not mutually exclusive. Top-down effects may be observed at all levels of scene processing.

In Experiments 2 and 3, we complemented our analyses of the vERPs with a machine learning approach to assess how categorical scene information emerges over time. Previous research has found that linear support vector machines (SVMs) can be used to identify a scene's category from participants' M/EEG and BOLD responses (Greene & Hansen, 2020; Ramkumar et al., 2016; Torralbo et al., 2013; Walther et al., 2009; Walther et al., 2011; and many others). Performance of time-resolved neural decoders typically peaks between 150 and 250 milliseconds after scene onset (Cichy et al., 2017; Greene & Hansen, 2020; Ramkumar et al., 2016). Importantly, patterns of responses made by neural decoders often mirrors the patterns of responses made by human observers (Ramkumar et al., 2016; Torralbo et al., 2013; Walther et al., 2009). By comparing decoding accuracy for scenes shown coherent and randomized sequences, we can determine if scene representations are richer when scenes are shown in coherent sequences. Early facilitation accounts propose that predictions facilitate the construction of a scene's structured visual description (Biederman et al., 1982; Palmer, 1975b). As such, divergence in decoding accuracy between coherent and randomized sequences should

occur early (0-150 ms). In contrast, matching accounts propose that predictions facilitate scene perception by limiting the number of categories to compare to the visual description (Bar, 2004; Bar & Ullman, 1996; Trapp & Bar, 2015). As such, divergence in decoding accuracy between coherent and randomized sequences should occur much later (150-249 ms).

One of the drawbacks with pattern classification techniques is that there is no guarantee that the algorithm used to decode the scene category information from the neural activity is using the same information that humans use to categorize a scene. From decoding accuracy alone, we do not know if a particular pattern is contributing to a participant's ability to categorize a scene. As such, we correlated confusions made by the pattern classification model trained to identify scene categories from EEG data with behavioral confusions made by participants across time. Analogous approaches have been used in both fMRI and MEG studies to examine what information humans use to categorize scenes (Torralbo et al., 2013; Walther & Shen, 2014) and its time course (Ramkumar et al., 2016). Critically, by manipulating a scene's spatiotemporal coherence, we can investigate if scenes presented in coherent sequences are recognized by the visual system more efficiently than scenes shown in randomized sequences. We did this by comparing correlations between coherent and randomized sequences.

Chapter 2 - Experiment 1

We first evaluated if scenes shown in coherent sequences were more predictable than scenes shown in randomized sequences. Smith and Loschky (2019) found evidence to suggest that they were; however, the current experiment involved 3 times the number of scenes, so the experiment was conducted to validate the new stimuli and replicate previous work. On each trial, participants viewed a series of 10 scenes, shown in either a spatiotemporally coherent or randomized sequence. Instead of categorizing one of the scenes as we later did in Experiments 2 and 3, participants predicted the category of one scene on each trial. The target that participants predicted varied in its temporal position within the sequence randomly across trials. One possibility, according to SPECT, is that the presentation of the first scene in a coherent sequence lays the foundation for the current event model in working memory (Loschky et al., 2020). Upcoming scenes become more predictable after the event model is constructed. This contrasts with having no event model for upcoming scene categories when scenes are presented in a randomized order. As such, we hypothesized that the ability to predict scene categories would be greater when the sequence is coherent than when it is randomized. In addition, if images in randomized sequences are less predictable, they should be predicted at no better than chance level, since viewers should be unable to construct coherent event models, and thus should be unable to predict upcoming scene categories.

Lastly, we also hypothesized that the ability to predict upcoming scene categories would increase as a function of the ordinal position of the scene on a trial. The event model should develop as the number of images shown on a trial increases. When the sequence is coherent, predictions should be much more accurate when participants are asked to predict the 10^{th} image on a trial than the 2^{nd}. Conversely, when the sequence is random, the ability to predict the scene

categories should not increase as a function of the ordinal position of the scenes on a trial since viewers should be unable to construct a coherent event model of the random sequences.

Method

Participants

One-hundred thirty-nine students (N = 93 females, N = 46 males) participated in Experiment 1 for course credit. Age of the participants ranged from 18 to 24 (M age = 19). Data from one participant was removed for not completing the experiment. The experiment was approved by the institutional review board at Kansas State University, and each participant signed an informed consent form electronically before participating.

To estimate the number of participants needed to find an effect of spatiotemporal coherence, we conducted a power analysis using data from Experiment 2 of Smith and Loschky (2019). The response variable was prediction accuracy, coded as a 0 or 1. General linear modeling-based approaches are insufficient for estimating an effect size for repeated measures binary response variables (Jaeger, 2008; Kumle et al., 2021). Thus, we used a method for conducting a power analysis of generalized linear mixed models (Green & MacLeod, 2016; Kumle et al., 2021). We conducted the analysis using the MixedPower package in R (Kumle et al., 2021). Power calculations using this method are based on Monte Carlo simulations. The power analysis begins by first fitting a statistical model to the data. Power is calculated by repeating three steps. First, new values for the response variable are simulated using the data provided. Second, the model refits to the simulated responses. Third, a statistical test is applied to the simulated data. In this step, the tested effect is ‘known to exist’, so every positive test is considered a true positive and every negative test is deemed a Type II error. Power is calculated

from the number of successes and the number of failures in the third step. Each of the three steps are then repeated for a given sample size that we specify.

The independent variable was spatiotemporal coherence. We sampled the effect with 1,000 iterations for 5 to 45 participants. Based on a two tailed hypothesis, with the effect size of d = .34, alpha = .05 and power of .90, we found that a total sample size of 20 would be needed to observe a behavioral difference in prediction performance between the coherent and randomized scene sequences. Participants in the experiment saw one of 3 different versions of the stimuli. Half of the participants saw the coherent sequences first and half saw the randomized sequences first. Given these additions in the experiment, we multiplied 20 by 6 to get the target sample size of 120 participants. We collected more data than we expected due to constraints in scheduling participants.

Materials

We collected a total of 1,440 photographs from Kansas State University (i.e., on-campus) and the local Manhattan, Kansas metropolitan area (i.e., off-campus). Half of the images were on-campus and half were taken off-campus. One-third of the total 1,440 images were used in our previous investigations of scene categorization (Smith & Loschky, 2019). The on-campus images were composed of four indoor scene categories (office, classroom, hallway, and stairwell) and four outdoor scene categories (parking lot, courtyard, sidewalk, and lawn). Two of the indoor scene categories (office and classroom) and two of the outdoor scene categories (parking lot and courtyard) were starting locations and destinations. These were locations observers appeared to be navigating to or from in their environment (e.g., going from an office to a parking lot) as determined from the actual locations of the scenes and how they were connected in the world. The remaining four categories were transitional scene categories between starting

locations and destinations (e.g., locations participants appear to be navigating through to get from a starting location to a destination). We had the same number of off-campus categories as on-campus categories. Four were indoor (store interior, bedroom, stairwell, and hallway) and four were outdoor (park, city center, sidewalk, and alley). Store interior, bedroom, park, and city center were starting locations and destinations. The remaining categories were transitional categories between destinations. We took special care when creating the stimuli to ensure that participants were unable to see the upcoming scene category from the currently viewed scene in the sequences (e.g., the door that separated the office from the hallway was shut).

Spatiotemporally coherent image sequences were photographed from a first-person viewpoint so that observers appeared to be navigating from one location in their environment to another (i.e., going from an office to a parking lot). We created forward versions of each coherent scene sequence (e.g., office to a parking lot) and their reverse (i.e., a parking lot to an office) from each starting location to each destination, but not along the same pathway (i.e., the office in the 'office to a parking lot' sequence was a different office than the one used in the 'parking lot to an office' sequence).

See Table 1 for a table containing all the sequences. There were 24 [(4 starting/destination locations choose 2) X 2 directions (forward or reverse) X 2 locations (on-campus and off-campus)] different base sequences. We created 3 versions of each base sequence (e.g., 3 different 'office to parking lot' sequences) for a total of 72 total sequences (24 base sequences X 3 versions of each sequence) to increase the number of exemplars per category. The versions were of the same base sequences (See Table 1), but with different scene exemplars. Each base sequence was composed of 20 different images, which were made up of 5 scene categories (e.g., office, hallway, stairwell, sidewalk parking lot) with 4 images per category (e.g.,

4 offices, 4 hallways, 4 stairwells, 4 sidewalks, 4 parking lots). An example full 20 image sequence is shown in Figure 2. Of the 20 total scenes in each base sequence, participants saw 10 scenes per trial. Participants viewed each base sequence twice (e.g., they appeared to be walking from the same office to the same parking lot twice).

We wanted our coherent sequences to be predictable based on participants' knowledge of the world, but not from artifacts due to the way we constructed our sequences. Thus, we created subsequences in which we randomly chose to show one to three images from each category before a category shift (e.g., three offices, one hallway, two stairwells, etc.). Specifically, our goal was to reduce the possibility of guessing when there would be a shift to a new category. If we were to always show the same number of scenes from each category before a transition, then a viewer would be able to accurately guess when a change to a new scene category would occur. By showing subsequences of 1-3 images from each category, this ensured that 1) participants saw each category of a base sequence on each trial, and 2) they were unable to guess when transitions would occur from the nature of the design of the experiment.

Given that we collected a total of 1,440 images (24 base sequences X 3 versions X 20 images per base sequence), we assigned participants to predict images in one of the three versions of each base sequences. Participants were randomly assigned to the version. Each participant viewed 960 scenes (480 in the coherent and 480 in the randomized condition) for 96 trials. This was done to keep the experiment length under one hour. For each participant, the same images were shown in the coherent and randomized conditions.[7]

[7] This is different from what was done in Smith and Loschky (2019). Smith and Loschky (2019) randomly assigned 10 of the 20 images within each base sequence to the coherent condition and the remaining 10 to the randomized condition.

Table 1. *Table of base sequences. Each base sequence begins with a starting location and terminates in a destination location. Observers appear to be navigating through the transitional categories when navigating between starting and destination locations. There were 3 versions of each base sequence showing the same starting and destination categories, but from different locations.*

Base Sequence Number	Starting Location	Destination	Transitional Categories	Location
1	Office	Parking Lot	Hallway, Stairwell, Sidewalk	On-Campus
2	Parking Lot	Classroom	Sidewalk, Hallway, Stairwell	On-Campus
3	Classroom	Courtyard	Hallway, Stairwell, Sidewalk	On-Campus
4	Office	Classroom	Hallway, Lawn, Hallway	On-Campus
5	Office	Courtyard	Hallway, Stairwell, Sidewalk	On-Campus
6	Courtyard	Parking Lot	Lawn, Hallway, Stairwell	On-Campus
7	Parking Lot	Office	Sidewalk, Stairwell, Hallway	On-Campus
8	Classroom	Parking Lot	Hallway, Sidewalk, Lawn	On-Campus
9	Courtyard	Classroom	Sidewalk, Lawn, Hallway	On-Campus
10	Classroom	Office	Hallway, Lawn, Sidewalk	On-Campus
11	Parking Lot	Courtyard	Sidewalk, Hallway, Stairwell	On-Campus
12	Courtyard	Office	Sidewalk, Stairwell, Hallway	On-Campus
13	Bedroom	City Center	Hallway, Stairwell, Sidewalk	Off-Campus
14	Bedroom	Park	Hallway, Stairwell, Sidewalk	Off-Campus

15	City Center	Bedroom	Sidewalk, Alley, Stairwell	Off-Campus
16	Park	Bedroom	Sidewalk, Alley, Stairwell	Off-Campus
17	Store Interior	City Center	Alley, Sidewalk, Alley	Off-Campus
18	Store Interior	Park	Sidewalk, Alley, Sidewalk	Off-Campus
19	City Center	Store Interior	Sidewalk, Alley, Stairwell	Off-Campus
20	Park	Store Interior	Sidewalk, Alley, Hallway	Off-Campus
21	Bedroom	Store Interior	Stairwell, Sidewalk, Alley	Off-Campus
22	Store Interior	Bedroom	Alley, Stairwell, Hallway	Off-Campus
23	Park	City Center	Alley, Stairwell, Sidewalk	Off-Campus
24	City Center	Park	Sidewalk, Alley, Sidewalk	Off-Campus

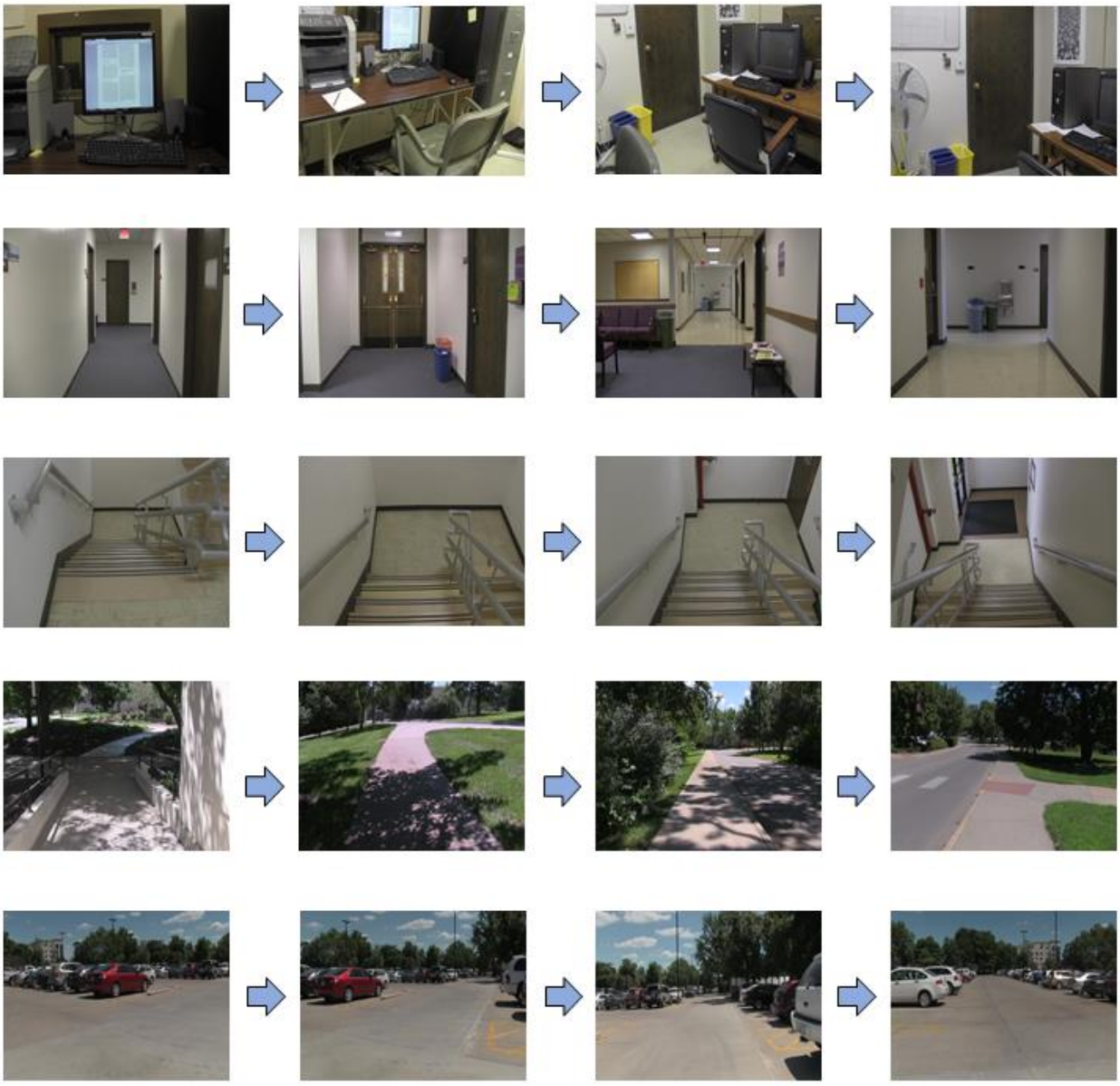


Figure 2. Example of a full 20 image base sequence for "Office to Parking Lot." There were 24 such base sequences, with 3 different exemplar base sequences for each, for a total of 72 such sequences (see Table 1 for full list). Each base sequence is from a first-person viewpoint, and is shown here in the spatiotemporally coherent order. Participants saw a subset of 10 images from

such a base sequence on each trial. The images shown within each subset were randomly determined for each participant.

Procedure

The experiment was programmed with custom software written in Python with the PsychoPy3 (Version 2020.2.2) libraries (Peirce & MacAskill, 2018). That was then, converted to jsPsych, and hosted on Pavlovia (https://pavlovia.org/). Participants completed the experiment online.

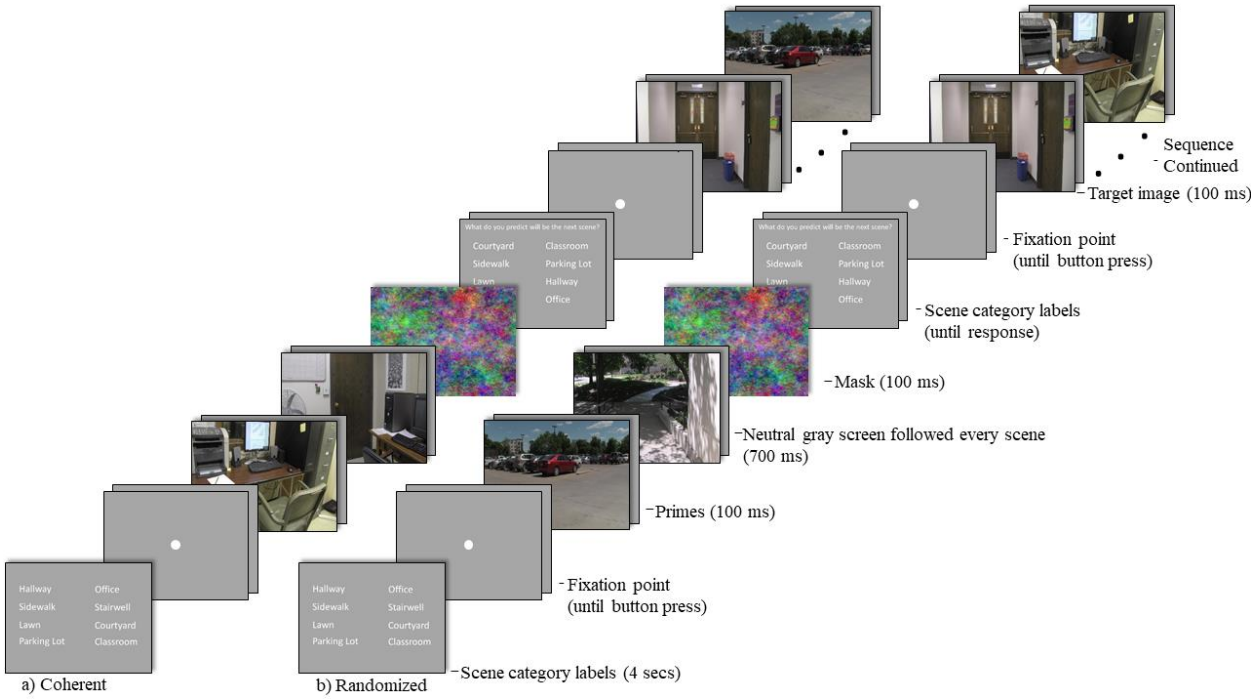


*Figure 3***.** This is a trial schematic. Scenes were shown in either a a) coherent or b) randomized sequence. Scenes in the coherent sequence represented here, begin in an office and end in a parking lot. Instead of pausing the sequence when the target was presented (as in Figure 1), the target was absent, so the sequence paused after presentation of a 700 ms blank screen followed by a 100 ms noise mask. The mask thus served as a cue to participants to predict what the next

scene would be. After participants selected their prediction from the 8-AFC response screen, they were shown a fixation dot, and they pressed a button to view the target they were asked to predict and the remaining scenes from each sequence. A continuation of the images were shown in all cases when the target that participants were asked to predict was not the 10^{th} (i.e., final) scene.

The trial schematic is shown in Figure 3. Prior to the beginning of each trial, participants saw a list of the eight category labels for the scenes in that sequence, listed in a randomized order, within a 4 X 2 grid, for 4 seconds. Participants then saw a fixation cross and were instructed to press a button on the keyboard while they fixated the cross to begin the trial. A series of 11 images (10 scenes and 1 noise mask) were each flashed for approximately 100 milliseconds (based on a 60 Hz monitor refresh rate), with each image followed by a 700 millisecond neutral gray screen (i.e., there was an 800 millisecond SOA). This was done to approximate the same processing time for each image in Experiment 2, which will be described later. Because it was an online study, we did not have control of participants' monitor refresh rate or their internet speed, which could have affected the SOAs. Each trial consisted of 11 images, 10 of which were real-world scenes, and one of which was a 1/f noise mask that preceded the target. The image sequence was shown until the 1/f noise mask appeared, and then participants were prompted to "*predict what the next image will be"* from among 8 scene category labels. The onset of the mask was not included to impair visual processing of the previous scene. Instead, it was included to alert participants for when they would need to make a prediction. In order to avoid contaminating the results by a bias toward responding to a favored location (e.g., always clicking the top left corner), the locations of the labels in the 8AFC array

changed randomly on each trial, even though this does make the task more difficult for the participants because they have to search for the target category label on each trial. Further, we counterbalanced the temporal location (2 – 10) of the target scene on each trial across trials and participants. Trial order was randomized for each participant. This allowed us to investigate prediction accuracy across time within a trial. After making their response, participants again fixated the cross, pressed the button, and continued to view the remaining images within the trial.

Participants performed the task using scenes from one of the three versions of base sequences. Each base sequence was composed of 480 scenes (24 base sequences X 20 images). Each trial consisted of 10 images. Participants completed 48 trials in the coherent and 48 in the randomized conditions. We blocked and counterbalanced the manipulation of spatiotemporal coherence across participants consistent with a previous exploration of a similar research question (Smith & Loschky, 2019)[8]. Half of the participants viewed the scenes in the coherent condition before viewing them in the randomized condition.

Importantly, we matched the specific image participants were asked to predict and its temporal location (2-10) within the trial across the coherent and randomized versions of each sequence. In this way, we could compare prediction accuracy, categorization accuracy, and vERPs in Experiments 2 and 3, for the *same* image, at the *same* temporal position, within each sequence.

Results

Full study results overview

[8] We blocked the sequences so that participants could set up the expectation that sequences they were viewing were consistently predictable or not.

We report many analyses throughout the results sections of Experiments 1, 2, and 3, and the large number of results and hypotheses may be difficult to keep track of. Thus, to aid in the interpretation of the results, we created a master summary table (Table 2) to convey the main results of each analysis in a single location for reference. To briefly summarize the results, analyses of prediction and categorization performance replicated the work of Smith and Loschky (2019). Analyses of the event related potentials supported matching and post-identification accounts of facilitation. The multivariate pattern classification analysis supported matching and post-identification accounts of facilitation and, more importantly, early facilitation accounts.

Table 2**.** *Summary of the results of Experiments 1, 2, and 3.*

Experiment	Dependent variable	Region	Window	Independent Variable(s)	Effect	Hypothesis Supported
Experiment 1: Prediction Experiment	Prediction Accuracy			Spatiotemporal Coherence, Location	Coherent > Randomized	SPECT
				Spatiotemporal Coherence, Location, Ordinal Position of Target	Ordinal position positive slope (off campus sequences)	SPECT
Experiment 2: EEG Experiment (Without Masking)	Categorization Accuracy			Spatiotemporal Coherence, Location	Coherent > Randomized	SPECT
				Spatiotemporal Coherence, Location, Ordinal Position of Target	Ordinal position positive slope	SPECT
	vERP of Target					
		Frontal & Central	50-149	Region, Location, Spatiotemporal Coherence	Coherent = Randomized	Feedforward
		Frontal & Central	150-249	Region, Location, Spatiotemporal Coherence	Coherent > Randomized	Matching
		Frontal & Central	250-449	Region, Location, Spatiotemporal Coherence	Coherent > Randomized	Post-identification
		Parietal/ Occipital	50-149	Region, Location, Spatiotemporal Coherence	Coherent = Randomized	Feedforward

	Parietal/ Occipital	150-249	Region, Location, Spatiotemporal Coherence	Coherent = Randomized	Feedforward
vERP of all scenes					
	Frontal & Central	50-149	Region, Location, Spatiotemporal Coherence	Coherent = Randomized	Feedforward
	Frontal & Central	150-249	Region, Location, Spatiotemporal Coherence	Coherent > Randomized	Matching
	Frontal & Central	250-449	Region, Location, Spatiotemporal Coherence	Coherent > Randomized	Post-identification
	Parietal/ Occipital	50-149	Region, Location, Spatiotemporal Coherence	Coherent = Randomized	Feedforward
	Parietal/ Occipital	150-249	Region, Location, Spatiotemporal Coherence	Coherent = Randomized	Feedforward
vERP Divergence					
	Frontal		Spatiotemporal Coherence		Matching, Post-identification
	Central		Spatiotemporal Coherence		Matching, Post-identification
	Parietal/ Occipital		Spatiotemporal Coherence		Feedforward
vERPs within a trial					
	Frontal & Central	50-149	Region, Location,	Coherent = Randomized	Feedforward

	Frontal & Central	150-249	Spatiotemporal Coherence Region, Location, Spatiotemporal Coherence	Coherent > Randomized	Matching
	Frontal & Central	250-449	Region, Location, Spatiotemporal Coherence	Coherent > Randomized	Post-identification
	Parietal/ Occipital	50-149	Region, Location, Spatiotemporal Coherence	Coherent = Randomized	Feedforward
	Parietal/ Occipital	150-249	Region, Location, Spatiotemporal Coherence	Coherent < Randomized	Matching
vERP of Target					
	Frontal		Spatiotemporal Coherence, Image Predictability	Positive Correlation between Amplitude and Predictability	Matching
	Central		Spatiotemporal Coherence, Image Predictability	Positive Correlation between Amplitude and Predictability	Matching
	Parietal/ Occipital		Spatiotemporal Coherence, Image Predictability	Negative Correlation between Amplitude and Predictability	Matching
Decoding Accuracy					
	Frontal		Spatiotemporal Coherence	Coherent > Randomized	Matching, Post-identification

		Central		Spatiotemporal Coherence	Coherent > Randomized	Matching, Post-identification
		Parietal/ Occipital		Spatiotemporal Coherence	Coherent > Randomized	Early, Matching, Post-identification
	Correlations between model and human responses					
		Frontal		Spatiotemporal Coherence	Coherent > Randomized	Matching, Post-identification
		Central		Spatiotemporal Coherence	Coherent > Randomized	Matching, Post-identification
		Parietal/ Occipital		Spatiotemporal Coherence	Coherent > Randomized	Early, Matching, Post-identification
Experiment 3: EEG Experiment (With Masking)	Categorization Accuracy			Spatiotemporal Coherence, Location	Coherent > Randomized	SPECT
				Spatiotemporal Coherence, Location, Ordinal Position of Target	Ordinal position positive slope	SPECT
	vERP of Target					
		Frontal & Central	50-149	Region, Location, Spatiotemporal Coherence	Coherent = Randomized	Feedforward
		Frontal & Central	150-249	Region, Location, Spatiotemporal Coherence	Coherent > Randomized	Matching
		Frontal & Central	250-449	Region, Location, Spatiotemporal Coherence	Coherent > Randomized	Post-identification

	Parietal/ Occipital	50-149	Region, Location, Spatiotemporal Coherence	Coherent = Randomized	Feedforward
	Parietal/ Occipital	150-249	Region, Location, Spatiotemporal Coherence	Coherent = Randomized	Feedforward
vERP of all scenes					
	Frontal & Central	50-149	Region, Location, Spatiotemporal Coherence	Coherent = Randomized	Feedforward
	Frontal & Central	150-249	Region, Location, Spatiotemporal Coherence	Coherent > Randomized	Matching
	Frontal & Central	250-449	Region, Location, Spatiotemporal Coherence	Coherent > Randomized	Post-identification
	Parietal/ Occipital	50-149	Region, Location, Spatiotemporal Coherence	Coherent = Randomized	Feedforward
	Parietal/ Occipital	150-249	Region, Location, Spatiotemporal Coherence	Coherent = Randomized	Feedforward
vERP Divergence					
	Frontal		Spatiotemporal Coherence		Matching, Post-identification
	Central		Spatiotemporal Coherence		Matching, Post-identification
	Parietal/ Occipital		Spatiotemporal Coherence		Feedforward

vERPs within a trial					
	Frontal & Central	50-149	Region, Location, Spatiotemporal Coherence	Coherent = Randomized	Feedforward
	Frontal & Central	150-249	Region, Location, Spatiotemporal Coherence	Coherent > Randomized	Matching
	Frontal & Central	250-449	Region, Location, Spatiotemporal Coherence	Coherent > Randomized	Post-identification
	Parietal/ Occipital	50-149	Region, Location, Spatiotemporal Coherence	Coherent = Randomized	Feedforward
	Parietal/ Occipital	150-249	Region, Location, Spatiotemporal Coherence	Coherent < Randomized	Matching
vERP of Target					
	Frontal		Spatiotemporal Coherence, Image Predictability	Positive Correlation between Amplitude and Predictability	Matching
	Central		Spatiotemporal Coherence, Image Predictability	Positive Correlation between Amplitude and Predictability	Matching
	Parietal/ Occipital		Spatiotemporal Coherence, Image Predictability	Negative Correlation between Amplitude and Predictability	Matching

Decoding Accuracy				
	Frontal	Spatiotemporal Coherence	Coherent > Randomized	Matching, Post-identification
	Central	Spatiotemporal Coherence	Coherent > Randomized	Matching, Post-identification
	Parietal/ Occipital	Spatiotemporal Coherence	Coherent > Randomized	Early, Matching, Post-identification
Correlations between model and human responses				
	Frontal	Spatiotemporal Coherence	Coherent > Randomized	Matching, Post-identification
	Central	Spatiotemporal Coherence	Coherent > Randomized	Matching, Post-identification
	Parietal/ Occipital	Spatiotemporal Coherence	Coherent > Randomized	Early, Matching, Post-identification

2

Behavioral Results

We began by examining if scenes shown in coherent sequences were more predictable than scenes shown in random sequences. We conducted all of the analyses in R (version 4.0.1). We used a logistic mixed effects model to assess if image predictability (correct = 1, incorrect = 0) was greater in the coherent than in the randomized sequences. We specified logistic mixed effects models using the lme4 library (Bates et al., 2014). Degrees of freedom were corrected with a Kenward-Roger correction (Kenward & Roger, 1997). Least squared means and their corresponding standard errors were obtained with the emmeans library (Lenth et al., 2018). We probed significant interactions between fixed effects using the emmeans library, and we adjusted *p* values associated with interactions with a Bonferroni correction.

Spatiotemporal coherence was dummy coded as a 1 for the randomized condition and a 0 for the coherent condition. Location of the image (on-campus vs. off-campus) was dummy coded as a 1 for on-campus and a 0 for off-campus prior to entry into the model. The location where the image was photographed (on-campus vs. off-campus) was treated as a nuisance variable in all of the analyses. Thus, we will only describe effects of location when it interacted with one of the other factors of interest though we still included it in the Tables of the model output. We determined the random effect structure of the models from the design of the experiment (Barr et al., 2013; Bello et al., 2016; Bello & Renter, 2018; Stroup, 2012). We fit the 'maximal' model, as opposed to either i) only treating the participant at their intercept as random effects, or ii) by comparing models with different random effect structures (Matuschek et al., 2017). The version of the base sequence (A, B, and C) and the participant number nested within each version were both treated at their intercept as random effects. Further, we allowed the main effect of spatiotemporal coherence (coherent vs. randomized), the location the images were photographed

(on-campus vs. off-campus), and their interaction to vary for each participant as a random effect (i.e., by-participant and by-version intercepts and a by-participant slope random effect). This random effect structure was chosen because it decorrelates the effects contributed by each subject and each version of the base sequence with the manipulation (Singmann & Kellen, 2019), and it is considered to be the most conservative model (Barr et al., 2013; Matuschek et al., 2017; Singmann & Kellen, 2019).

We first assessed if scenes presented in coherent sequences were more predictable than scenes presented in randomized sequences consistent with hypotheses generated from SPECT. As shown in Figure 4, most of the scenes were more predictable in the coherent than the randomized sequences. Prediction accuracy was significantly greater for scenes in coherent sequences (M = 0.31, SE = 0.02), β = -.71, SE = 0.07, z = -10.30, p < .001, BF > 1,000.[9] Further, scenes presented in the randomized (M = 0.18, SE = 0.01, 95% CI = [0.12 0.21]) sequences were very close to chance (12.5%). Thus, our results confirmed our hypothesis.

To assess how well the logistic mixed model was able to classify correct from incorrect trials, and thus to provide a descriptive statistic that is analogous to an R^2 for the overall model fit, we fit a receiver operating characteristic (ROC) curve using the model's predictions and the

[9] Bayes factors for all mixed effects models are reported in favor of the alternative hypothesis. We used uninformed priors to calculate Bayes factors. We calculated Bayes factors by taking the exponential of the difference between the Bayesian information criterion values of the intercept-only *null* model that did not contain the effect of interest and the model containing the effect of interest divided by negative 2 (Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192-196.) The random effect structures were the same between the two models, so that models only differed in the inclusion or exclusion of the main effect of interest. Values greater than 1 suggest that the model containing the effect (i.e., the alternative hypothesis) was a better model, and values less than 1 provide evidence in favor of the intercept only, null hypothesis Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving*, *7*(1), 2. .

raw values, and then estimated the area under the ROC curve (Kuhn & Johnson, 2013) using the pROC package in R (Robin et al., 2011). A model with perfect discrimination ability has a ROC curve close to 1 and a model with poor ability to discriminate correct from incorrect trials has a ROC curve close to 0.5. The logistic model we fit to prediction accuracy did well at discriminating correct from incorrect trials, with an AUC = 0.74.

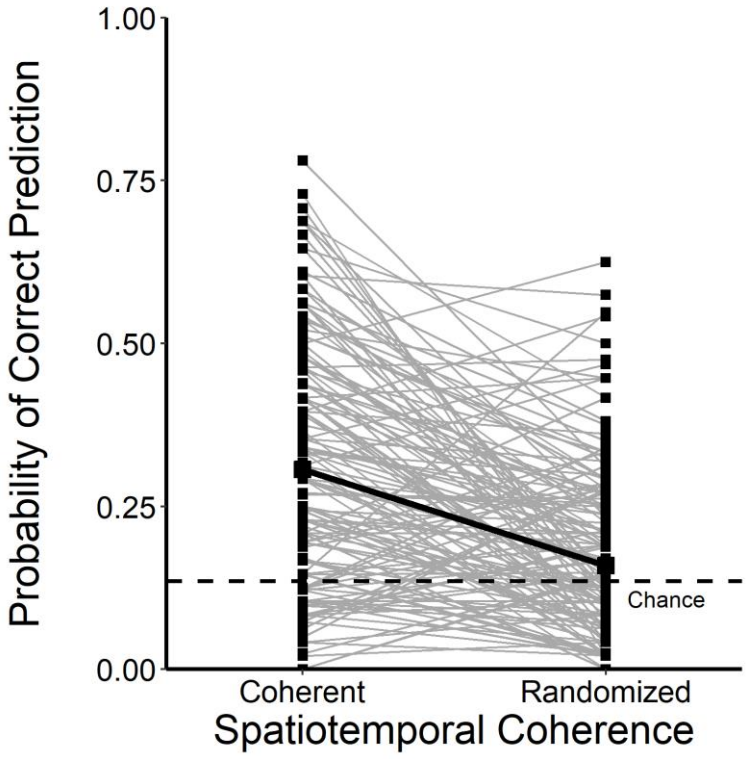


Figure 4. Exp 1: Image predictability as a function of the spatiotemporal coherence of the image sequences. The proportion of times each **image** was accurately predicted is represented by the gray lines. Least square means generated from the estimated regression equation are represented by the thick black line. The dashed line at 12.5% represents chance performance.

We also explored how prediction accuracy changed as a function of the temporal position of the target (2-10) on each trial. According to SPECT, scene categories should become more predictable as the event model constructs a representation of the sequences across time. When we know where we are, or where we are going to be, we can use this knowledge to predict the range of the most likely scenes that will appear next. As such, images presented in the 10^{th} position on a trial should be more predictable than images presented in the 2^{nd} position, as the event model should be much more developed by the 10^{th} scene on a given trial.

The logistic mixed effects model contained the same fixed effects as before, with the addition that the ordinal position of the target (centered at its mean) was also included as a continuous fixed effect and as a by-participant slope random effect. Again, we observed a significant main effect for spatiotemporal coherence, β = -0.71, *SE* = 0.07, z = -10.32, $p < .001$, BF > 1,000. We also found a marginally significant positive effect for the ordinal position of the target, β = 0.03, *SE* = 0.02, z = 1.95, p = .05, BF = .06, which was notably associated with a small Bayes factor in favor of the alternative hypothesis. As evident in Figure 5, the location where the images were photographed interacted with the ordinal position of the target, β = -0.08, *SE* = 0.02, z = -3.70, p = .0002, BF = 8.91. We probed this significant interaction and found that predictability of the target scene increased as a function of the ordinal position for off-campus scenes as we hypothesized, β = 0.03, *SE* = 0.02, z = 2.10, p = .02; but prediction accuracy decreased as a function of ordinal position in the on-campus sequences, β = -0.05, *SE* = 0.01, z = 4.12, $p < .001$. The explanation for the decrease in prediction accuracy in both the coherent and randomized sequences over time in the on-campus sequences is unclear; though it appears in Figure 5, to be driven by very high prediction accuracy (*M* = .49, *SE* = 0.02) for observations in the 2^{nd} position in the on-campus sequences. In fact, the interaction between the ordinal position

and location disappeared when we removed observations from the 2nd position in an exploratory analysis. Instead, we found a significant interaction between spatiotemporal coherence and the ordinal position of the target as we predicted, β = -0.05, *SE* = 0.03, *z* = -2.08, *p* = .03; whereby the slope for coherent sequences was positive, β = 0.04, *SE* = 0.01, *z* = 3.16, *p* < .001 but the slope for the randomized sequences was not significant, β = -0.03, *SE* = 0.02, *z* = -2.11, *p* = .98. Further, the initial model was able to discriminate correct from incorrect trials, but not very well, AUC = 0.67.[10]

[10] The exploratory model with observations from the 2nd position excluded was also unable to discriminate correct from incorrect trials very well, AUC = .67. We also reran the initial model using a probit link function, but the ability of the model to discriminate correct from incorrect trials did not improve, AUC = .63. We also treated the ordinal position of the scene as a categorical variable. Again, we found that the ability of the model to discriminate correct from incorrect trials did not improve, AUC = .66.

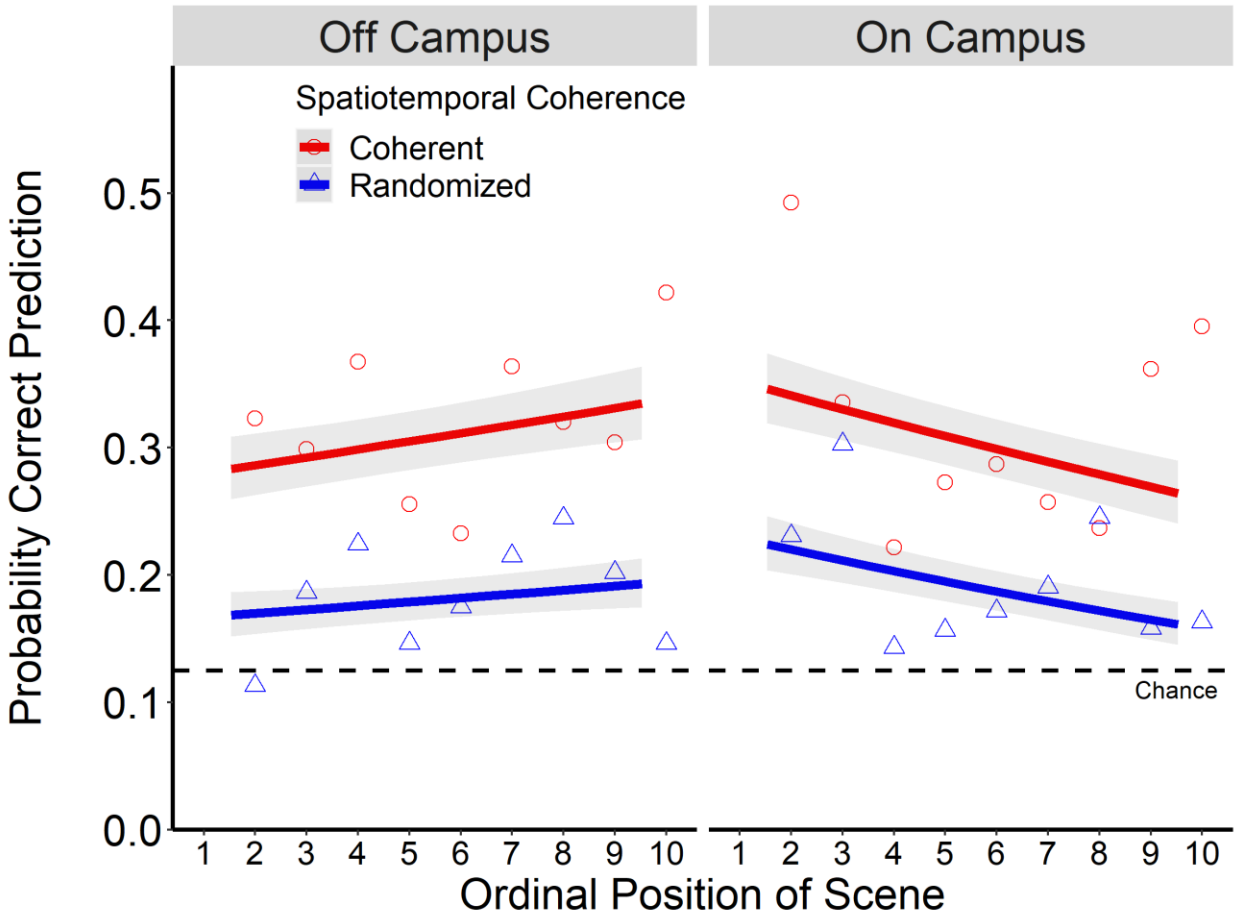


Figure 5 Exp 1: Image predictability as a function of the ordinal position (2-10) of the target on each trial, the spatiotemporal coherence of the image sequences, and the location the images were photographed. The proportion of instances when the target was correctly predicted at each ordinal position (2-10) is represented by the dots. The lines reflect the least square means calculated from the estimated regression equation. Error ribbons reflect 1 standard error to the estimated means. The dashed line at 12.5% represents chance performance.

Discussion

We found that scenes presented in coherent sequences were more predictable and that prediction accuracy for images presented in randomized sequences was only slightly above

chance performance. Participants could have been at above-chance performance in the randomized condition for a variety of reasons. Sidewalks far outnumber the other categories in the sequences (See Table 1), so guessing a sidewalk on any given trial could result in significantly greater than chance performance in the randomized condition. Further, there were only five categories per trial even though participants had 8 options to choose from on each trial, so participants could be using a process of elimination and statistical learning to correctly guess the category of the image they were going to see next. Most importantly, however, images in coherent sequences were more predictable than images shown in randomized sequences, which replicated the results reported by Smith and Loschky (2019) after adding 960 new image sequences, thus tripling their number.

We also found that image predictability increased as the event model was being constructed in the images that were taken off-campus, but we did not find this increase over time in the on-campus sequences. In fact, image predictability over time decreased in the on-campus images. An explanation for this result is unclear though an evaluation of the individual data points reported in Figure 5 show that prediction accuracy for the on-campus sequences was greater in the coherent than in the randomized scene sequences in all other ordinal positions of the target except for the 8^{th} position, and the slope of the function became positive in both locations when we reran the analysis after removing observations in the 2^{nd} position in an exploratory analysis. This could be an issue with how some of the coherent sequences were created. Not all of the scenes in a coherent sequence are predictable (See Figure 4 for prediction accuracy for each scene), though the majority are more predictable when the sequence is randomized.

Chapter 3 - Experiment 2

Having identified images with a range of predictability, we next investigated if perception of a scene changes when it is shown in a spatiotemporally coherent versus randomized sequence. We had two competing hypotheses. According to feed-forward theories of scene perception, vERPs time locked to the onset of the scenes should not differ from images presented in randomized sequences (Serre et al., 2007; VanRullen, 2007; VanRullen & Thorpe, 2001b, 2002), or if they differ, it should occur very late in the epoch (consistent with post-identification accounts) (Ganis & Kutas, 2003; Hollingworth & Henderson, 1998, 1999). According to feed-forward theories of perception, predictions for an upcoming scene do not influence scene perception because scenes are recognized based on their features. As such, vERPs in response to scenes presented in spatiotemporally coherent versus randomized sequences should not differ, or they should only do so very late in the epoch, such as in the N400 time window or later. Alternatively, SPECT proposes that the viewer's back-end event model for the sequences should feedback and facilitate front-end information extraction. Furthermore, according to SPECT, because the event model will have already been constructed prior to seeing the target image, such feedback activation could logically be present prior to onset of the target. Thus, the typical longer latencies reported for feedback from higher-order cortical areas (e.g., prefrontal cortex [PFC]) to earlier cortical areas (e.g., V1, V2, V4) (e.g., Kar & DiCarlo, 2021; Kar et al., 2019) does not apply in this situation. Accordingly, predictions for an upcoming scene, made prior to viewing it, should influence vERPs either around or before 150 milliseconds.

Motivated from SPECT, we had 3 hypotheses for when in the time course of scene processing differences in vERPs may arise. Early facilitation accounts (Biederman et al., 1982;

Palmer, 1975b) predict that differences in vERPs between coherent and randomized sequences appear in the earliest components associated with scene processing (0-149 ms post-scene onset). Such early vERP components have been shown to be affected by perceptual attributes of the stimulus (Anllo-Vento et al., 1998; Clark et al., 1994; Kenemans et al., 1993), and task manipulations that affect how information is attended (Hillyard et al., 1998). Alternatively, matching accounts of facilitation predict that differences in vERPs arise in later components (150-249 ms). Such a difference could suggest that predictions for an upcoming scene influences the process of comparing scene structural descriptions with expected scene categories (Bar, 2004; Bar & Ullman, 1996; Friedman, 1979; Schendan, 2019; Trapp & Bar, 2015). Finally, post-perceptual accounts argue that predictions do not influence scene perception, but rather predictions influence processes involved in response selection or in integrating current with prior information (Ganis & Kutas, 2003; Hollingworth & Henderson, 1998, 1999).

Though analyses of vERPs were the primary focus of Experiment 2, we also complemented the vERP analyses with a machine learning approach to investigate the temporal dynamics of scene category representations and how emerging categorical representations contributed to behavior. By showing scenes in either spatiotemporally coherent or randomized sequences, we can determine if information diagnostic of a scene category, can be more accurately decoded from participants' brain activity when scenes are shown in predictable sequences. If such an EEG decoding advantage for scenes shown in coherent sequences exists, it would suggest that participants' brains can more efficiently and/or effectively represent scene category information when scenes are predictable. If the brain can more *efficiently* represent a scene category, it should do so more quickly, and with fewer resources. If the brain can more

effectively represent a scene category, it should produce greater accuracy[11]. Again, the time course for when decoding accuracy diverges between coherent and randomized sequences would suggest when the event model begins to facilitate scene categorization.

As mentioned previously, decoding often mirrors behavioral measures. For instance, decoding and rapid scene categorization accuracy are reduced when scenes are inverted (Walther et al., 2009) or when participants categorize poor exemplars of scene categories (Torralbo et al., 2013). In addition, the distribution of decoding errors over time as well as in various brain regions, correlate with distributions of errors made by participants (Ramkumar et al., 2016). By showing scenes in either spatiotemporally coherent or randomized sequences, we can determine if patterns of responses made by neural decoders agrees more strongly with patterns of responses made by human observers when the sequence is predictable. Such an effect provides direct neural evidence that scenes shown in coherent sequences are represented in more detail.

Method

Participants

Twenty-seven (N = 16 females, N = 11 males) students participated in Experiment 2 for course credit. All the participants were right-handed. Age of the participants ranged from 18 to 20 (M age = 18). We removed EEG data from 3 participants due to excessive movement artifacts. More than half of the trials were removed from these participants after cleaning the data with visual inspection. We retained their behavioral data in the analysis of rapid scene

[11] Decoding accuracy can also be low for a variety of reasons that do not correspond to the brain's efficiency to process information. For instance, if the neural signals that produce the activity are deep within the brain (as in the case of many scene-related brain regions), then the signal can be weak relative to the noise. Decoding accuracy will be low if the signal is harder to detect with EEG.

categorization performance. Participants began the experiment by signing an electronic informed consent form, authorized by Kansas State University's Institutional Review Board. We screened participants for normal or corrected to normal vision (< 20/30 Snellen acuity) prior to participating in the experiment using the Freiburg Visual Acuity and Contrast Test (Bach, 2006). None of the participants were aware of the experiment's purpose, and they were instructed not to discuss the purpose of the experiment with others. The experiment lasted approximately 2.5 hours, including the time to set up the EEG cap. Participants were encouraged to take a break halfway through the experiment, after which the electrodes were reset onto the participants' scalp.

We determined the sample size from a power analysis using an effect size estimate reported in McLean et al. (2021). Specifically, we used the effect size, partial $\eta^2 = .25$, McLean et al. (2021) reported for the difference between expected and unexpected scenes over the parietal and occipital regions. This effect size was chosen because it was the smallest statistically significant effect reported by McLean et al. (2021), and facilitation during the P200 time window demonstrates the clearest evidence that predictions made prior to viewing a scene may facilitate scene perception. With an alpha = .05, power of 0.95, 4 groups (Image Congruency: Expected vs. Unexpected X Hemisphere), and an assumed correlation among repeated measures = .1, G*Power indicated that a total sample of 20 participants was needed to find an effect of this magnitude or larger. Thus, a sample size of 24 participants was assumed to be adequately powered to detect the effects of interest on the P200 component in the parietal/occipital region.

EEG Data Acquisition and Preprocessing

We recorded EEG signals using a 64-channel electrode system (Net Station, Electrical Geodesics, Inc., Eugene, OR, USA) at a sampling rate of 1,000 Hz. We used a Net Amps 400

(NA 400) amplifier and tin electrodes mounted in a HydroCel Geodesic Sensor Net. Select electrodes on the net correspond to locations in accordance with the international 10-20 coordinate system. Data were referenced from Cz during data acquisition and impedances were maintained below 50 kΩ. The system is a high impedance system. Data were referenced offline to the average of the left and right mastoids. We placed electrodes underneath the eyes to monitor for eye movement and blink artifacts.

We conducted offline data preprocessing using the open-source EEGLAB toolbox (Delorme & Makeig, 2004) and custom MATLAB 2019a scripts. We first down sampled the data to 256 Hz. We high passed filtered the data at a 0.1 Hz cut-off frequency to remove DC offset, and then low-pass filtered at 50 Hz to eliminate 60 Hz noise.

We manually removed channels and portions of the continuous data if visual inspection indicated that they were contaminated by noise or excessive movement. Specifically, data from 12% percent of all of the scenes were removed (Range = 9% - 23%). We then submitted the remaining data to Independent Component Analysis (ICA) to identify ocular artifacts, lateral eye movements, EKG, and channel noise. We used the ICLabel plugin for EEGLAB to identify problematic ICAs (Pion-Tonachini et al., 2019). Components classified as brain activity with a probability of 80% or above were retained in the final dataset. We rejected any component ICLabel classified as eye, muscle, heart, line noise, or channel noise with a probability at or above 80%. We used each component's spectra and scalp distribution to identify other components that were problematic, and we reconstructed the data without them. On average, we removed 22.71 (*SD* = 6.03), components across participants (Range = [12 - 35]).

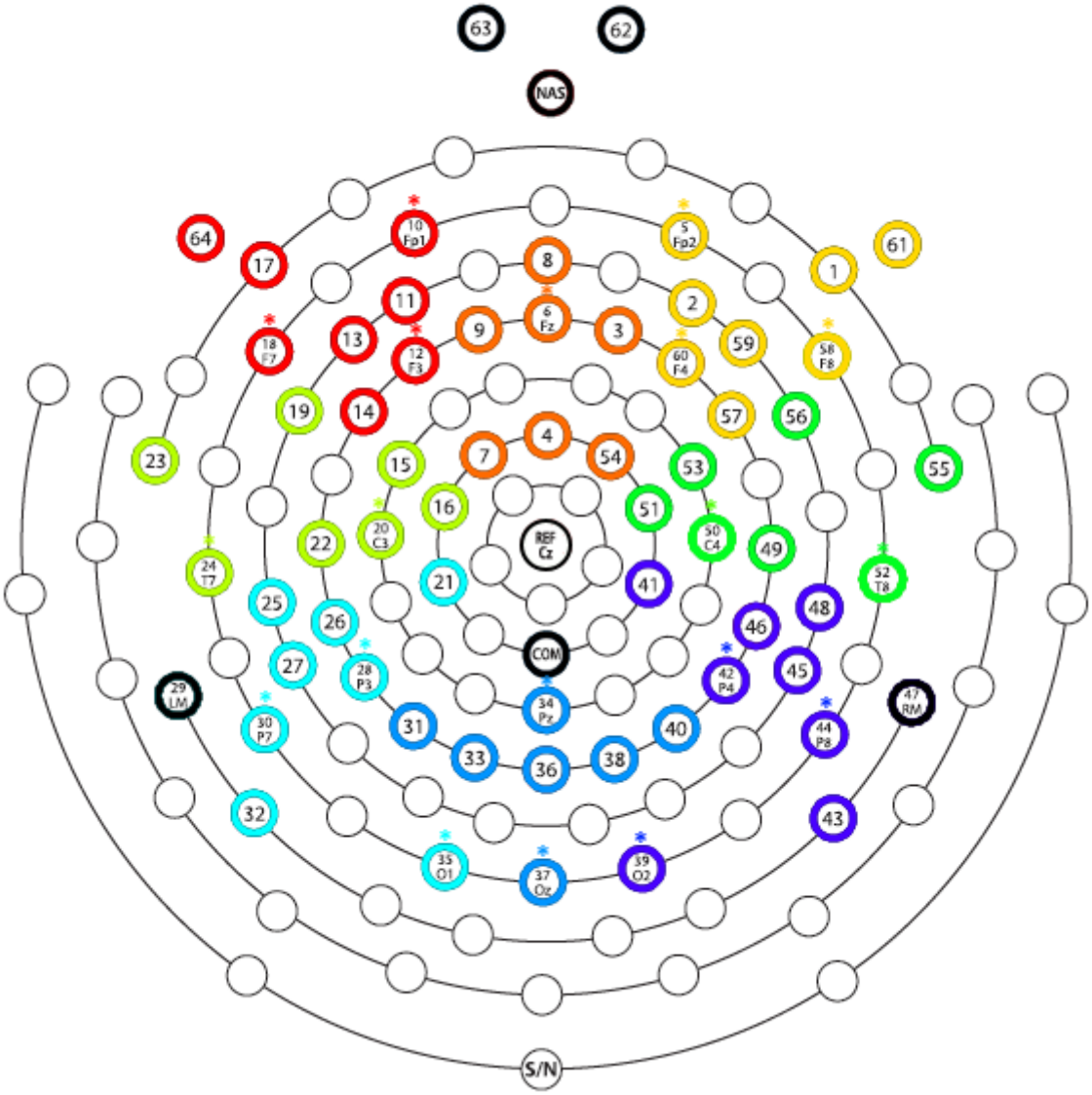
Evoked Potential Recordings

We then epoched the data for 800 ms, from 200 ms prior to the onset of the images, as recommended by Luck (2014), to 600 ms after scene onset. We applied a baseline correction on the vERPs by subtracting the mean voltage in the 200 milliseconds prior to scene onset from voltage at every time point in the epoch (Luck, 2005). To reduce the number of comparisons, we pooled the 64 electrodes into eight different regions of interest (see Figure 6). These regions of interest were arranged by hemisphere and area using the following division according to the 10-20 system: Left Frontal [Fp1, F7, F3]; Middle Frontal [Fz]; Right Frontal [Fp2, F4, F8]; Left Central [T7, C1]; Right Central [C4, T8]; Left Parietal/Occipital [P3, O1, P7]; Middle Parietal/Occipital [Pz, Oz]; Right Parietal/Occipital [P4, O2, P8]. We chose these ROIs because they have previously been found to display maximal amplitude of the components of interest and they are similar to what was used in previous research (see Table 17 in the Appendix). The remaining electrodes within each ROI that did not correspond to an electrode in the 10-20 system were grouped according to their location and the distributions of neural activity from prior work (See Table 17 in the Appendix). Each ROI contained 7 or 8 electrodes.

We determined the EEG-dependent measures for analysis (mean amplitudes over given stretches of time) in two different ways: 1) a priori windows in time and space determined from the prior literature, and 2) a data-driven permutation-based analytical method to determine the first time point at which vERPs in coherent and randomized conditions differed. For the component-based analyses, we used time windows and electrode selections that have been used in the prior literature (see Table 17 in the Appendix). Amplitudes of the N400 were taken as the mean of all data points between 250-449 ms over frontal and central cites for each participant, respectively (Kutas & Federmeier, 2000, 2011).

We also analyzed the data by splitting the large window of the N400 into two smaller windows to capture both the N300 and N400. See the Appendix for details. We used 250-349 as the window of the N300 (Hamm et al., 2002; Kumar et al., 2021; Lauer et al., 2018; Smith & Federmeier, 2020; Võ & Wolfe, 2013), and 350-449 ms as the window of the N400. Consistent with prior work, we did not find differences in the results when we analyzed the components separately, so those results are provided in the Appendix.

Less standardization exists regarding the P200. As such, we used a window of 150-249 ms because it covered the 220 ms timepoint previously identified as showing maximal amplitudes for scene processing (Harel et al., 2016), and it ensured that all three components of interest (including the N300) were the same size (100 ms). We also explored early differences, arising in the 50-149 ms window. Details of the permutation-based simulation are reported in the Analysis of vERP divergence section.



259

260 *Figure 6*. EGI 64-channel sensor layout. The EGI 64 channel HydroCell Geodesic Sensor Net is
261 displayed above. Each region of interest for the analysis is color coded. Red, orange, and yellow
262 electrodes represent frontal electrodes, green electrodes represent central electrodes, and blue
263 electrodes represent parietal/occipital electrodes. Electrodes indicated with an astericks
264 correspond to locations of the 10-20 system.

Neural Decoding Procedure

To examine how scene category representations emerged over time, we submitted vERP data at each time point within the epoch to a pattern classification analysis using a linear kernel support-vector machine (SVM) with the e1071 (V1.7.6) implementation of libsvm in R (Meyer et al., 2019; Meyer & Wien, 2015). We obtained one decoding matrix for each time point within each ROI for each participant. We performed decoding separately for each of the 8 regions of interest (See Figure 6) for on- and off-campus sequences shown in the coherent and the randomized conditions, respectively. Thus, we ran 4 total sets of classifiers [2 locations (on- vs. off-campus) X 2 spatiotemporal coherence conditions (coherent vs. randomized)] at each time point. Voltage at each electrode within each region was the independent variable input to the SVM algorithm and the scene category predicted by the SVM algorithm was the dependent variable. We used the one-against-one approach for each classifier, where, based on using 8 scene categories, we trained a total of $8(8-1)/2 = 28$ binary classifiers at each time point in the epoch. The appropriate 1 class out of 8 was determined by a voting scheme using the default functionality of the e1071 SVM implementation in R (Meyer et al., 2019; Meyer & Wien, 2015).

We started by first dividing the data into a training and a test set. We trained an SVM on each time point in the epoch using data from all the images, except for target scenes, and we tested the model on the corresponding time point within the epoch time locked to the onset of the target. We evaluated accuracy of the model on the target scenes. We chose to run the analyses this way, as opposed to testing the accuracy of the models using k-fold cross validation, as is typically done (i.e., Greene & Hansen, 2020; Ramkumar et al., 2016), because participants only categorized the target image on each trial. This ensured that there was a one-to-one correspondence between responses made by the neural decoders and the participants (i.e., the

participant and the model categorized the same scenes). In addition, we improved the signal-to-noise ratio when training the SVM on the single trial data by creating 'super trials' for each time point (Cichy et al., 2016; Greene & Hansen, 2020). To do this, we chose 10 images from each category at random with replacement and averaged their signals, time locked to the onset of the scenes. We repeated this process separately for each time point in the epoch 100 times for each scene category except that the model was not tested on super trials. Thus, models were trained on (100 super trials X 8 categories) 800 data points at each time point and tested on the number of target scenes within each of the on-and off-campus sequences (e.g., 144 in participants without artifacts on the target image) for each participant. We then evaluated if spatiotemporal coherence modulated the ability of the model to decode the scene categories from the EEG with a linear mixed effects model.

One concern with using algorithms to categorize scene categories from neural signals is whether the information used by the machine learning technique corresponds to the same information used by humans for scene categorization. Thus, we investigated if human participants use that same information for scene categorization by correlating SVM responses at each time point within the epoch with behavioral confusion matrices from each participant. If both humans and the neural decoder rely on the same category-specific signal, then there should be a correspondence between the responses made by humans and the responses made by the neural decoder. Prior to running these series of correlations, we removed observations from participants' behavioral responses when EEG data corresponding to the image was removed due to artifacts. We did this to ensure a one-to-one correspondence between responses made by the SVM and responses made by participants.

Apparatus

All images were presented in color on a 17-inch Samsung SyncMaster 957 MBS CRT computer monitor with an image resolution of 1024 X 768 pixels running at an 85-Hz refresh rate. We did not stabilize the participants' head from the computer monitor though participants were approximately 53 cm from the screen. Single pixels subtended approximately 0.04 degrees of visual angle.

Procedure

The experiment was programmed in Experiment Builder (version 2.2.1) (SR Research, Mississauga, ON, Canada) and with custom Python scripts.

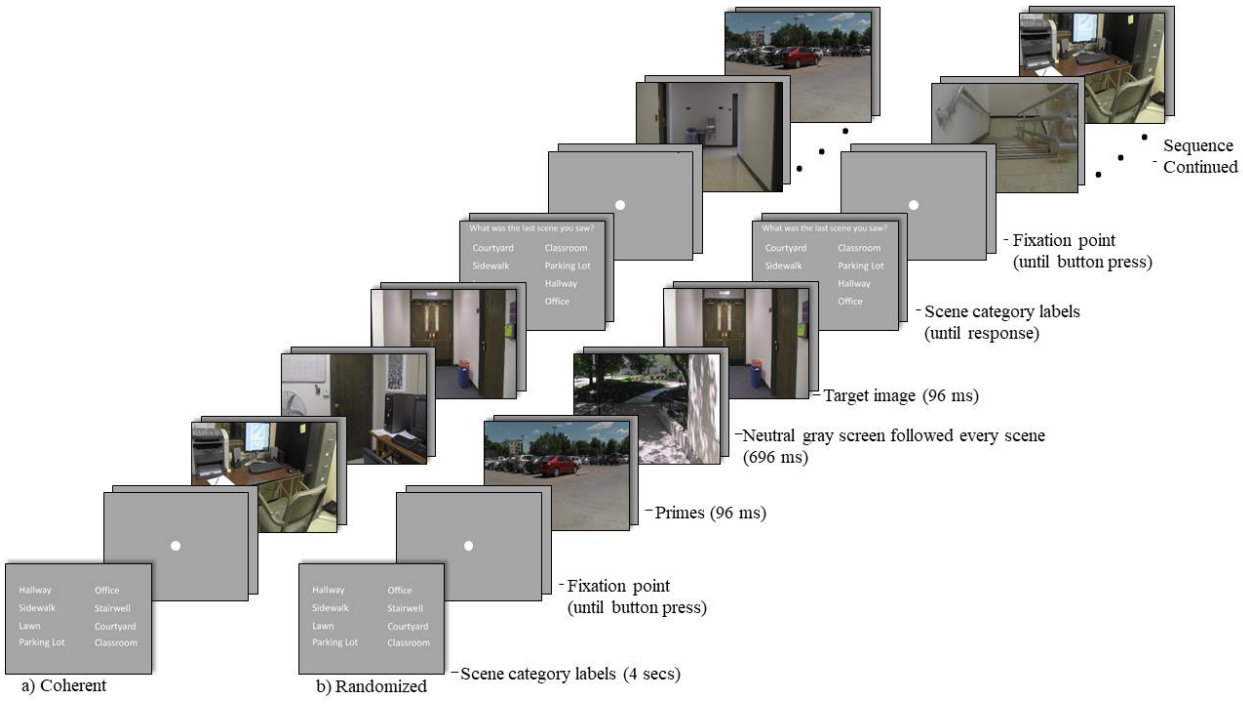


Figure 7. This is a trial schematic of Experiment 2. Scenes were shown in either a a) coherent or b) randomized sequence. Participants were asked to categorize 1 target on each trial in an 8 alternative forced choice task. The ordinal position of the target (2-10) was randomly chosen on

each trial. The target scene and its ordinal position was the same in the coherent and randomized sequences.

As shown in Figure 7, we yoked the experimental design from Experiment 1 with the following exceptions. In Experiment 1, participants only viewed a subset of the total set of images. In Experiment 2, participants viewed all 1,440 images twice, once in each condition. Second, we adjusted the duration of the image presentation to fit a monitor with a refresh rate of 85 Hz. Images in Experiment 2 were flashed for 96 ms each, interleaved by a 696 ms neutral gray screen (i.e., 792 ms SOA). The target image within each trial was the same between experiments. However, instead of asking participants to predict what the next scene will be, after briefly flashing the target scene, we asked participants in Experiment 2 to categorize it from an 8-AFC array of scene-category labels.

Participants read a series of "Yes"/"No" questions after the experiment concerning whether they noticed the differences in coherence between sequences. Participants were asked the following four questions: 1) "Did anyone tell you anything about this study prior to participating in it?" 2) "Did you notice anything in the experiment?" 3) "Did you notice anything about the image sequences?" and 4) "Did you notice that some of the image sequences appeared as if you were walking from one location to another?" If participants responded "Yes" to any of the questions, they were asked to describe what they were told, or what they noticed, by typing in a text box on the computer screen. Based on participants' written responses, two raters independently judged if each participant reported anything meaningfully related to the coherence manipulation of the image sequences. Raters produced strong interrater reliability (Cohen's $k = 0.88$). Discrepancies between raters were resolved through thoughtful discussion to produce the

final coding of the participant responses. Few participants reported that they noticed the manipulation on the second question (3%). Many more reported they noticed the coherence of some of the image sequences on the third question (30%), and all of participants reported that they noticed the coherence by the final question.

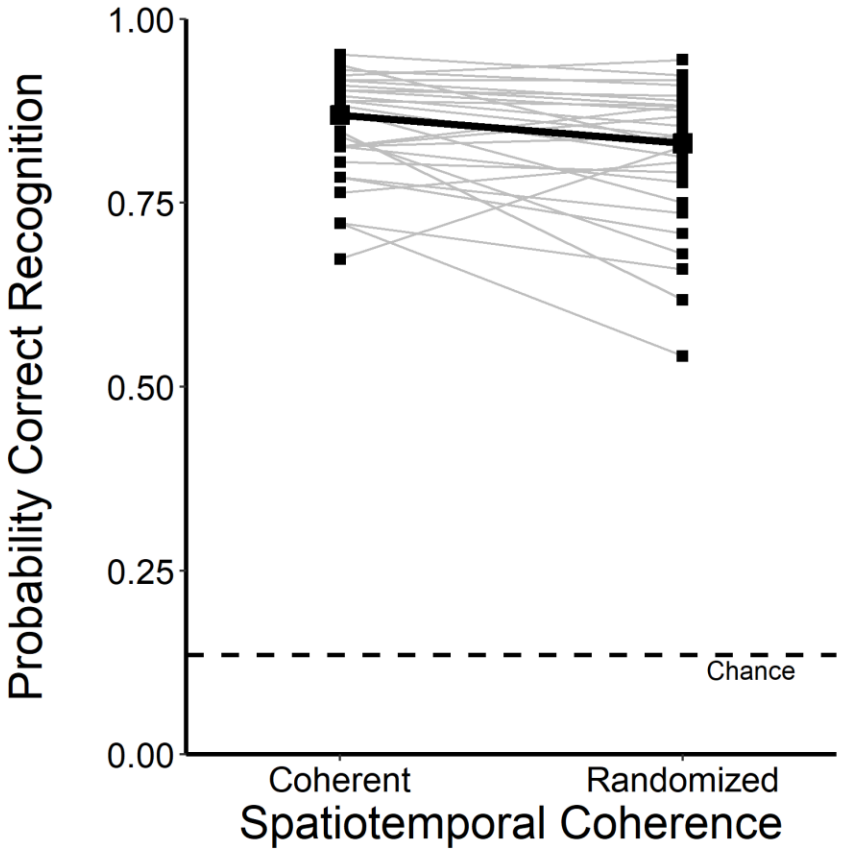
Results

We will first describe the behavioral results. These will be followed by the vERP results time locked to the target and then time locked to all of the images within the trial, regardless of whether the scene was a target. These results will be followed by a set of exploratory analyses examining the possible source of the observed vERP differences and an analysis for how each of the components changed as a function of the ordinal position of the scene within a trial. We will then report an exploratory analysis examining the unique role of image predictability, as measured in Experiment 1 in modulating the vERP components of interest in Experiment 2. We will conclude by presenting the neural decoding analysis.

Behavioral Results

We used a logistic mixed effects model to predict the probability of correctly identifying the target scene from the fixed effects of spatiotemporal coherence (coherent vs. randomized), the location the images were photographed (on-campus vs. off-campus), and their interaction. Again, we specified the random effect structure of the model from the design of the experiment. The participant intercept was allowed to vary as a random effect, and the main effects of spatiotemporal coherence, location, and their interaction were allowed to vary as random effects (by-participant intercept and by-participant slope random effects). Spatiotemporal coherence (Coherent = 0, Randomized = 1) and image location (Off-campus = 0, On-campus = 1) were both dummy coded prior to entry into the model in the same way as Experiment 1.

We hypothesized that rapid scene categorization would be accurate when scenes were shown in coherent sequences, inconsistent with feed-forward accounts of scene perception (Serre et al., 2007; VanRullen, 2007; VanRullen & Thorpe, 2001a). Despite such long SOAs (i.e., 800 ms), participants were significantly better at categorizing the targets when they were shown in coherent (M = 0.87, SE = 0.01) than in randomized (M = 0.83, SE = 0.02) sequences, β = -0.23, SE = 0.11, z = -2.09, p = .03, BF = 3.97, replicating the work of Smith & Loschky (2019) though the difference observed here was notably smaller than the difference observed in that earlier study. This effect is consistent with hypotheses generated from SPECT, namely that processes involved in the back-end feed-back to influence front-end rapid scene categorization. As shown in Figure 8, almost all of the participants in the experiment showed the behavioral advantage for the coherent sequences though overall accuracy was not as good as we would have anticipated given the long SOAs. Categorization performance usually reaches asymptote level performance of 90% with masked SOAs of 200 ms (Hansen & Loschky, 2013). The logistic model was successfully able to discriminate between accurate and inaccurate trials, AUC = 0.72.



383

384 *Figure 8*. Exp 2: Rapid scene gist categorization performance as a function of the spatiotemporal
385 coherence of the image sequences. The proportion of times each participant accurately
386 categorized the target images is represented by the thin gray lines. Least square means generated
387 from the estimated regression equation are represented by the thick black line and dots.

388

389 We also explored how categorization performance changed as a function of the ordinal
390 position of the target on each trial. See Figure 9. According to SPECT, the presentation of the
391 first scene should lay the foundation in the event model in working memory. Once an event
392 model is constructed, subsequent scene categories should become easier to predict and

subsequently easier to categorize. Furthermore, as more in-coming information is mapped onto the event model, it should become richer, and more helpful in predicting up-coming new information. To examine this possibility, we fit a logistic mixed effects model that included the spatiotemporal coherence and the location the image was photographed (on-campus vs. off-campus) as before. We also added a third variable to the model for the ordinal position of the target image (centered at its mean), which ranged from 2 to 10. The model contained the participant intercept as a random effect and the random slopes of spatiotemporal coherence, location, the ordinal position of the target scene on each trial, and all of their interactions. Again, we observed a significant effect for spatiotemporal coherence, β = -0.22, *SE* = 0.11, *z* = -1.93, *p* <.001, BF = 3.97 consistent with the previous analysis. Importantly, we also observed a significant positive effect of the ordinal position of the target (2-10), β = 0.03, *SE* = 0.005, *z* = 6.69, *p* < .001, BF = 3.45. Thus, the ability to categorize the scenes improved as the event model developed across subsequent scenes consistent with the hypothesis generated from SPECT. Interestingly, we also observed an unexpected interaction between the spatiotemporal coherence and the ordinal position of the image on the trail, β = 0.06, *SE* = 0.03, *z* = 2.13, *p* = .02, BF = 2.15; whereby the slope of categorization performance increased at a steeper rate when the images were shown in the randomized sequences, β = 0.10, *SE* = 0.02, *z* = 4.98, *p* < .001, as compared to when they were shown in coherent sequences, β = 0.04, *SE* = 0.02, *z* = 2.02, *p* = .02 (*p* values were adjusted with a Bonferroni correction). The explanation for the steeper increase in performance as a function of the ordinal position of the target scene in the randomized condition is unclear. It is possible that participants could be improving in the randomized condition over time because of educated guessing or statistical learning of the types of categories that could be shown. If a participant knows the full range of possible categories and the total sequence length

on each trial, then they may use this knowledge to keep track of how many instances of each category are on a trial. They may have been able to calculate the probability that a remaining image would be one of each of the categories on a given trial (i.e., a very rapid form of "card counting"). Of course, this explanation can also explain why categorization accuracy improves across time in the coherent sequences; however, it does not explain why categorization accuracy was better in the coherent than randomized sequences by the 2nd scene on a trial. None of the other interactions in the model were statistically significant. In addition, the logistic model was successfully able to categorize correct from incorrect trials, AUC = 0.70.

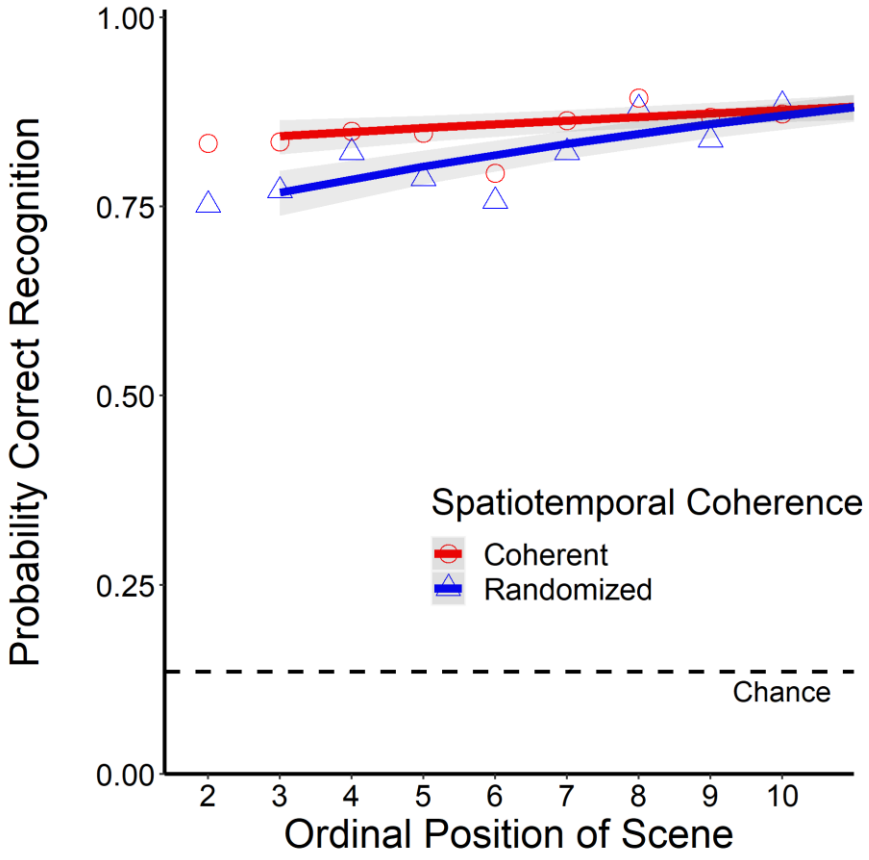
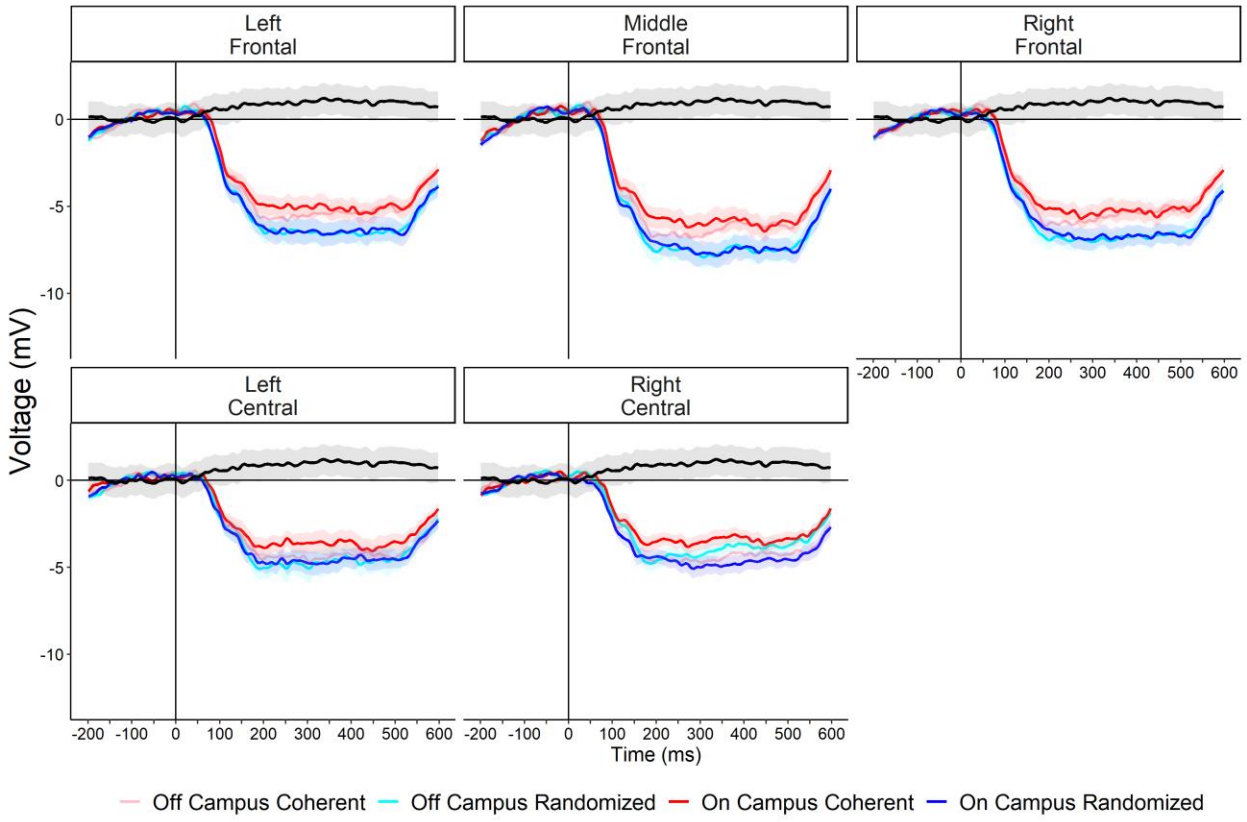


Figure 9. Exp 2: Rapid scene gist categorization accuracy as a function of the ordinal position (2-10) of the target scene on each trial and the spatiotemporal coherence of the image sequences. The proportion of instances when the target image was correctly categorized is represented by

dots in the figure. The lines reflect the least square means calculated from the estimated regression equation.

vERPs to the Target Image

Having established that rapid scene categorization accuracy is indeed worse when scenes are shown in randomized sequences, we asked what effect presenting scenes in coherent and randomized sequences would have on vERPs to the target scenes on each trial. An analysis of the target scene is important because participants were shown the same target scene in the same ordinal position (2-10) between the two conditions. Thus, targets shown in the coherent and randomized sequences only differed in the scenes that came before them. Grand averages are reported in Figures 10 and 11, time locked to the onset of the target for the frontal, central and parietal/occipital regions, respectively. Visually evoked potentials show a similar topography of vERPs reported in prior work (Greene & Hansen, 2020). The windowed analysis assessed the time course of prediction-based facilitation on scene gist perception.



441

442 *Figure 10*. Exp 2: Grand average vERP waveforms time locked to the ***target*** image for the

443 frontal and central regions. Responses to target scenes in the coherent sequences are represented

444 in red and responses to scenes in the randomized sequences are in blue. The difference between

445 the waveforms of the coherent and randomized sequences are represented by the black line. Error

446 ribbons correspond to 1 standard error to raw means.

447

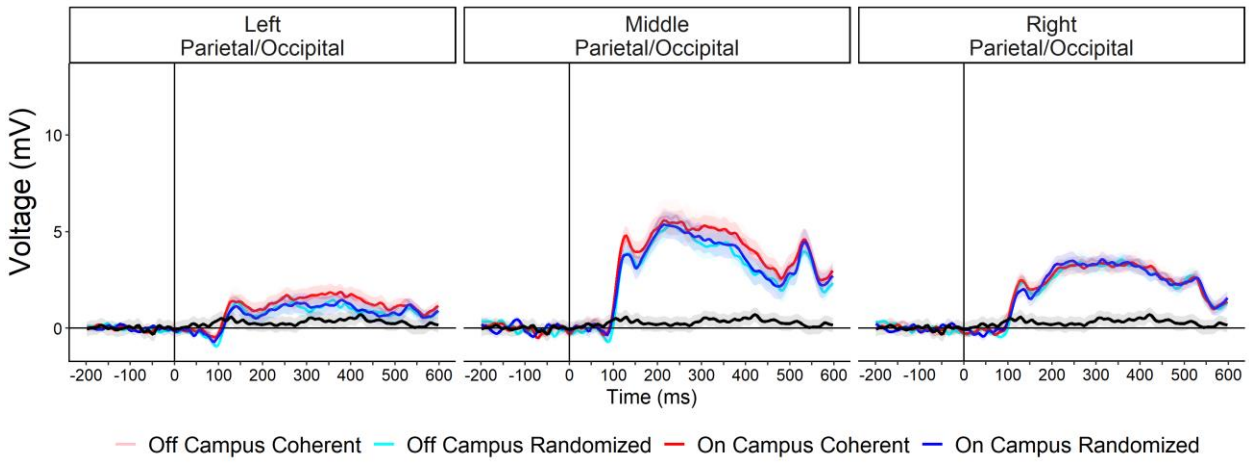


Figure 11. Exp 2: Grand average vERP waveforms time locked to the ***target*** image on a trial for the Parietal/Occipital regions. Responses to target images in the coherent condition are represented in red and responses to the images in the randomized condition are in blue. The difference between waveforms of the coherent and randomized sequences are represented by the black line. Error ribbons correspond to 1 standard error to raw means.

Frontal and Central Electrodes.

ERP components found in frontal and central electrodes reflect higher-level cognitive and executive functions (Key et al., 2005). Amplitudes were averaged at each time point in the epoch, for all trials, electrodes, regions, and participants, excluding behaviorally incorrect trials. Linear mixed effects models included the fixed effects of the region (Left Frontal, Middle Frontal, Right Frontal, Left Central, and Right Central), the location where the image was photographed (on-campus vs. off-campus), the effect of spatiotemporal coherence (coherent vs. randomized), and all of their interactions. Models contained the participant intercepts and the random slope effects of spatiotemporal coherence, the regions of interest, the image location, and

all of their interactions. Least square means of amplitude at each window are shown in Figure 12, and results from the individual models at each time window are in Table 3.

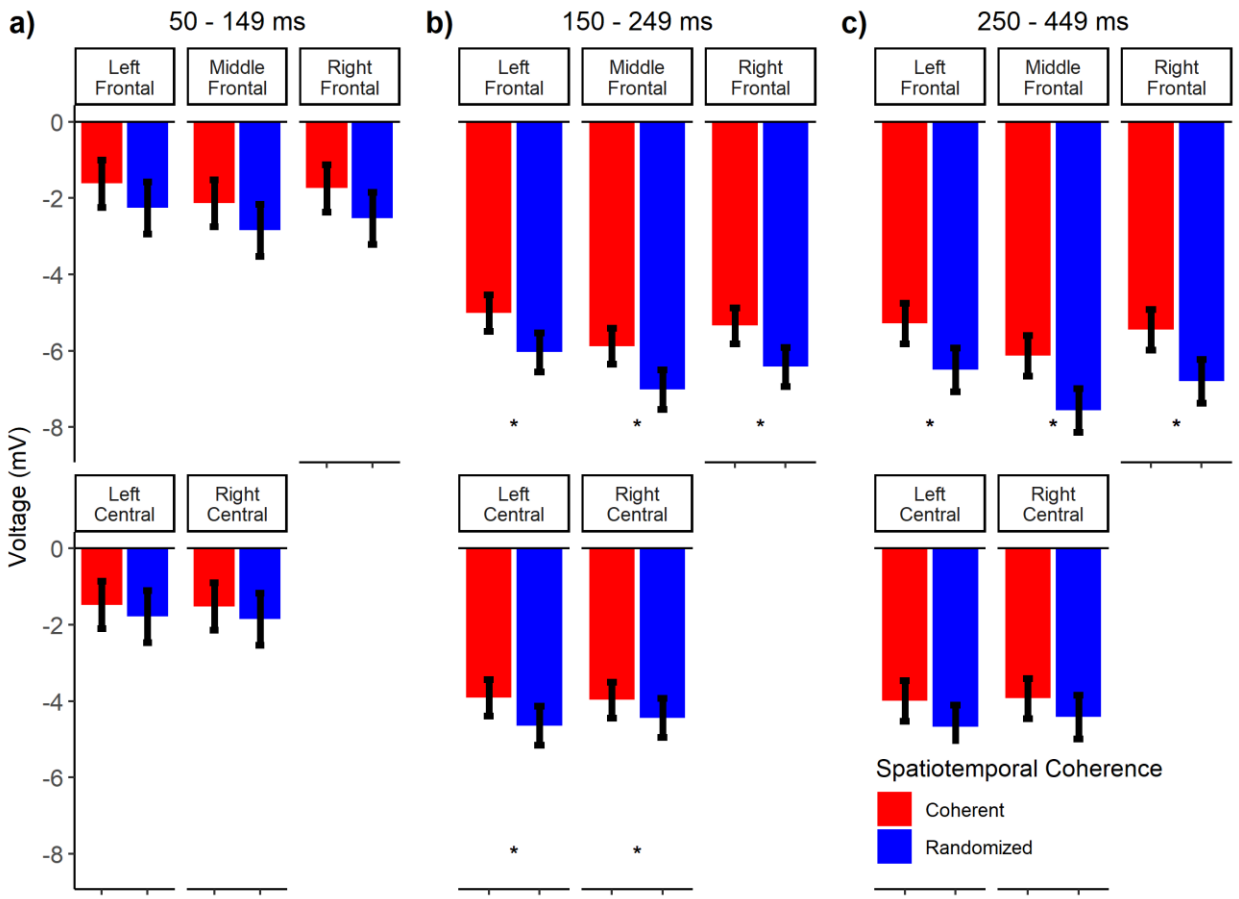


Figure 12. Exp 2: Least square means of amplitudes in response to the ***target*** at the frontal and central regions. Amplitudes are reported for the a) 50-149, b) 150-249, and c) 250-449 ms windows averaged across the location factor. Error bars correspond to 1 standard error around the estimated means.

Table 3 Exp 2: *Summary of the results for the frontal and central regions for each of the three windows (starting from 50 ms to 449 ms). Amplitudes were time locked to the onset of the target scene.*

Window	Factor	*df*	*F*	*β*	*t*	*p*
50-149 ms	Region	4,384	13.81			<.001*
	SC	1,24	0.14	0.55	3.77	.72
	Location	1,24	0.1	-0.01	-0.11	.75
	Region*SC	4,384	0.43			.79
	Region*Location	4,384	0.48			.49
	SC*Location	1,24	0.21			.65
	Region*SC*Location	4,384	0.75			.56
150-249 ms	Region	4,384	102.24			<.001*
	SC	1,24	31.84	0.89	5.64	<.001*
	Location	1,24	15.43	-0.46	-3.93	<.001*
	Region*SC	4,384	2.04			.09
	Region*Location	4,384	0.23			.92
	SC*Location	1,24	3.77			.06
	Region*SC*Location	4,384	0.52			.72
250-449 ms	Region	4,384	108.74			<.001*
	SC	1,24	24.40	-0.36	4.94	<.001*
	Location	1,24	2.17	0.72	1.47	.15
	Region*SC	4,384	3.61			.01*
	Paired t-tests (for Region*SC)					

Left Frontal			1.22	4.20	<.001*
Middle Frontal			1.43	4.94	<.001*
Right Frontal			1.35	4.67	<.001*
Left Central			0.67	2.33	.11
Right Central			0.48	1.67	.50
Region*Location	4,384	0.29			.89
SC*Location	1,24	5.20			.03*
Paired t-tests (for SC*Location)					
On-campus			0.72	3.29	.01*
Off-campus			1.35	4.83	<.001*
Region*SC*Location	4,384	2.32			.06

Note: SC = Spatiotemporal coherence of scene sequences. * denotes p < .05.

50-149 ms window.

See Figure 12a). We averaged the amplitudes at each time point within each of the regions of interest within the predetermined window and then submitted the averages to a linear mixed effects model to determine if facilitation influenced early perceptual analysis of the scenes. The corresponding least square means from the analysis are shown in Figure 12a). Responses to scenes were significantly more positive at the central channels than the frontal channels [Right Frontal (*M* = -2.13, *SE* = 0.30); Middle Frontal (*M* = -2.49, *SE* = 0.30); Left Frontal (*M* = -1.94, *SE* = 0.30); Right Central (*M* = -1.68, *SE* = 0.30); Left Central (*M* = -1.63, *SE* = 0.30)] as supported by a significant main effect of region, $F(4,384) = 13.81$, $p < .001$, BF > 1,000. Importantly we failed to find support for the early facilitation hypothesis. Average

amplitudes in the coherent (M = -1.70 SE = 0.60) and randomized (M = -2.25, SE = 0.60) sequences did not significantly differ, $F(1,24) = 0.14$, $p = 0.72$, BF = .0004. This effect was supported by a small Bayes factor in favor of the alternative hypothesis, and a large Bayes factor in support of the null, BF = 2,695[12]. This is consistent with feed-forward accounts (Serre et al., 2007; VanRullen & Thorpe, 2002), which would not predict any differences in vERPs due to spatiotemporal coherence. None of the other effects were statistically significant. See Table 3 for details.

150-249 ms window.

See Figure 12b), which shows support for the hypothesis that the event model facilitates matching processes. Consistent with the results from the earlier window, we observed a main effect of region, $F(4,384) = 102.24$, $p < .001$, BF >1,000. Importantly, however, unlike the 50-149 ms window, we also found a significant main effect for spatiotemporal coherence in the 150-249 ms window, which was supported by a Bayes factor greater than 3 in favor of the alternative hypothesis, $F(1,24) = 31.84$, $p < .001$, BF = 36.48, and thus, a small Bayes factor in support of the null, BF = 0.03. Average amplitudes in response to scenes shown in coherent sequences (M = -4.82, SE = 0.45) were significantly more positive than neural responses to images shown in randomized sequences (M = -5.71, SE = 0.50), consistent with matching accounts of facilitation (Bar, 2004; Mudrik et al., 2010; Truman & Mudrik, 2018). Thus, ***in***consistent with purely feed-

[12] The evidence in favor of the null hypothesis is equal to 1 divided by the Bayes factor in favor of the alternative. We calculated Bayes factors by taking the exponential of the difference between the Bayesian information criterion values of the intercept-only *null* model and the model containing the effect of interest divided by negative 2 (Wagenmakers & Farrell, 2004). The BIC of the null model of spatiotemporal coherence here was equal to -2260.65 and the BIC of the model that contained spatiotemporal coherence (i.e., the alternative hypothesis) was equal to -2244.85; thus, there was more evidence in favor of the null than the alternative hypothesis (smaller BIC values indicate a better model fit).

forward accounts of scene perception (Serre, Oliva, & Poggio, 2007; VanRullen, 2007), predictions made prior to viewing a scene facilitated perception of the target between 150 and 249 milliseconds post-target onset. Finding this important difference as early as 150 ms, or earlier, was predicted by the results of Smith and Loschky (2019, Exp 3), which showed that the facilitation due to spatiotemporal coherence was perceptual/sensory in nature. However, we did not expect to observe such a difference so early in the frontal and central regions. Specifically, if the EEG dipoles for these components are close to their scalp distributions, then they would represent activity in relatively higher-level cognitive areas by 150-249 ms post-stimulus. None of the interactions were significant. See Table 3 for a summary of the results.

250-449 ms window.

See Figure 12c), which shows support for the hypothesis that coherent sequences were easier to integrate into the event model. Again, we found a significant main effect for region, $F(4,384) = 108.74$, $p < .001$, BF >1,000; and an effect for spatiotemporal coherence such that amplitudes were more positive in coherent ($M = -4.96$, $SE = 0.52$) than randomized ($M = -5.99$, $SE = 0.56$) sequences, $F(1,24) = 24.40$, $p < .001$, BF = 37.64. We also observed a significant interaction between the region and spatiotemporal coherence, $F(4,384) = 3.61$, $p = .01$, BF = 16.97. The difference between the targets shown in the coherent and randomized sequences was larger at the frontal sites [Right Frontal, $\beta = 1.35$, $SE = 0.29$, $t = 4.67$, $p = .0001$; Middle Frontal, $\beta = 1.43$, $SE = 0.29$, $t = 4.94$, $p < .001$; Left Frontal, $\beta = 1.22$, $SE = 0.29$, $t = 4.20$, $p = .0003$] than at central sites [Right Central, $\beta = 0.48$, $SE = 0.29$, $t = 1.67$, $p = .50$; Left Central, $\beta = 0.67$, $SE = 0.29$, $t = 2.33$, $p = .11$]. The larger difference between the coherent and randomized conditions over the frontal region could be expected given that the N400 has a more frontal distribution when people view pictures (Ganis, Kutas, & Sereno, 1994). We also observed a significant

interaction between spatiotemporal coherence and the location where the images were photographed, $F(1,24) = 5.20$, $p = .03$, BF = 3.57. The N400 was more positive in response to scenes shown in the coherent than randomized sequences in both the on-, $\beta = 0.72$, $SE = 0.22$, $t = 3.29$, $p = .01$, and off-campus sequences, though the difference was larger in the off-campus sequences, which tend to be more difficult to categorize (i.e., lower categorization accuracy), $\beta = 1.35$, $SE = 0.28$, $t = 4.83$, $p < .001$. The larger difference in the off-campus sequences is consistent with the notion that top-down effects on perception may have a larger influence when the sensory input is ambiguous (Gregory, 1990). None of the remaining interactions were significant. See Table 3 for details.

Parietal/Occipital Electrode Sites.

ERP components found in parietal and occipital electrodes are known to be highly involved in visual processing functions (Anllo-Vento et al., 1998; Clark et al., 1994; Hillyard et al., 1998). Linear mixed effects models for the early component analysis (50-149 ms) and the analysis of the P200 (150-249 ms) included the same fixed and random effects as the analysis of the frontal and central regions. Results are reported in Table 4, and least square means of amplitude for the parietal/occipital electrode regions are reported in Figure 13.

According to early facilitation accounts, predictions influence the construction of the structural description of the scene. This would lead to the prediction that that differences between coherence conditions would be found early (e.g., 50-149 ms) in primarily visual processing areas of cortex (e.g., parietal & occipital areas). Alternatively, matching accounts propose that predictions influence processes that arise when matching the structural description to a representation stored in semantic memory somewhat later (e.g., 150-249 ms). In either case, we hypothesized that vERPs would be more positive when the target was in a randomized than a

coherent sequence considering that prior work showed a reduction in amplitudes of the P200 when participants were shown expected scenes (McLean et al., 2021), possibly because it was easier to group the visual features when the target was expected (Han et al., 2005). Either account would be consistent with the results of Smith and Loschky (2019, Exp 3), which showed that facilitation due to spatiotemporal coherence was sensory or perceptual in nature. Conversely, feedforward accounts of rapid scene categorization would not predict any differences in the vERPs for such early processes.

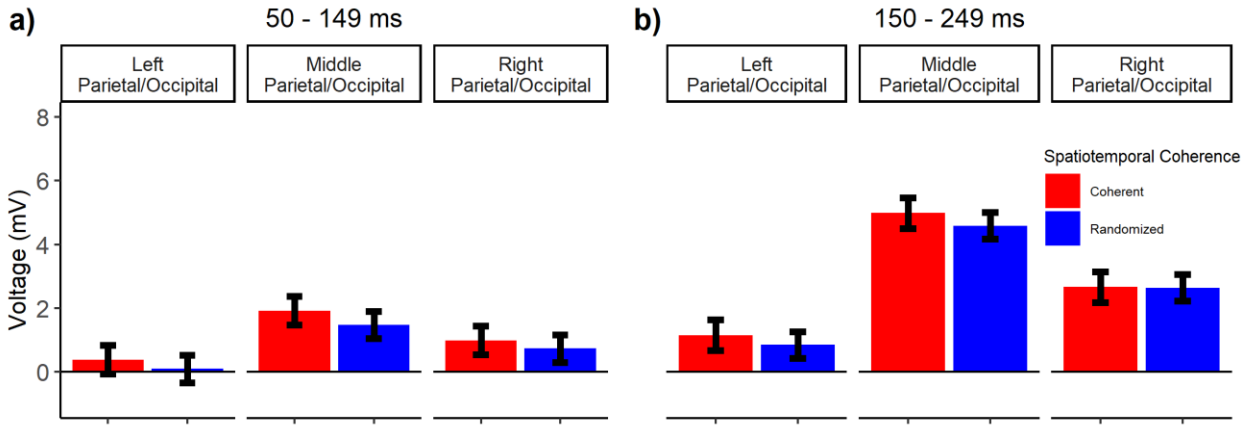


Figure 13. Exp 2: Least square means of amplitudes in response to the ***target*** image at the parietal and occipital sites from a) 50-149 ms and b) 150-249 ms. Amplitudes in the coherent condition were not significantly different from amplitudes in the randomized condition in either the early component or in the analysis of the P200. See the text for details.

Table 4. Exp 2: *Summary of the results for the parietal/occipital regions. Amplitudes were time locked to the target scene.*

Window	Factor	*df*	*F*	*β*	*t*	*p*
50-149 ms	Region	2,192	3.90			.02*
	SC	1,24	0.001	0.33	1.88	.97

	Location	1,24	0.83	0.006	0.05	.37
	Region*SC	2,192	0.22			.80
	Region*Location	2,192	0.91			.40
	SC*Location	1,24	0.02			.89
	Region*SC*Location	2,192	0.09			.92
150-249 ms	Region	2,192	146.98			<.001*
	SC	1,24	1.50	0.24	1.22	.23
	Location	1,24	0.06	-0.05	-0.24	.81
	Region*SC	2,192	0.40			.67
	Region*Location	2,192	0.07			.93
	SC*Location	1,24	0.10			.76
	Region*SC*Location	2,192	0.10			.91

Note: SC = Spatiotemporal coherence of scene sequences. * denotes p < .05.

50-149 ms window.

We averaged the amplitudes within the predetermined time window of 50-149 ms to assess the possibility of early facilitation at the parietal/occipital regions. See Figure 13a). Again, we failed to find support for early facilitation accounts. We observed a significant effect for the region, $F(2,192) = 3.90$, $p = .02$, BF = 2.18; such that Middle Parietal/Occipital electrodes were significantly more positive than both the left, $\beta = -1.46$, $SE = 0.12$, $t = -11.94$, $p < .001$ and right Parietal/Occipital sites, $\beta = -0.62$, $SE = 0.12$, $t = -5.07$, $p < .001$. In addition, the right parietal/occipital electrodes were significantly more positive than the left parietal/occipital electrodes, $\beta = 0.84$, $SE = 0.12$, $t = 6.87$, $p < .001$. If predictions made prior to viewing a scene facilitates early perceptual analysis, then we would expect to find a main effect for the spatiotemporal coherence manipulation. However, consistent with the analyses reported at the frontal and central regions, responses to targets in coherent ($M = 1.09$ $SE = 0.20$) and randomized

(*M* = 0.77, *SE* = 0.20) sequences did not significantly differ. This lack of an effect was supported by a small Bayes factor in favor of the alternative hypothesis, $F(1,24) = 0.001$, $p = 0.97$, BF <.001; and thus support for the null. Thus, we did not find evidence to suggest that predictions made prior to viewing a scene facilitates scene processing very early in perceptual analysis at any of the regions on the scalp. Again, this is consistent with predictions of feed-forward accounts (Serre et al., 2007; VanRullen, 2007). None of the remaining effects were significant. See Table 4 for details.

150-249 ms window.

An analogous linear mixed effects model was conducted on the average amplitudes of the waveforms during the 150-249 ms window to capture the P200 (Harel et al., 2016). Again, Middle Parietal/Occipital electrodes were significantly more positive than both the left, $\beta = -3.79$, *SE* = 0.22, $t = -17.10$, $p < .001$ and right Parietal/Occipital sites, $\beta = -1.66$, *SE* = 0.22, $t = -7.47$, $p < .001$, and the right parietal/occipital electrodes were significantly more positive than the left parietal/occipital electrodes, $\beta = 2.14$, *SE* = 0.22, $t = 9.63$, $p < .001$. McLean et al. (2021) found a reduction in the P200 when observers viewed an expected scene after a series of primes (e.g., the inside of a house vs. the inside of a parking garage after multiple views from outside a house). As such, we expected that the P200 would be larger (i.e., more positive) to targets in the randomized condition. However, as evident in Figure 13b), we found that responses to targets in coherent (*M* = 2.93, *SE* = 0.45) and randomized (*M* = 2.69, *SE* = 0.37) sequences did not significantly differ, $F(1,68.19) = 1.50$, $p = 0.23$, BF = .06. Thus, we found evidence that presenting scenes in coherent sequences facilitated scene perception at the frontal electrode sites, but we did not observe the same pattern over the parietal/occipital region. The discrepancy in the results between what we observed and what was found by McLean et al. (2021) may have been

due to differences in the experimental design. McLean et al. (2021) showed an expected or an unexpected target after multiple primes. The P200 is sensitive to differences in a scene's spatial layout and global scene properties (Harel et al., 2016). Target scenes in the expected and unexpected conditions used in McLean et al. (2021) belonged to different scene categories. Thus, layout features that differed between targets in the expected and the unexpected conditions may have produced differences in the P200, regardless of one's predictions for the upcoming scene. This is unlike what we did here. Namely, we used the exact same target images in both the coherent and randomized conditions, so that any effects due to image characteristics would be held constant. We did not explore the N400 or any N400-like component over parietal/occipital regions given that the N400 is localized over frontal and central electrodes (see Kutas & Federmeier, 2011 for a review).

vERPs to all of the images

Having found both behavioral categorization differences and evidence of facilitation on scene perception at frontal and central electrodes as early as 150-249 ms after the onset of the target scene, we next ran a series of analyses to examine differences in vERPs to all of the scenes presented within coherent and randomized sequences. Averaging the voltage at each time point in the epoch across multiple scenes should provide more precise estimates and should therefore be a more powerful analysis of the P200 component, where we failed to find a significant difference between the coherent and randomized sequences, which was inconsistent with our hypotheses and the results of McLean et al. (2021). We removed the first image within a trial since amplitudes in response to them cannot differ based on spatiotemporal coherence. We also removed behaviorally incorrect trials. Figure 14 shows scalp maps of voltage differences across the conditions. Consistent with the analyses time locked to the target scene on each trial,

amplitudes do not appear to differ in the parietal/occipital electrodes, but amplitudes in coherent sequences appear much more positive than amplitudes in the randomized sequences over frontal and central regions.

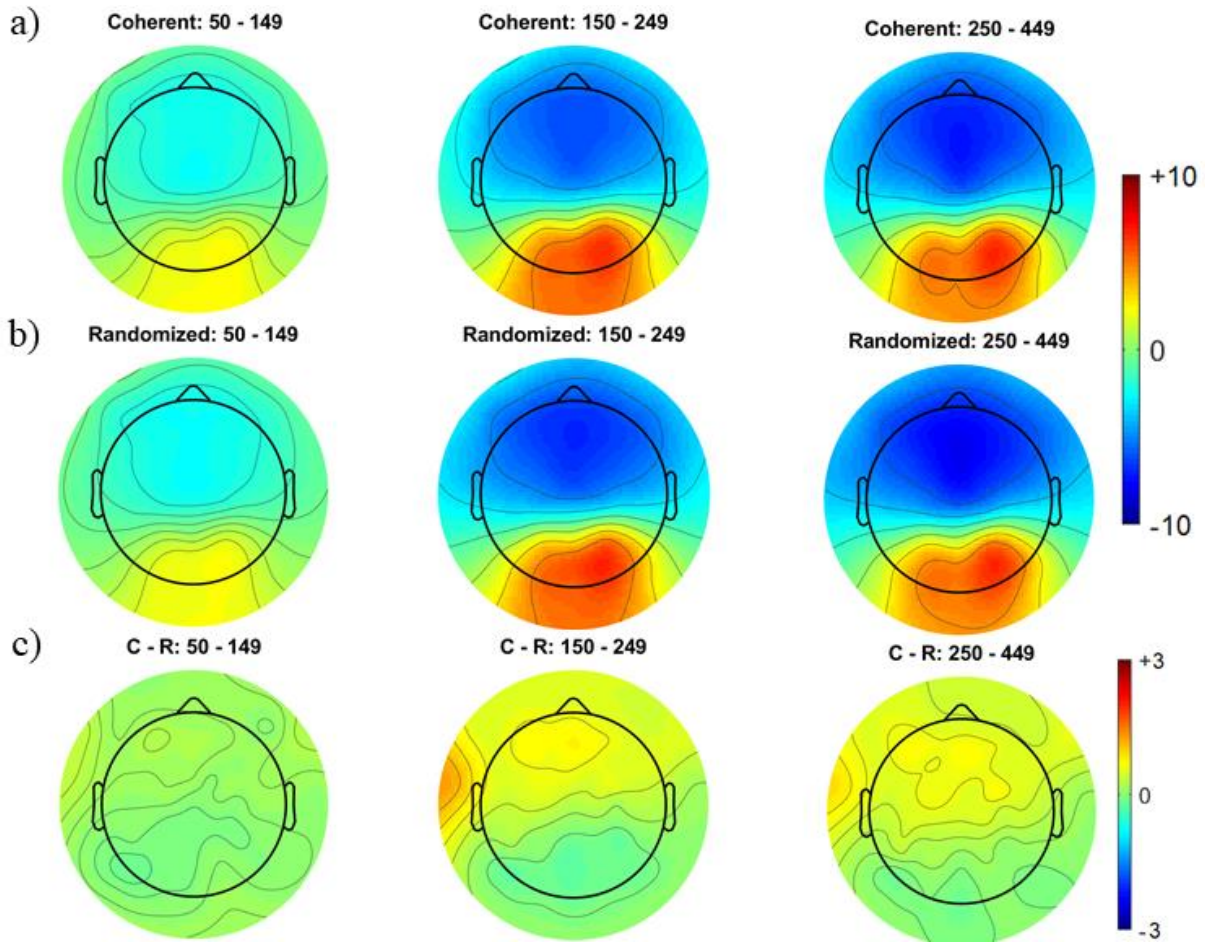


Figure 14. Exp 2: Scalp maps of the mean voltage time locked to the onset of scenes within the a) coherent, b) randomized sequences. The difference between the coherent and randomized conditions are represented in c). Scalp maps do not include behaviorally incorrect trials or responses to the first scene within a trial. Voltage ranged from -10 to +10 microvolts in the coherent and randomized sequences and -3 to +3 in the difference maps.

Frontal and Central Electrodes.

Linear mixed effects models for the early component analyses (50-149 ms), the P200 (150-249 ms), and N400 (250-449) analyses included the same fixed and random effects as were used when we analyzed vERPs after the onset of the target. Least square means from the models are reported in Figure 15.

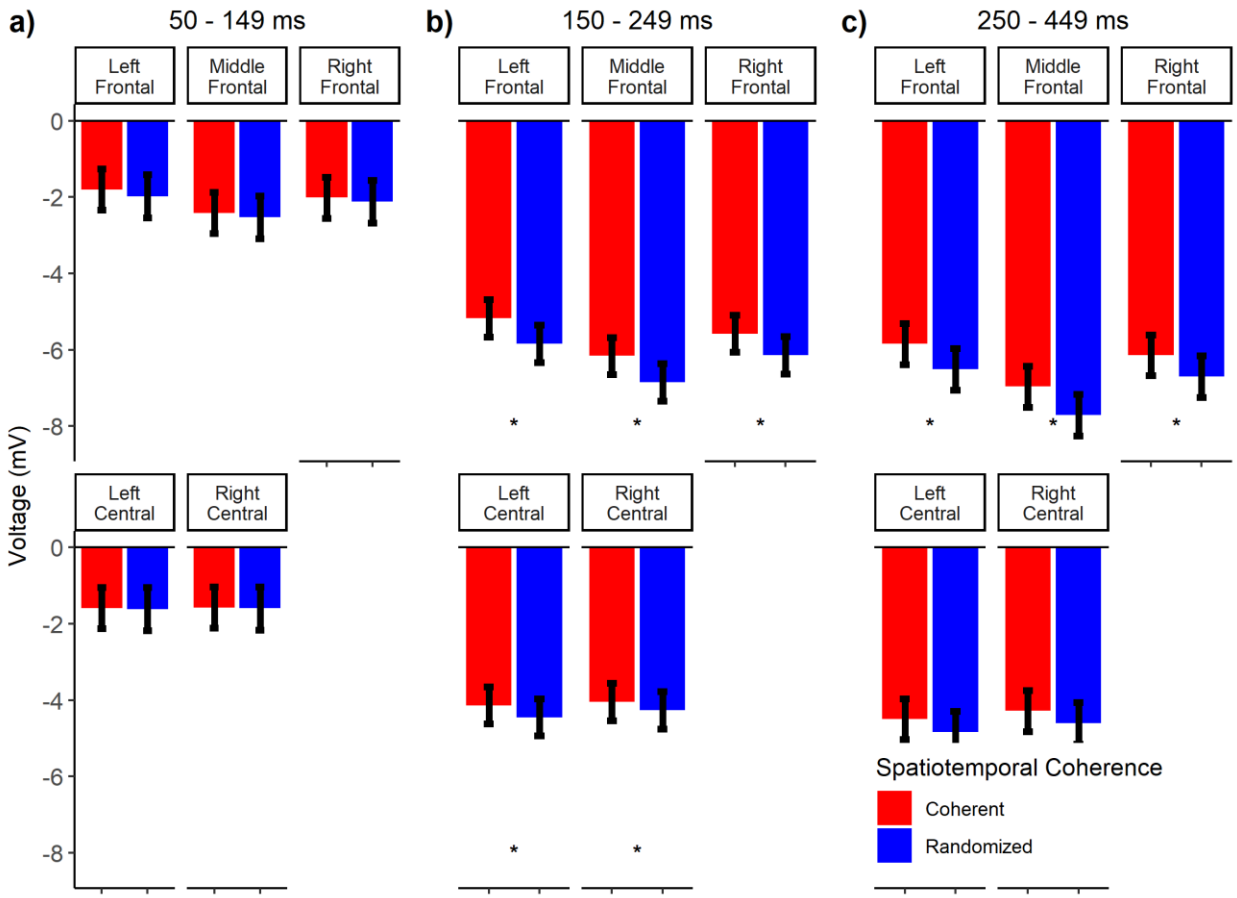


Figure 15. Exp 2: Amplitudes in response to all of the images shown in the coherent and randomized conditions excluding behaviorally incorrect trials and cases when the image was the first scene within a trial at the frontal and central regions. Amplitudes are averaged across the location factor. Responses to images presented in coherent sequences differed significantly from responses to images in randomized sequences in the 150-249 and 250-449 windows.

649

650 Table 5 Exp 2: *Summary of the results for the frontal/central electrodes. Amplitudes were time*
651 *locked to the onset of the scenes in the experiment. Observations from the first image within each*
652 *sequence were removed from the analyses as well as behaviorally incorrect responses to the*
653 *target.*

Window	Factor	*df*	*F*	*β*	*t*	*p*
50-149 ms	Region	4,384	67.13			<.001*
	SC	1,24	1.38	0.09	1.18	.25
	Location	1,24	0.07	-0.02	-0.26	.80
	Region*SC	4,384	0.54			.71
	Region*Location	4,384	0.10			.98
	SC*Location	1,24	1.47			.24
	Region*SC*Location	4,384	0.16			.96
150-249 ms	Region	4,432	151.64			<.001*
	SC	1,24	14.35	0.49	3.78	<.001*
	Location	1,44	19.22	-0.38	-4.38	<.001*
	Region*SC	4,432	1.67			.16
	Region*Location	4,432	0.3			.88
	SC*Location	1,189	1.12			.29
	Region*SC*Location	4,432	0.2			.94
250-449 ms	Region	4,432	185.50			<.001*
	SC	1,24	12.84	-0.42	3.58	<.001*
	Location	1,77	21.14	0.16	4.60	<.001*
	Region*SC	4,432	1.13			.34
	Region*Location	4,432	0.28			.89
	SC*Location	1,147	0.28			.60
	Region*SC*Location	4,432	0.27			.90

654 Note: SC = Spatiotemporal coherence of scene sequences. * denotes $p < .05$.

655

50-149 ms window.

See Figure 15a). Results were consistent with the vERP results time locked to the target, and thus consistent with feed-forward accounts. There was a significant main effect of region, $F(4,384) = 67.13$, $p < .001$, BF > 1,000. Importantly, we found that responses to the scenes presented in coherent sequences ($M = -1.88$ $SE = 0.26$) did not significantly differ from scenes presented in randomized sequences ($M = -1.97$, $SE = 0.27$), $F(1,24) = 1.38$, $p = 0.25$, BF = .004. Thus, we have no evidence from this analysis to suggest that scenes shown in coherent and randomized sequences were processed any differently in the earliest window, as predicted by feed-forward accounts of rapid scene categorization (Serre et al., 2007; VanRullen, 2007; VanRullen & Thorpe, 2001a). See Table 5 for a summary of the results.

150-249 ms window.

See Figure 15b). Our results supported the matching account of facilitation. As before, we found a significant main effect for region, $F(4,432) = 151.64$, $p < .001$, BF >1,000. Importantly, inconsistent with feed-forward accounts, but consistent with the results of Smith and Loschky (2019, Exp 3), we also observed a significant main effect of spatiotemporal coherence, such that vERPs in the coherent sequences ($M = -5.02$, $SE = 0.48$) were significantly more positive than they were to scenes shown in the randomized sequences at this relatively early time point ($M = -5.51$, $SE = 0.48$), $F(1,24) = 14.35$, $p < .001$, BF = 5.47. None of the interactions were significant. See Table 5 for details.

250-449 ms window.

See Figure 15c). Results were consistent with the hypothesis that it is easier to integrate scenes into the current event model when the sequence is spatiotemporally coherent than when it is randomized. Again, we found a main effect of region, $F(4,432) = 185.50$, $p < .001$, BF >

1,000. Importantly, the N400 was significantly more positive for scenes shown in the coherent (*M* = -5.55, *SE* = 0.53) than randomized sequences (*M* = -6.07, *SE* = 0.53), *F*(1,24) = 12.84, *p* = 001, BF = 1.22 as we hypothesized; though the Bayes factor in favor of the alternative hypothesis was notably smaller than what it was when the voltages were time locked to the target. Again, none of the interactions were statistically significant. See Table 5.

Parietal/Occipital Electrodes.

Linear mixed effects models for the early component analysis (50-149 ms) and the analysis of the P200 (150-249 ms) contained the same fixed and random effects as the models time locked to the target. Results from the individual analyses at each time window are shown in Figure 16 and Table 6.

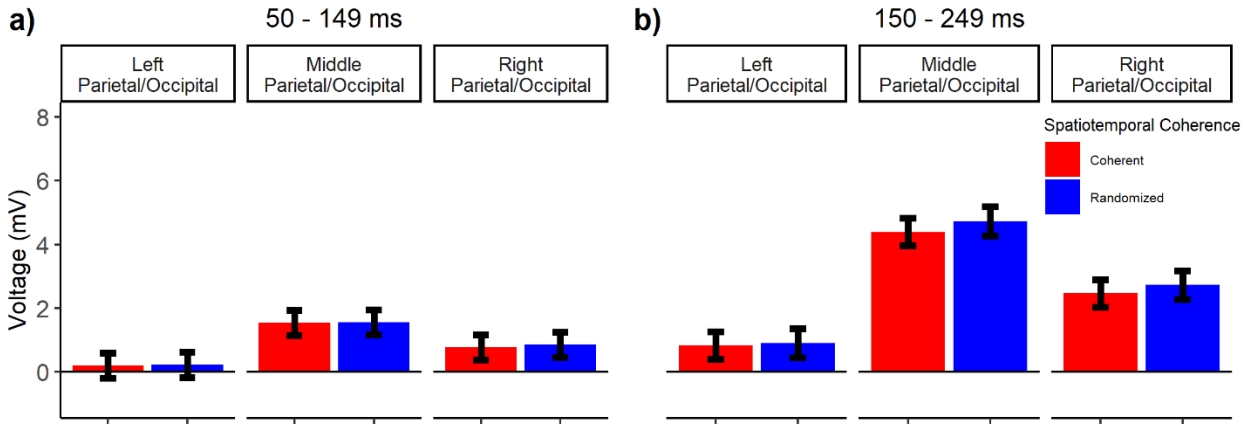


Figure 16. Exp 2: Amplitudes in response to all of the scenes shown in the coherent and randomized conditions at the Parietal/Occipital regions, excluding behaviorally incorrect trials and cases when the image was the first scene within a trial at the parietal/occipital sites. Responses to images shown in the coherent and randomized conditions did not significantly differ in either the early component (50-149) or in the P200.

Table 6. Exp 2: *Summary of the results for the parietal/occipital electrodes. Amplitudes were time locked to the onset of the scenes in the experiment. Observations from the first scene within each sequence and behaviorally incorrect responses to the target were removed from the analysis.*

Window	Factor	*df*	*F*	*β*	*t*	*p*
50-149 ms	Region	2,240	96.76			<.001*
	SC	1,25.13	0.24	-0.04	-0.49	.63
	Location	1,97.16	0.99	-0.08	-1.00	.32
	Region*SC	2,240	0.06			.94
	Region*Location	2,240	0.5			.61
	SC*Location	1,191.47	0.35			.56
	Region*SC*Location	2,240	0.02			.98
150-249 ms	Region	2,264	166.55			<.001*
	SC	1,179.72	1.81	-0.23	-1.65	.18
	Location	1,104.66	1.11	0.18	1.05	.29
	Region*SC	2,264	0.24			.78
	Region*Location	2,264	0.11			.89
	SC*Location	1,209	0.48			.49
	Region*SC*Location	2,264	0.04			.97

Note: SC = Spatiotemporal coherence of scene sequences. * denotes p < .05.

50-149 ms window.

See Figure 16a). The amplitudes of vERPs to scenes were averaged within the predetermined time window of 50-149 ms to assess if predictions for to-be-presented scenes facilitated early perceptual analysis of the scenes recorded at the Parietal/Occipital regions. Consistent with the responses to the target, we observed a significant effect for the region, $F(2,239.99) = 96.76$, $p < .001$, BF > 1,000 such that Middle Parietal/Occipital electrodes were

significantly more positive than both the left, β = -1.34, *SE* = 0.10, *t* = -13.89, *p* < .001 and right Parietal/Occipital sites, β = -0.60, *SE* = 0.10, *t* = -6.28, *p* < .001. The right parietal/occipital electrodes were significantly more positive than the left parietal/occipital electrodes, β = 0.73, *SE* = 0.10, *t* = 7.61, *p* < .001. Consistent with feedforward accounts, amplitudes in coherent (*M* = 0.84, *SE* = 0.18) and randomized (*M* = 0.88, *SE* = 0.18) sequences did not significantly differ, *F*(1,25.13) = 0.24, *p* =0.63, BF = 0.02. Again, this effect was associated with a small Bayes factor in support of the alternative. Thus, we have no evidence from the analysis of the vERPs to suggest that predictions facilitate early perceptual analysis of the scenes shown in coherent sequences. None of the remaining effects were statistically significant. See Table 6 for details.

150-249 ms window.

See Figure 16b). Results of the P200, excluding behaviorally incorrect responses and responses to the first image on a trial were analogous to responses to the target scenes. Again, we observed a significant main effect for region, *F*(2,264) = 166.55, *p* < .001, BF >1,000. We hypothesized that the P200 would be more positive in response to scenes shown in randomized sequences. However, amplitudes in coherent (*M* = 2.56, *SE* = 0.40) and randomized (*M* = 2.79, *SE* = 0.42) sequences were numerically, but not statistically different, *F*(1,179.72) = 1.81, *p* = .18, BF = 0.05. Thus, we found that predictions made prior to viewing a scene affects scene perception in the frontal and central regions within the 150-249 ms window, but not in the parietal/occipital region. None of the remaining effects were statistically significant. See Table 6.

Analysis of vERP divergence

Although there were a priori reasons to believe that differences between the experimental manipulations would be found in the early windows (50-149 ms), the P200, and N400 windows,

we were also interested in characterizing the pattern obtained when no a priori choices were made about the window for the analysis. To do this, we conducted a point-by-point t-test at each moment within the epoch. ERPs from all of the images in the experiment were included in this analysis, excluding responses to the first image on each trial and all of the incorrect behavioral responses. The analysis was conducted three different times for each of the areas of interest (Frontal, Central, and Occipital). To run this analysis, we averaged the voltage across the electrode regions (Left, Middle, and Right) within the frontal, central, and parietal/occipital sites since we did not observe any reliable interactions with region in the component-based analysis. See Tables 5 and 6.

For the permutation-based simulation analysis, we calculated t-statistics comparing the amplitudes of pairs of vERPs (coherent vs. randomized) at each time point within the epoch. Given that this entailed calculating 205 different t-values after downsampling, the possibility of falsely rejecting the null hypothesis is very high. To account for this possibility, we sought a way to distinguish between spurious significant comparisons at random moments in the epoch from meaningful differences that should be characterized by many consecutive time points that also show a significant difference in the amplitude in the same direction. To do this, we calculated the number of consecutive significant time points that would be expected by chance in a series of paired comparisons where no difference was assumed to exist using a Monte Carlo simulation. We randomly assigned data at each time point in the epoch from each participant to one of two arbitrary conditions, regardless of whether the data were from the coherent or randomized sequences. We then ran a paired samples t-test at each time point in the epoch, saved the number of consecutive statistically significant t values in the epoch, and then repeated the entire process 1,000 times for each of the regions (Frontal, Central, and Parietal/Occipital). The time points

754 passing the threshold were deemed to be significant purely by chance since the data were
755 assigned to the two arbitrary conditions at random. The simulation revealed that a run length of
756 15 or greater occurred in 5% of the simulations in the frontal sites, 15 or greater occurred in 5%
757 of the simulations in the central sites, and 18 or greater occurred in 5% of the simulations in the
758 parietal/occipital sites. Thus, these 3 values were used as criteria for considering a given pairwise
759 comparison statistically meaningful. Thus, the number of consecutive significant time points
760 required to surpass the threshold were 15 (frontal, central), and 18 (parietal/occipital). We also
761 calculated Bayes factors for each of the statistical comparisons using the BayesFactor R package
762 (Morey et al., 2018) with a default non-informative Cauchy prior distribution imposed on the
763 effect sizes (Rouder et al., 2012). Bayes factors in favor of the alternative hypothesis greater than
764 3 are generally considered to be substantial evidence in favor of the alternative hypothesis (Kass
765 & Rafferty, 1995). Bayes factors are not influenced by the multiple comparison problem and the
766 inflation of Type I error rates associated with running multiple tests (Dienes, 2016; Simmons et
767 al., 2011).

768 As shown in Figure 17a), results for the frontal regions converged with the component-
769 based approach. Waveforms differed significantly between coherent and randomized sequences
770 beginning at 144 milliseconds post scene onset consistent with matching accounts of facilitation.
771 The effect remained significant until 539 milliseconds. Thus, the effect contained the 150-
772 millisecond time point, previously associated with the time point when people begin to activate
773 higher level representations of scenes (VanRullen & Thorpe, 2001c), as well as the N400,
774 previously associated with semantic access and integration processes (Hagoort et al., 2009; Kutas
775 & Federmeier, 2000). This effect was supported by Bayes factors for each of the individual t-
776 tests. Bayes factors were greater than 3, indicating substantial evidence in favor of the alternative

hypothesis at every time point in the epoch. We obtained analogous results at the central electrode sites. See Figure 17b). Waveforms differed significantly at 164 milliseconds and the difference remained statistically significant until 515 milliseconds. Bayes factors were greater than 3 at every time point in the epoch. Thus, predictions for upcoming scenes begin to facilitate scene perception approximately 150 milliseconds after scene onset.

Results of the analysis of amplitudes for the parietal/occipital electrodes also converged with the component-based approach. ERP waveforms in the coherent and randomized sequences differed significantly at only 2 time points: 222 milliseconds, BF = 1.86 and again at 558 milliseconds, BF =1.87; however, these differences were associated with small Bayes factors in favor of the alternative hypothesis, and they were not followed by at least 18 consecutive significant differences. None of the Bayes factors computed at parietal/occipital regions were greater than 3, indicating anecdotal evidence in favor of both the null and the alternative. The lack of a difference was surprising given that McLean et al. (2021) found differences in vERPs between expected and unexpected scenes in the parietal/occipital region.

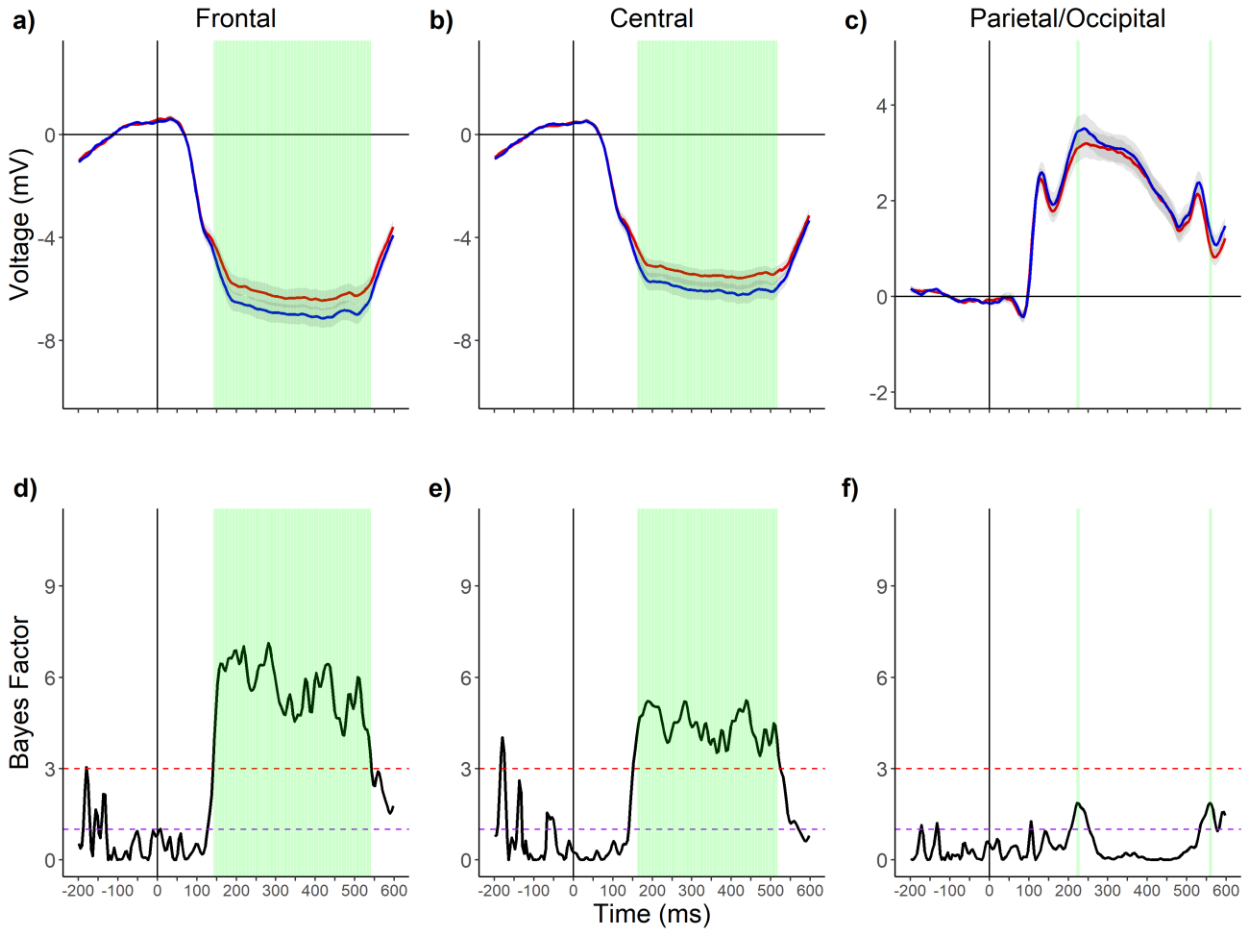


Figure 17. Exp 2: Grand average vERPs time locked to the onset of the scenes at time 0. Scene were presented in either coherent or randomized sequences. Average waveforms at a) Frontal b) Central, and c) Parietal/Occipital sites are on the top row. Bayes factors for each of the paired sample t tests within the epoch for d) Frontal e) Central, and f) Parietal/Occipital electrodes are provided in the bottom row. Green patches represent statistically significant comparisons. Red dashed lines in the Bayes factors plots represent a Bayes factor of 3 and purple lines represent a Bayes Factor of 1 and -1 respectively.

Exploratory analyses of source localization

We hypothesized that we would find differences in vERPs as early as 150 to 249 milliseconds in parietal/occipital regions consistent with a P200, and that differences would not emerge in frontal and central regions until 250 milliseconds or later. In contrast, we found evidence of facilitation over frontal and central regions beginning ~150 milliseconds after the onset of the scenes and very little evidence of facilitation in the P200 at parietal/occipital sites. Note that our predictions were based on the implicit assumption that the scalp regions showing vERP effects were at least globally spatially consistent with the cortical regions that generated them (e.g., vERPs at parietal/occipital locations would be generated by parietal/occipital cortex). However, due to the *inverse problem*, it is well-known that EEG signals at the scalp are often generated by dipoles at distal cortical locations. As such, we next ran an exploratory analysis to identify the possible neural sources of the difference between vERPs over frontal and central regions.

The goal of this analysis was to find some number of equivalent current dipoles whose summed projections most clearly resemble the observed vERP differences between the coherent and randomized conditions. To identify the possible source of the difference between the conditions, we first identified the top 12 independent components that contributed the most to the average scalp distribution between 150 and 249 milliseconds after the onset of the images for each participant. We then fit single equivalent current dipoles to each of the 12 components using the Autofit option within the DIPFIT function in EEGLAB (Oostenveld & Oostendorp, 2002). We did this within a boundary element head model based on the MNI (Montreal Neurological Institute, Quebec, Canada) brain. This resulted in 300 (25 participants X 12 components) independent components across participants. We clustered the 300 components

using a k-means cluster analysis to identify common independent components across participants (Onton & Makeig, 2006). We clustered the independent components using the similarity between each of their vERPs, scalp distributions, and dipole locations. Thus, we assume that sources with similar vERPs, scalp topographies, and dipole locations are due to the same neurophysiology across participants, and that differences within any cluster are due to variations among participants' cortical folds and minor discrepancies in electrode placement from participant to participant. In addition, components that were further than 3 standard deviations from a centroid were categorized into an outlier cluster and omitted from the analyses.

There is no consensus in the literature on what value to set the number of clusters (k), so we followed the instructions outlined by previous researchers who used a similar method (Maruyama et al., 2020). We set k to a value that we considered to yield plausible results in terms of the consistency of characteristics within clusters and distinctiveness between clusters. We checked the estimated centroid of the dipoles for each cluster, the scalp topographies and vERPs of the clusters, and the components that made up each cluster. We also assessed the total number of participants that contributed to each cluster, and the number of components contributed by each participant to each cluster. We found that the 12 brain clusters each contained components from approximately half of the participants (~12). Coordinates of the centroid from each cluster and their labels are shown in Table 7. These 12 clusters were retained as regions of interest.

To compare vERPs in the coherent and randomized conditions, we back projected data from each of the independent components to the individual channels (Zeman et al., 2007). Finally, we repeated the analyses reported in the Analysis of vERP divergence section of this document using the back projected data from each of the components within each cluster. We

assume that back projected vERPs from each of the components that made up each cluster correspond to the activity originating from the centroid of each cluster.

Table 7 Exp 2: *Montreal Neurological Institute (MNI) coordinates and labels of the centroids of independent component clusters. Clusters are in no particular order.*

Cluster Index	Brodmann Area of Centroid	MNI Coordinates (X, Y, Z)	Number of Components	Number of Participants
1	Brodmann Area 21	(-66 -17 -16)	18	10
2	Brodmann Area 40	(42 -47 42)	29	20
3	Brodmann Area 18	(20 -95 8)	23	18
4	Brodmann Area 10	(23 48 18)	16	12
5	Brodmann Area 6	(4 -11 69)	37	23
6	Brodmann Area 37	(42 -60 -14)	23	18
7	Brodmann Area 4	(66 -7 23)	17	12
8	Brodmann Area 7	(0 -66 51)	43	22
9	Brodmann Area 11	(-21 46 -12)	14	11
10	Brodmann Area 40	(-57 -34 44)	20	15
11	Brodmann Area 19	(-42 -80 5)	22	15
12	Brodmann Area 23	(-7 -32 30)	29	18

Visually evoked potentials from the frontal and central electrodes back projected from each of the 12 clusters of independent components are provided in Figures 18 and 19, respectively. This analysis revealed the involvement of two clusters in the brain that contributed to the difference at frontal electrode sites and one cluster that contributed to the difference at central electrode sites between 150 and 249 milliseconds. Consistent with the analysis of the raw waveforms, waveforms at the frontal electrodes began to diverge significantly in Cluster 8 (localized in Brodmann Area 7) at 140 milliseconds and they remained significantly different

until 402 milliseconds. The medial portion of Brodmann area 7, where the centroid of Cluster 8 was localized (See Figure 20), is considered to be the precuneus. The precuneus has a variety of different functions (Cavanna & Trimble, 2006). It is involved in retrieving and elaborating on episodic memories (Daselaar et al., 2008; Hassabis et al., 2007; Shallice et al., 1994), mental navigation (Malouin et al., 2003), encoding and retrieving information about spatial locations (Wagner et al., 2005), and coherence building in discourse comprehension (Ferstl et al., 2008; Moss et al., 2011). It is part of the poster medial network (Inhoff & Ranganath, 2017; Ranganath & Ritchey, 2012), and it is considered the 'hub' of the default mode network (Utevsky et al., 2014). Importantly, it is involved in integrating information across large temporal windows into the event model (Hassabis & Maguire, 2009; Hasson et al., 2008; Stawarczyk et al., 2019), and some researchers have been able to successfully decode scene categories from activity patterns in the precuneus (Choo & Walther, 2016; Ramkumar et al., 2016).

Waveforms at the frontal electrode sites additionally began to significantly diverge in Cluster 12 (Brodmann Area 23) at 144 milliseconds and the difference between the coherent and randomized condition remain statistically significant (all p values < .05) until 539 milliseconds. Brodmann area 23, the location of the centroid of cluster 12, lies within the posterior cingulate. The posterior cingulate plays a very important cognitive role, since its metabolic rate is approximately 40% greater than the brain regions that surround it (Leech & Sharp, 2014; Raichle et al., 2001). Like the precuneus, the posterior cingulate is part of the default mode and posterior medial networks (Inhoff & Ranganath, 2017; Ranganath & Ritchey, 2012), which also includes the parahippocampal cortex and retrosplenial cortex – two areas that are generally considered to be scene-selective regions (Epstein & Kanwisher, 1998; Epstein & Baker, 2019; Harel et al., 2013; Kravitz et al., 2011; O'Craven & Kanwisher, 2000). Neurophysiological evidence and

882 theory suggest that regions in the posterior medial network, such as the precuneus and posterior
883 cingulate are involved in representing meaningful, continuous, contextual information (Hassabis
884 et al., 2007; Hasson et al., 2015; Hasson et al., 2008; Stawarczyk et al., 2019). For instance,
885 according to the Posterior Medial-Anterior Temporal (PM-AT) framework, the precuneus and
886 posterior cingulate may represent the *situation* or the event model itself (Ranganath & Ritchey,
887 2012) [13]. Dipoles from each of the independent components that make up clusters 8 and 12 and
888 their centroid are represented in Figure 20.

889 We assume that observers are better able to develop a coherent event model when they
890 view scenes in a coherent sequence. Consistent with this assumption, we found that differences
891 in amplitudes between 150-249 milliseconds originate from sources that have previously been
892 linked to the construction and application of internal event models (Inhoff & Ranganath, 2017;
893 Ranganath & Ritchey, 2012; Stawarczyk et al., 2019). In addition, it is important to highlight the
894 fact that the cluster analysis conducted on the independent components was data driven. We did
895 not constrain the localization of the dipoles to these regions. Nevertheless, we found evidence to
896 suggest that the differences in the raw frontal vERPs originate from sources implicated in
897 constructing the event model.

898 Waveforms back projected from the independent components that made up both clusters
899 were significantly different at the central region as well, along with responses from an additional
900 cluster. Waveforms in Cluster 8 (Brodmann Area 7) began to significantly differ at 140

[13] Situation models are a special type of event model (Radvansky & Zacks, 2014). Situation models typically refer to event models derived from language. These are in contrast to experience models, which are representations of events derived from interactive experiences. Researchers who work with narratives generally assume that situation models share properties with experience models, and most work is consistent with this assumption Radvansky, G. A., & Zacks, J. M. (2014). *Event Cognition*. Oxford University Press.

milliseconds, and they remained significantly different until 207 milliseconds. Further, waveforms in Cluster 12 (Brodmann Area 23) began to differ significantly from 292 milliseconds and remained significantly different until 476 milliseconds. These comparisons were additionally associated with differences in Cluster 5 (Brodmann Area 6). Waveforms in Cluster 5 significantly differed from 304 milliseconds until 480 milliseconds. Brodmann area 6 is composed of the primary motor and supplementary motor areas. It becomes active when people copy the body movements of others (Molenberghs et al., 2009) or imagine performing a body movement (Raffin et al., 2012). It is also involved in navigation of virtual environments and in the detection of unexpected events when navigating a familiar pathway (Iaria et al., 2008). Thus, it is likely that participants imagined themselves walking through the environment in the coherent sequences as the individual who took the photographs; though none of the participants reported doing so in the post-experiment surveys. Further, though it is typically considered a motor area, it is has also been shown to be involved in updating spatial and verbal mental representations even when participants do not have to make a motor response (Tanaka et al., 2005).

It is also important to note that back projected vERPs from components within clusters that had centroids localized in regions that are known to process low-level visual information (Clusters 3 and 11, which were localized to Brodmann areas 18 and 19, respectively) did not show differences between the coherent and randomized sequences at frontal or central electrodes. This could suggest that facilitation of the event model on scene gist perception does not occur at very early stages in scene processing such as in regions that are involved in initial processing of a scene's low-level features. Instead, the event model may feed-back and influence scene perception at higher levels of scene processing, consistent with matching accounts (Bar,

2004; Bar & Ullman, 1996; Friedman, 1979; Mudrik et al., 2010; Palmer, 1975a; Schendan, 2019).

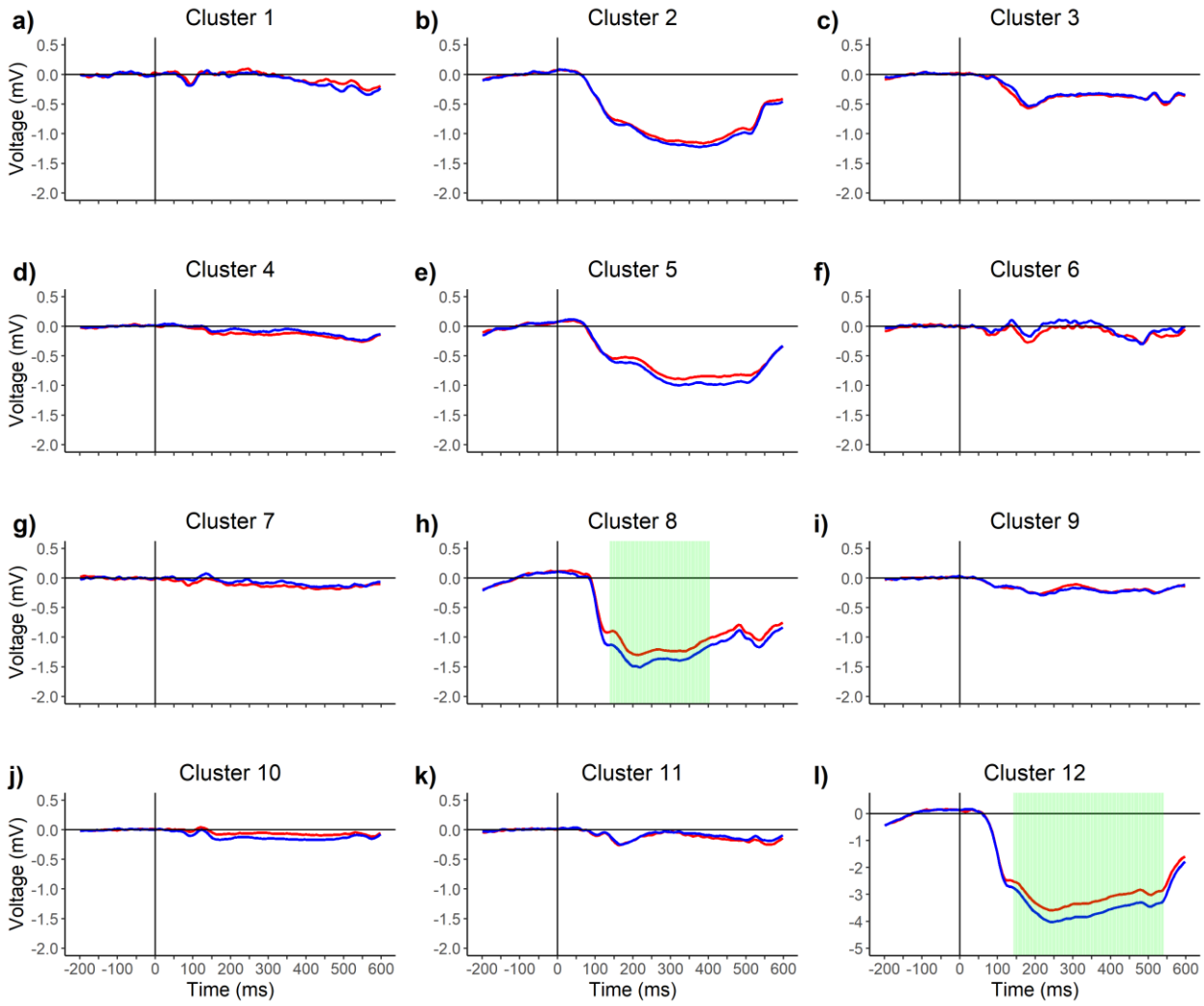


Figure 18. Exp 2: Grand average waveforms at *frontal* electrodes time locked to the onset of the scenes back projected from each of the 12 clusters of independent components. The cluster indices are represented in a) through l). Waveforms in response to scenes in the coherent sequences are in red, and waveforms from the randomized sequences are in blue. Green patches represent significant comparisons.

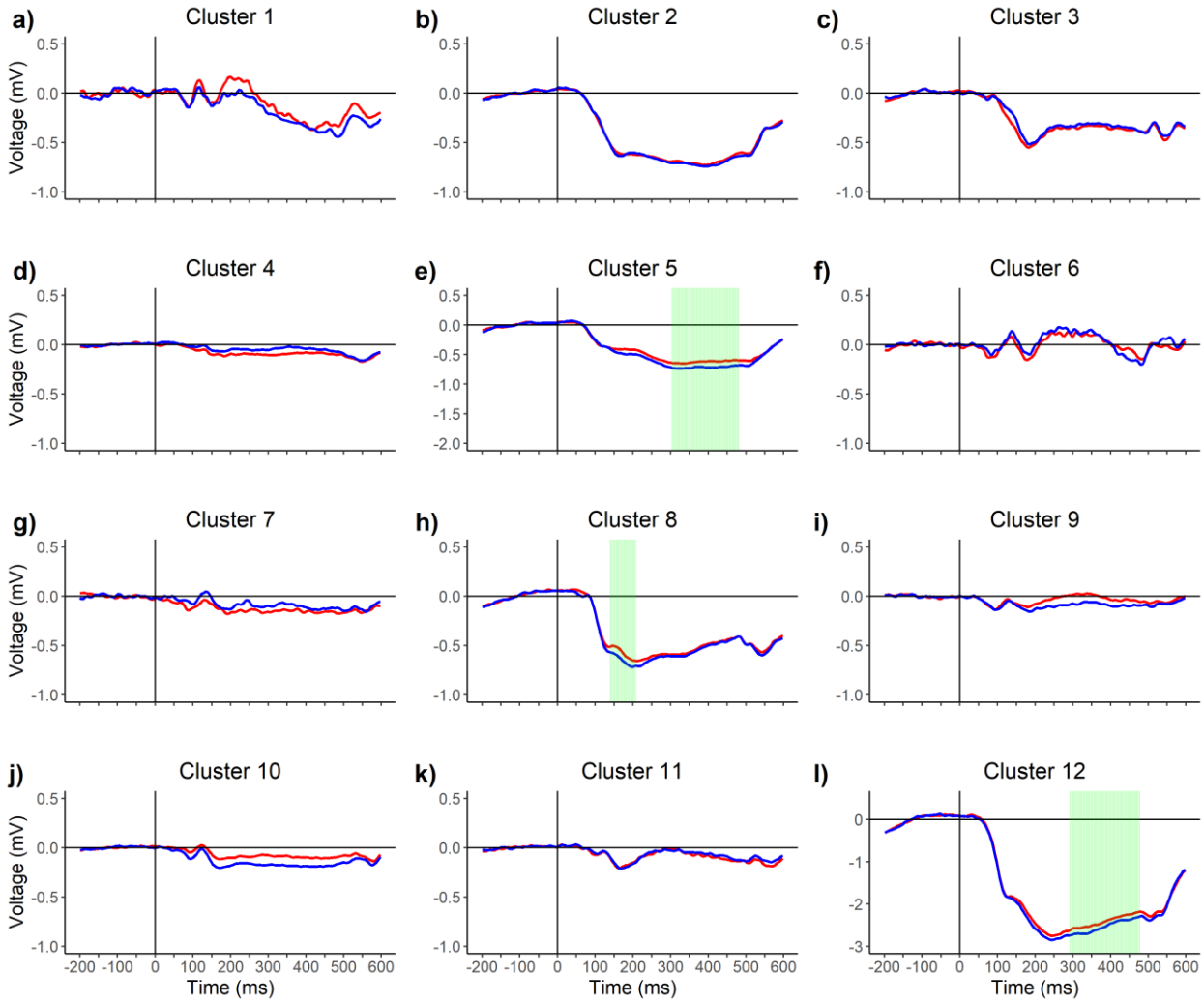


Figure 19. Exp 2: Grand average waveforms at *central* electrodes time locked to the onset of the scenes back projected from each of the 12 clusters of independent components. The cluster indices are represented in a) through l). Waveforms in response to scenes in the coherent sequences are represented in red, and waveforms in response to scenes shown in the randomized sequences are in blue. Green patches represent significant comparisons.

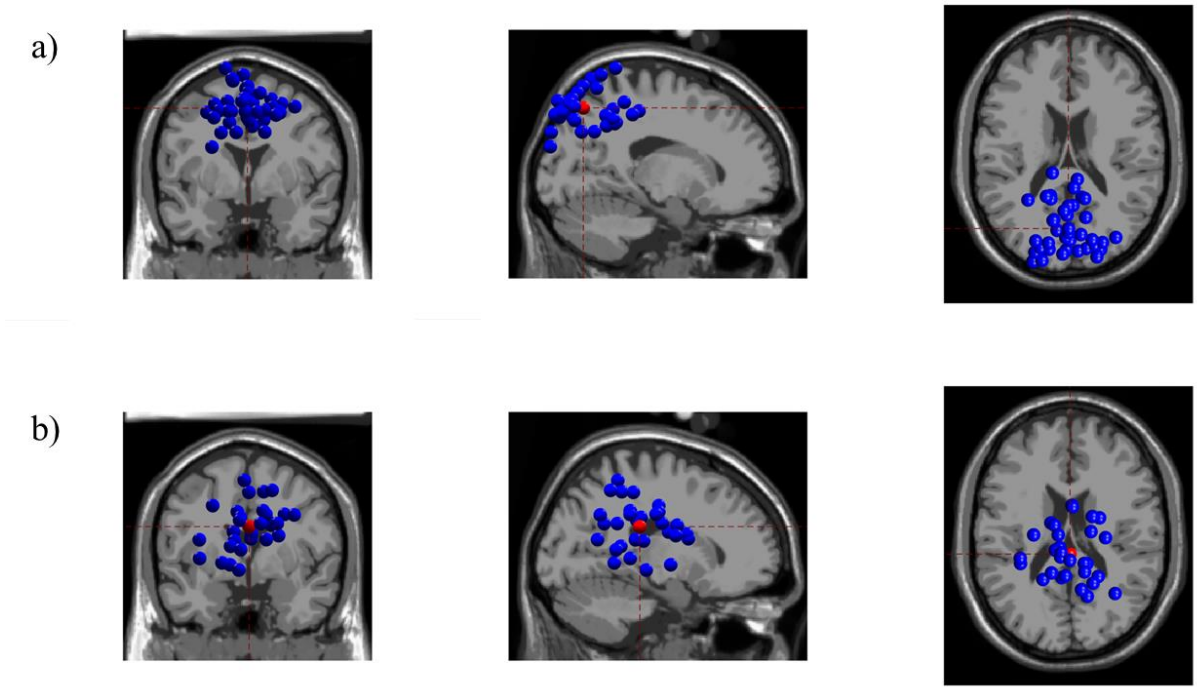


Figure 20. Exp 2: Clusters of sources of independent components for all subjects across trials and conditions for clusters a) 8 (Brodmann area 7: Precuneus) and b) 12 (Brodmann are 23: Posterior Cingulate Cortex).

Changes in vERPs within a trial

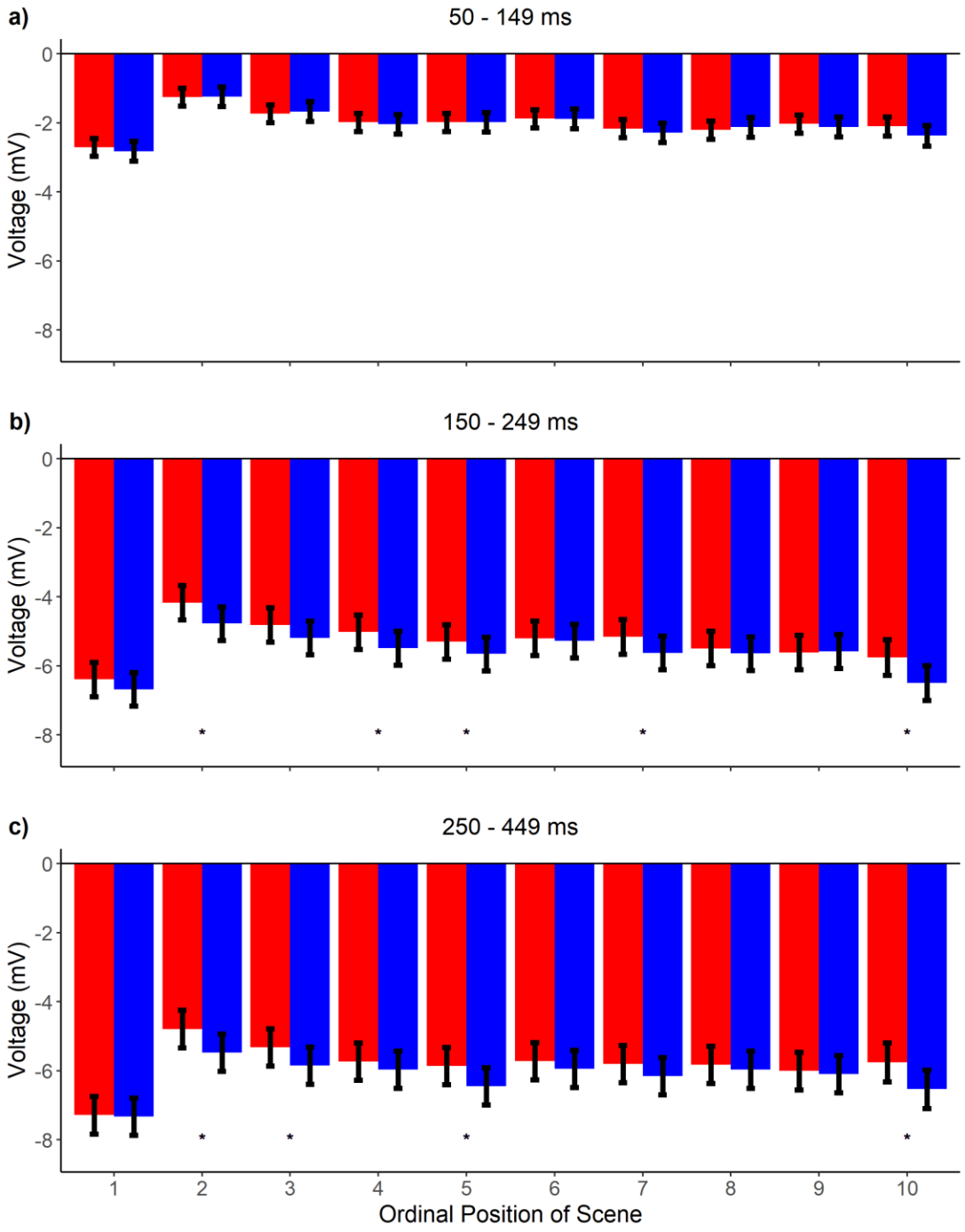
As mentioned previously, SPECT assumes that viewers lay the foundation of the event model after the first image is shown in a coherent sequence on a trial. Consistent with the hypothesis that viewers create an event model of the coherent sequences, our source localization results suggest that differences in vERPs between the coherent and randomized conditions originate from brain regions that have previously been implicated in construction of the event model (Hasson et al., 2015; Inhoff & Ranganath, 2017; Stawarczyk et al., 2019). Further, we assume that scene categories become more predictable as the event model is constructed over time within a trial. Behaviorally, we found that both the ability to predict upcoming scenes (See Figure 5 for the off-campus image sequences) and the ability to categorize the target scenes (See Figure 9) improved as a function of the ordinal position of the target, though the increase in

behavioral categorization was unexpectedly larger for scenes shown in the randomized sequences. As such, we also explored changes in vERPs as a function of the ordinal position of the target scenes (1-10). We hypothesized that amplitudes would not differ between the coherent and randomized sequences on the first scene within a trial, but they would afterwards. We removed all behaviorally incorrect trials before running the analysis. Results are reported in Tables 8 and 9 for each of the 3 windows (50-149, 150-249, 250-449), and least square means of amplitude for each window are shown in Figures 21 and 22. We probed all significant interactions by adjusting with a Bonferroni correction.

Frontal and Central Electrodes

Linear mixed effects models of amplitudes within each time window included the region of the electrodes, the main effect of spatiotemporal coherence, the location the images were photographed (on-campus vs. off-campus) as the previous analyses of the different vERP components. We also included the ordinal position of the scenes (1-10) on each trial, and all of their interactions as fixed effects. The effect of spatiotemporal coherence, the location the images were photographed, the region, the ordinal position of the scenes, and their interactions varied for each participant as random effects.

We attempted to run the analysis by treating the ordinal position of the scenes on each trial as a continuous predictor of amplitude; however, the majority of those models failed to converge with the complex random effect structure we specified. As evident in Figures 21 and 22, issues with convergence were likely due to the nonlinear relationship between the ordinal position of the scenes and voltage. As such, we treated the ordinal position of the scenes as a categorical predictor in the analysis at each window for consistency. Model output is provided in Table 8.



982

983 *Figure 21.* Exp 2: Amplitudes at each ordinal position (1-10), excluding behaviorally incorrect

984 trials. Responses to images in coherent sequences are in red, and responses to images in

985 randomized sequences are in blue.

986

Table 8. Exp 2: *Summary of the results for the frontal/central electrodes. Amplitudes were time locked to the onset of the scenes in the experiment in the 1st through the 10th position.*

Window	Factor	*df*	*F*	*p*
50-149	Region	4.4408	109.99	<.001*
	SC	1,24	0.24	.63
	Location	1,25	0.6	.45
	Ordinal Position	94,408	71.22	<.001*
	Region*SC	44,408	0.3	.88
	Region*Location	44,408	0.31	.87
	SC*Location	14,408	0.004	.98
	Region* Ordinal Position	364,408	0.69	.92
	SC* Ordinal Position	94,410	0.81	.60
	Location* Ordinal Position	94,394	3.66	<.001*
	Region*SC*Location	44,408	0.11	.98
	Region*SC* Ordinal Position	364,408	0.29	.99
	Region*Location* Ordinal Position	364,408	0.37	.99
	SC*Location* Ordinal Position	94,408	1.69	.08
	Channels*SC*Location* Ordinal Position	364,408	0.29	.99
150-249	Region	4.4408	468.65	<.001*
	SC	1,24	4.26	.04*
	Location	1,25	13.76	.001*
	Ordinal Position	94,408	71.42	<.001*
	Region*SC	44,408	1.35	.25
	Region*Location	44,408	0.50	.74
	SC*Location	14,408	3.61	.06
	Region* Ordinal Position	364,408	0.69	.92
	SC* Ordinal Position	94,410	2.77	.003*
	Location* Ordinal Position	94,394	7.40	<.001*
	Region*SC*Location	44,408	0.20	.94

	Region*SC* Ordinal Position	364,408	0.32	.99
	Region*Location* Ordinal Position	364,408	0.34	.99
	SC*Location* Ordinal Position	94,408	3.89	<.001*
	Channels*SC*Location* Ordinal Position	364,408	0.31	.99
250-449	Region	4,408	529.54	<.001*
	SC	1,24	3.58	.07
	Location	1,25	28.68	<.001*
	Ordinal Position	94,408	59.19	<.001*
	Region*SC	44,408	1.46	.21
	Region*Location	44,408	0.48	.75
	SC*Location	14,408	0.98	.32
	Region* Ordinal Position	364,408	0.71	.90
	SC* Ordinal Position	94,410	2.68	.004*
	Location* Ordinal Position	94,394	5.95	<.001*
	Region*SC*Location	44,408	0.19	.94
	Region*SC* Ordinal Position	364,408	0.21	.99
	Region*Location* Ordinal Position	364,408	0.32	.99
	SC*Location* Ordinal Position	94,408	2.40	.01*
	Channels*SC*Location* Ordinal Position	364,408	0.29	.99

Note: SC = Spatiotemporal coherence of scene sequences. * denotes $p < .05$.

50-149 ms window.

Least square means from each model are shown in Figure 21a). We again observed a significant main effect of region. $F(4,4408) = 109.99$, $p < .001$, BF > 1,000. Consistent with feed-forward accounts, we did not observe an effect for spatiotemporal coherence, $F(1,24) = 0.24$, $p = .63$, BF = 0.002. However, we did observe a significant main effect for the ordinal position of the scene, $F(9, 4408) = 71.22$, $p < .001$, BF > 1,000 such that the average amplitudes

997 in the 50-149 ms window of the first (M = -2.76, SE = 0.27) image were significantly more
998 negative than amplitudes in response to the remaining images after we applied a Bonferroni
999 correction to account for multiple comparisons [Second (M = -1.25, SE = 0.27); Third (M = -
1000 1.70, SE = 0.27); Fourth (M = -2.01, SE = 0.27); Fifth (M = -1.98, SE = 0.27); Sixth (M = -1.88,
1001 SE = 0.27); Seventh (M = -2.22, SE = 0.27); Eighth (M = -2.17 , SE = 0.27); Ninth (M = -2.07,
1002 SE = 0.27); Tenth (M = -2.24, SE = 0.27)]. This effect could be expected, given that prior work
1003 using analogous RSVP paradigms have found that amplitudes are reduced when one stimulus
1004 immediately follows another (Lu et al., 1992; Woodman, 2010). In addition, there was also a
1005 significant interaction between the location the images were photographed and the ordinal
1006 position of the scene, $F(9,4394) = 3.66$, $p < .001$, BF > 1,000, such that amplitudes to off-campus
1007 sequence were significantly more positive than on-campus sequences when the image was the
1008 10^{th} image on trial, β = -0.55, SE = 0.17, t = 3.35, p = .008. This effect was not observed for any
1009 of the other positions within a trial after we adjusted the p values for the family wise error rate
1010 using a Bonferroni correction. None of the remaining effects were significant. See Table 8.
1011 ***150-249 ms window.***
1012 See Figure 21b). Again, we observed a significant main effect for region, $F(4,4407) =$
1013 468.65, $p < .001$, BF >1,000. More importantly and consistent with matching accounts of
1014 facilitation, we also observed a significant main effect for spatiotemporal coherence, $F(1,24) =$
1015 4.26, $p < .04$, BF = 2.92; such that amplitudes where significantly more positive in coherent (M =
1016 -5.29, SE = 0.48) than randomized (M = -5.64, SE = 0.48) sequences. Importantly, the main
1017 effect of spatiotemporal coherence depended upon the ordinal position of the scene, as evident
1018 from a significant interaction between spatiotemporal coherence and the ordinal position,
1019 $F(9,4410) = 2.77$, p = .003, BF > 1,000. In addition, this two-way interaction was moderated by a

significant three-way interaction with the location the images were photographed (on-campus vs. off-campus), $F(9,4408) = 3.89$, $p < .001$, BF = 2.98. As shown in Figure 21, amplitudes did not significantly differ between the coherent and randomized conditions when the image was the first scene on a trial, and this was true for on-, $\beta = -0.08$, $SE = 0.24$, $t = -0.32$, $p = .75$ and off-campus sequences $\beta = 0.40$, $SE = 0.24$, $t = 1.64$, $p = .10$. This is consistent with the hypothesis that the first scene lays the foundation of the event model in working memory, and thus should not be facilitated by one's predictions for an upcoming scene. Even though the ability to predict the scene categories within each sequence decreased as a function of the image on a trial in the on-campus sequences in Experiment 1 (See Figure 5), amplitudes were significantly more positive in the coherent sequences in the 2nd, 4th, 5th, and 7th positions, and they were numerically greater in the remaining positions. For off-campus images, amplitudes in coherent sequences were significantly more positive than responses to images in randomized sequences in the 2nd, 7th, and 10th positions (all Bonferroni corrected p values < .05). Amplitudes were numerically more positive in the remaining 7 positions. Thus, we observed evidence that the event model facilitated matching processes after the first scene on a trial consistent with predictions of SPECT. The remaining effects were consistent with the effects observed in the previous window.

250-449 ms window.

As shown in Figure 21c), our results supported the hypothesis that scenes presented in coherent sequences were easier to integrate into an event model than scenes presented in randomized sequences. We again observed a significant three-way interaction between spatiotemporal coherence, location, and the ordinal position of the scenes on the trial, $F(9, 4408) = 2.40$, $p = .01$, BF = 3.11. Amplitudes of the N400 did not significantly differ in the first position for on-, $\beta = -0.07$, $SE = 0.28$, $t = -0.25$, $p = .80$ or off-campus sequences, $\beta = 0.16$, $SE =$

0.28, $t = 0.58$, $p = .56$. As mentioned previously, this is likely because the first scene of a sequence lays the foundation of the event model in working memory. Amplitudes differ after the first scene. Amplitudes were significantly more positive in coherent on-campus sequences in the 5th and 7th positions, and they were numerically more positive in the remaining positions. Coherent off-campus sequences were significantly more positive than randomized off-campus sequences in the 2nd, 3rd, 6th, and 10th positions and they were numerically more positive in the remaining positions except for the 9th. Thus, we observed evidence of facilitation after the first scene in both on- and off-campus sequences consistent with predictions of SPECT. The remaining significant interactions were consistent with the previous 3 analyses. See Table 8 for details.

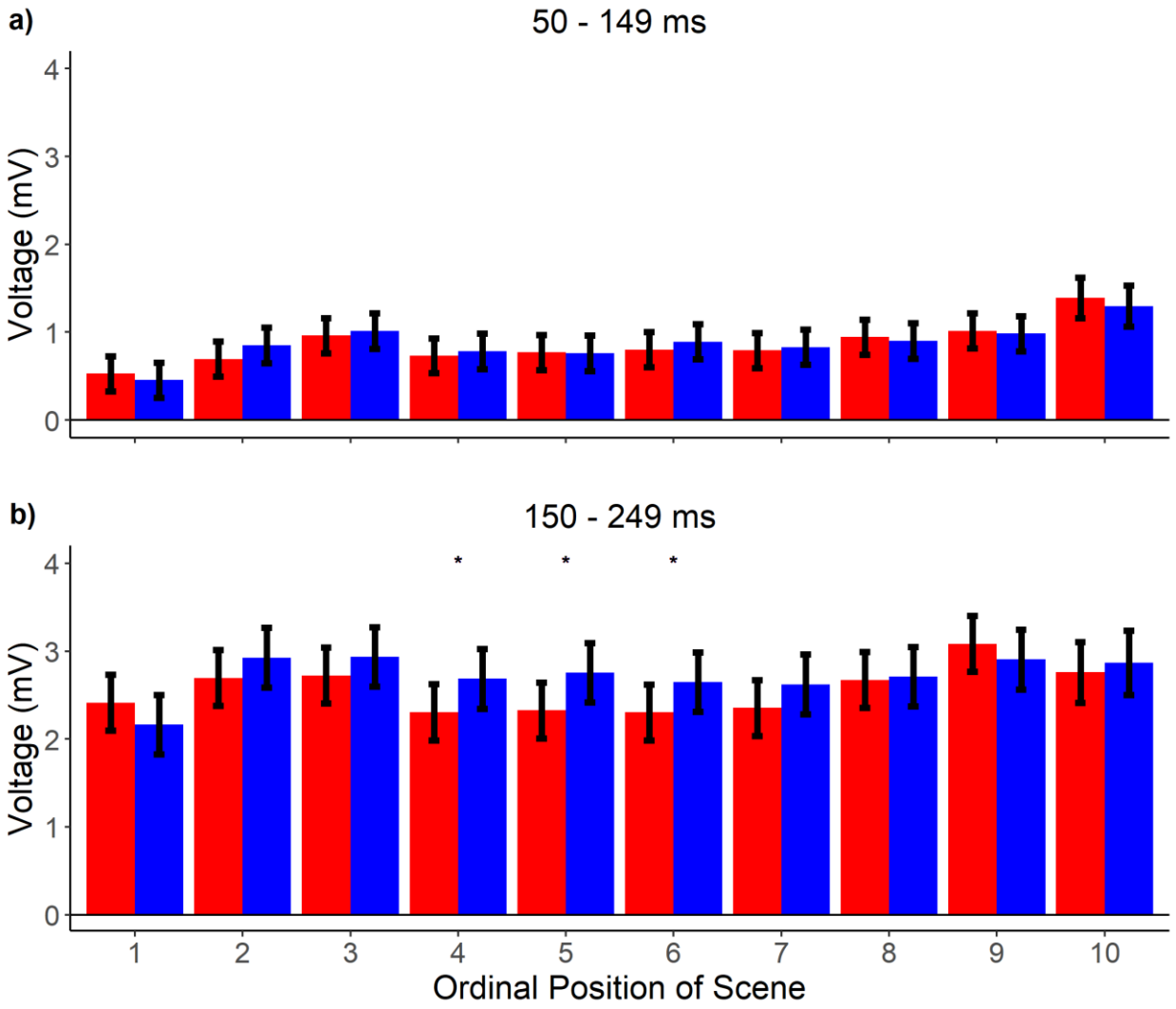


Figure 22. Exp 2: Amplitudes in response to each ordinal position (1-10) of a scene on a trial, excluding behaviorally incorrect trials. Amplitudes in response to scenes in coherent sequences are in red, and amplitudes in randomized sequences are in blue.

Table 9. Exp 2: *Summary of the results for the parietal/occipital electrodes. Amplitudes were time locked to the onset of the images in the experiment in the 1st – 10th position.*

Window	Factor	*df*	*F*	*p*
50-149	Region	22,640	372.40	<.001*
	SC	1,24	0.02	.90
	Location	1,36	4.52	.04
	Ordinal Position	92,640	9.36	<.001*
	Region*SC	22,640	0.03	.97
	Region*Location	22,640	4.01	.02*
	SC*Location	12,640	1.45	.23
	Region* Ordinal Position	182,640	1.11	.34
	SC* Ordinal Position	92,643	0.39	.94
	Location* Ordinal Position	92,640	1.80	.06
	Region*SC*Location	22,640	0.64	.53
	Region*SC* Ordinal Position	182,640	0.38	.99
	Region*Location* Ordinal Position	182,640	0.51	.96
	SC*Location* Ordinal Position	92,640	1.86	.05
	Channels*SC*Location* Ordinal Position	182,640	0.33	.99
150-249	Region	22,640	1100.72	<.001*
	SC	1,24	1.60	.22
	Location	1,36	2.97	.09
	Ordinal Position	92,640	13.08	<.001*
	Region*SC	22,640	2.35	.10
	Region*Location	22,640	0.22	.81
	SC*Location	12,640	1.90	.17
	Region* Ordinal Position	182,640	1.00	.45
	SC* Ordinal Position	92,643	1.65	.10
	Location* Ordinal Position	92,640	2.11	.03*
	Region*SC*Location	22,640	0.39	.68

Region*SC* Ordinal Position	182,640	0.23	.99
Region*Location* Ordinal Position	182,640	0.34	.99
SC*Location* Ordinal Position	92,640	1.79	.07
Channels*SC*Location* Ordinal Position	182,640	0.22	.99

50-149 ms window.

See Figure 22a). Consistent with the previous analyses of the amplitudes in the parietal/occipital regions, we observed a significant main effect for the region, $F(2,2640) = 372.40$, $p < .001$, BF > 1,000. Consistent with feed-forward accounts, we did not observe a significant main effect for spatiotemporal coherence, $F(1,24) = 0.02$, $p = .90$, BF = 0.002. However, we did a significant effect for the ordinal position of the scenes, $F(9,2640) = 9.36$, $p < .001$, BF >1,000. Consistent with the previous analyses at the parietal/occipital region, the effect of spatiotemporal coherence was the same at each ordinal position, as evident from a nonsignificant interaction between spatiotemporal coherence and the ordinal position of the scene (1-10) on each trial, $F(9,2643) = 0.39$, $p = .94$, BF < 1,000. We also found a marginally significant three-way interaction between spatiotemporal coherence, the location where the scenes were photographed, and the ordinal position of the images on a trial, $F(9.2640) = 1.86$, $p = .05$, BF = 1.05; however, this effect was associated with a small Bayes factor and none of the differences were statistically significant after we probed this interaction and corrected each test for multiple comparisons with a Bonferroni adjustment. Thus, our results are consistent with feed-forward accounts, which propose that amplitudes between coherent and randomized sequences should not differ. None of the remaining interactions were statistically significant. See Table 9.

150-249 ms window.

Results were surprisingly different from the previous analyses when we focused our analysis on the P200 over parietal/occipital regions. As evident in Figure 22b). We observed a significant main effect for the region, $F(2,2640) = 1100.72$, $p < .001$ as before. We did not observe a significant main effect for spatiotemporal coherence, $F(1,24) = 0.92$, $p = .35$, BF = .005, but we did for the ordinal position of the scenes, $F(9,2640) = 4.99$, $p < .001$, BF > 1,000. We predicted that the P200 would be more positive when the sequences were randomized than coherent. Consistent with this hypothesis, we found a significant two-way interaction, supported by a Bayes factor greater than 3, between the spatiotemporal coherence manipulation and the ordinal position of the scenes, $F(9,2640) = 2.11$, $p = .02$, BF = 3.79. This effect was consistent with results from McLean et al. (2021). As hypothesized, the P200 did not differ significantly between the coherent and randomized sequences in the 1st position on the trial**,** $\beta = 0.25$, $SE = 0.24$, $t = 1.03$, $p = .30$. However, **t**here was a significantly larger P200 for scenes in the randomized sequences in the 4th, 5th, and 6th positions, and the P200 was numerically larger for scenes in the randomized sequences in the remaining positions, except the 9th position. Thus, inconsistent with feed-forward accounts, but consistent with SPECT, we found evidence to suggest that predictions made prior to viewing a scene may facilitate the P200, which is the earliest marker to reflect scene-specific processing (Harel et al., 2016). None of the remaining interactions were statistically significant. See Table 9.

Exploratory analyses of image predictability and image similarity

We also explored an alternative explanation for the advantage we observed for scenes shown in coherent sequences. It is possible that the facilitation effects we observed in the component-based analyses and cluster simulation analyses may have been purely the result of

facilitation from low-level visual information shared between successive scenes rather than feedback from the event model (Bar & Biederman, 1998; Shafer-Skelton & Brady, 2019; Sperber et al., 1979). Afterall, sequential pairs of photographs shown in the coherent sequences share many more features and objects with one another than sequential pairs of scenes shown in randomized sequences. This is apparent from visually inspecting the scenes provided in Figure 1. The two office photographs in Figure 1 share many of the same colors, lighting conditions, objects, and spatial layout. This is in comparison to the parking lot and stairwell scenes shown in Figure 1. Furthermore, even as one navigates from an office into a hallway many more features across the pair of scenes will be similar than between randomly paired scene categories in the randomized sequences.

We assume that various feature detectors become activated along the ventral visual pathway when a scene is perceived. If a subsequent scene activates the same, or many of the same, feature detectors, then the combined activity in sensory memory from the first and second scenes may facilitate identification of the second scene. As mentioned previously, prior work has suggested that scene layout priming may be due to such a mechanism (Shafer-Skelton & Brady, 2019). Thus, regardless of predictions informed from one's event model, scenes may have been perceived more efficiently because primes and targets may have activated similar feature detectors along the ventral visual pathway (Bar & Biederman, 1998).

In contrast, both predictability and visual similarity between successive scenes may have facilitated scene processing. Prior work has also found that the predictability of a word in a sentence (DeLong et al., 2005; Van Petten & Luka, 2012) or panel in a comic strip (Coderre et al., 2020) positively correlates with the N400 such that voltage increases as pictures become more predictable. For instance, Coderre et al., (2020) showed participants comic strips that

contained a critical panel that was either highly predictable, moderately predictable, or unexpected. They found that N400 amplitudes increased as a function of the predictability of the panel.

Predictability has also been shown to facilitate the P200 component as well. McLean et al. (2021) found that the P200 is more positive when viewers encounter an unexpected scene after a sequence of coherent scenes using a similar paradigm to the one we used here; though it is unclear if the facilitation effects observed by McLean et al. (2021) were due to predictability driven from feed-back mechanisms in the event model or facilitation from the overlap of visual features in the feed-forward sweep.

To examine these alternative possibilities in our study, we correlated how predictable each target was in the sequence using data from Experiment 1 with the average amplitudes in response to the target scenes in Experiment 2, while controlling for the effects of visual similarity in a partial correlation analysis. Likewise, we also examined partial correlations between visual similarity and voltage controlling for image predictability (as explained below). This was done to examine the unique contribution of image predictability and image similarity on vERP amplitudes. We quantified image predictability by averaging prediction accuracy in Experiment 1 for each image, so predictability ranged from 0% to 100% predictable. We hypothesized that image predictability and image similarity would correlate positively with amplitudes in the frontal and central regions and negatively in parietal/occipital regions. These analyses were limited to the target because participants in Experiment 1 were only asked to predict the category of the missing target. The design of Experiment 2 was yoked from Experiment 1, such that the target scenes participants predicted in Experiment 1 were the same scenes that a different sample of participants categorized in Experiment 2.

We quantified visual similarity between the target and the image that immediately preceded the target using the spatial envelope model (Oliva & Torralba, 2001). Windowed spectral information can be used to categorize scenes at the basic and superordinate level, and it has been shown to be diagnostic of global scene properties such as naturalness and openness (Greene & Oliva, 2009b), which can be decoded from neural activity in scene selective areas (Cichy et al., 2017; Harel et al., 2013; Park et al., 2011). Spatial envelope features also correlate with performance of neural decoders in scene-selective regions (Watson et al., 2017; Watson et al., 2014). To quantify image similarity, we extracted spectral energies from four fixed windows for each image of size 1024 X 768. Within each window of size 256 X 192, we extracted spectral information by calculating the responses to Gabor filters at four spatial frequencies and eight orientations. Filter responses were concatenated to obtain an 8 X 4 X 4 X 4 = 512 feature vector for each image. The reciprocal of the Euclidean distance between pairs of images was used as a measure of image similarity between the prime immediately before the target and the target. Given its strong positive skew, we took the natural log of image similarity prior to entering it into the analyses. We also removed behaviorally incorrect categorization responses before running the analyses.

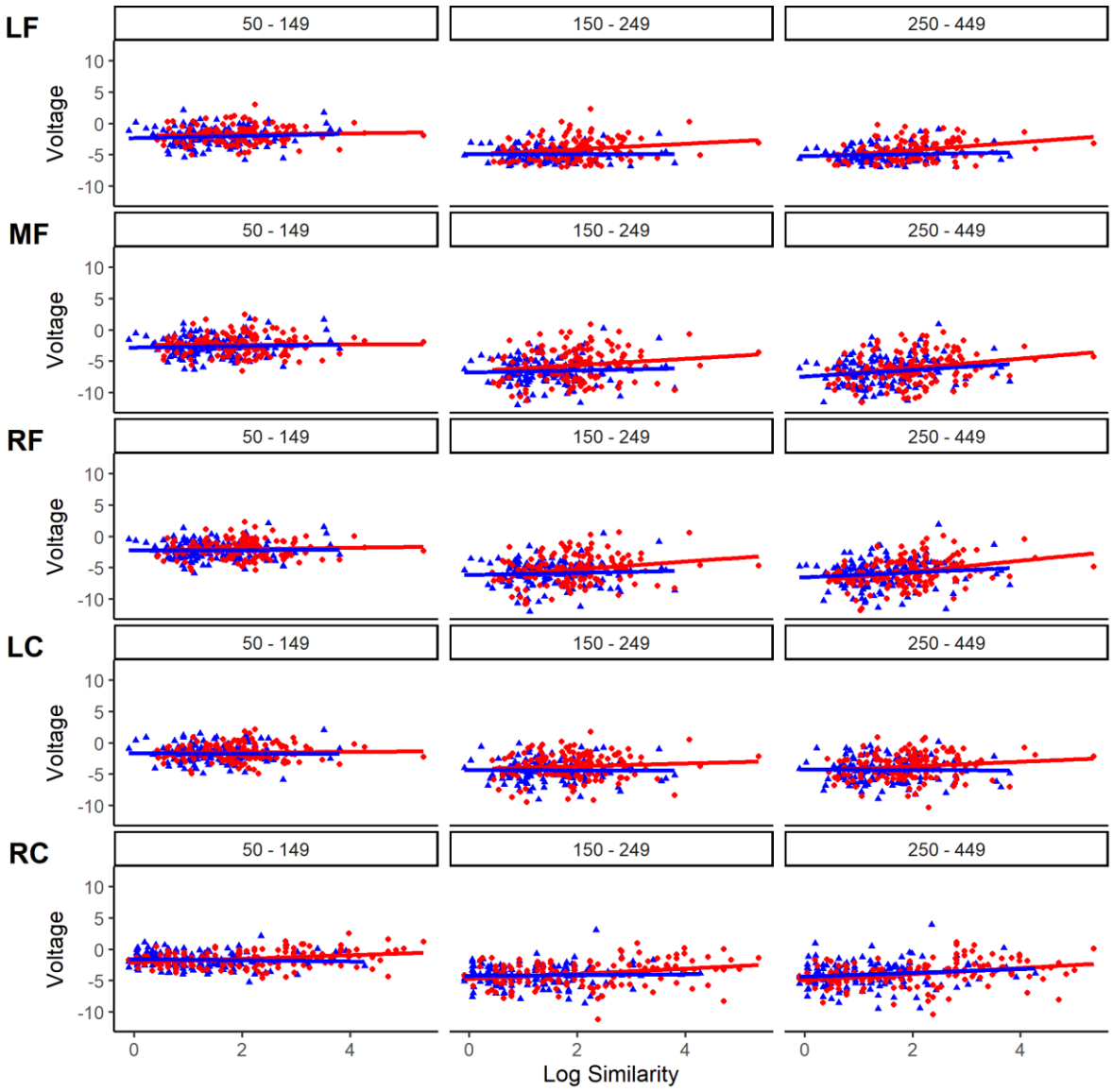
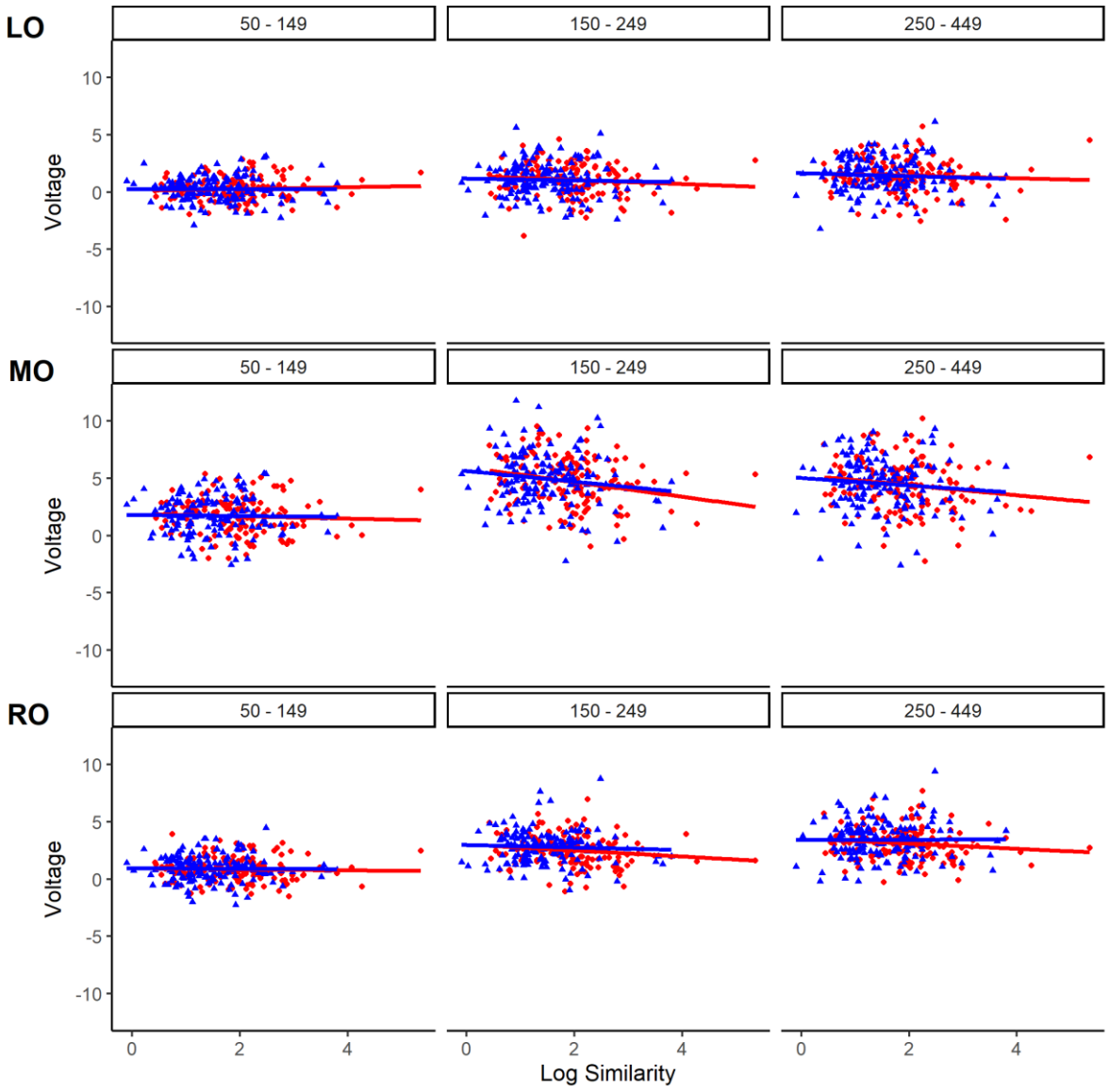


Figure 23. Exp 2: Scatterplots between the log of ***image similarity*** and voltage at LF) left, MF) middle, and RF) right ***frontal*** as well as LC) left and RC) right ***central*** regions.

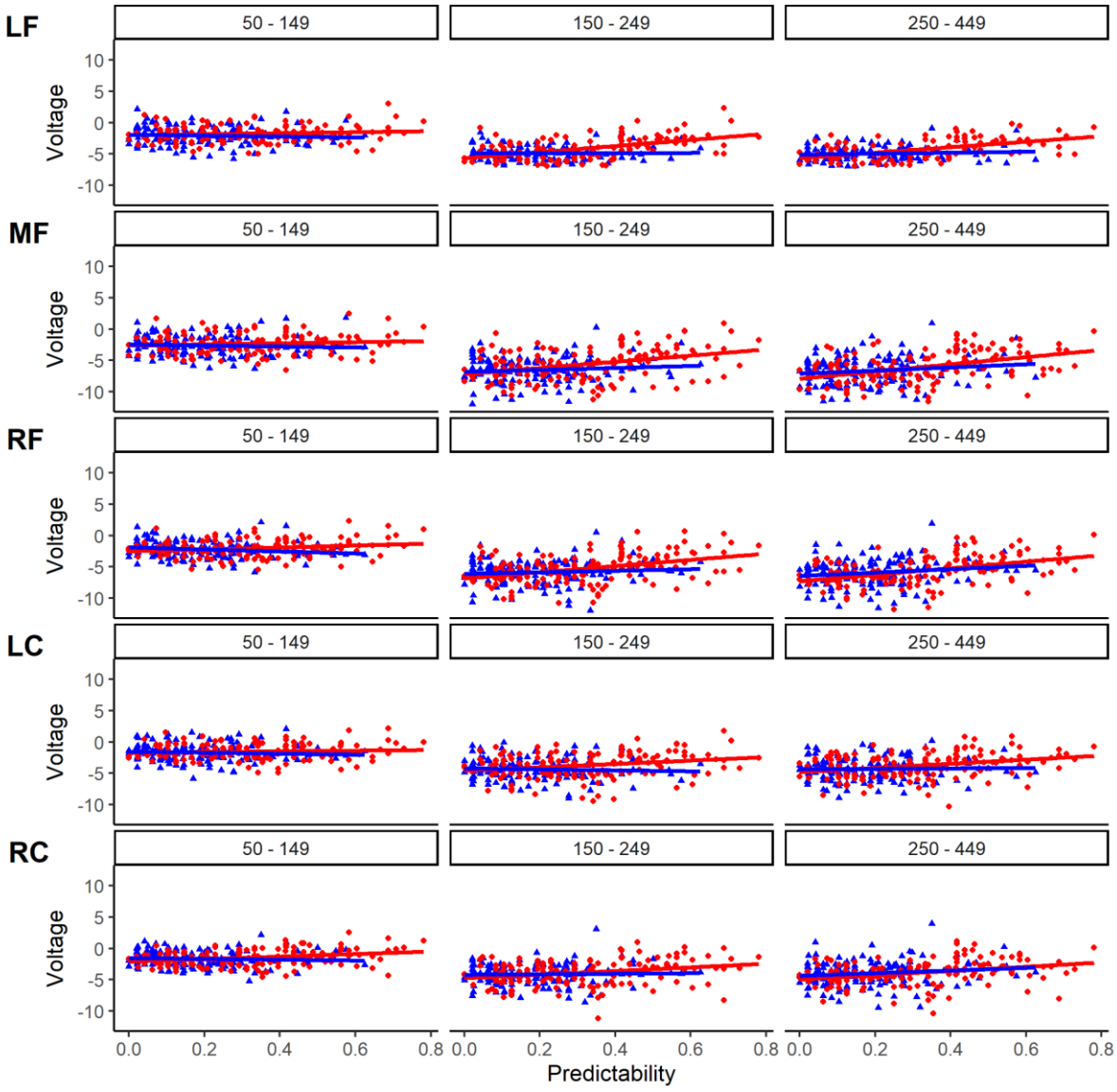


1171

1172 *Figure 24*. Exp 2: Scatterplots between the log of ***image similarity*** and voltage at LO) left, MO)

1173 middle, and RO) right ***parietal/occipital*** regions.

1174



1175

1176 *Figure 25*. Exp 2: Scatterplots between ***image predictability*** and voltage at LF) left, MF) middle,

1177 and RF) right ***frontal*** as well as LC) left and RC) right ***central*** regions.

1178

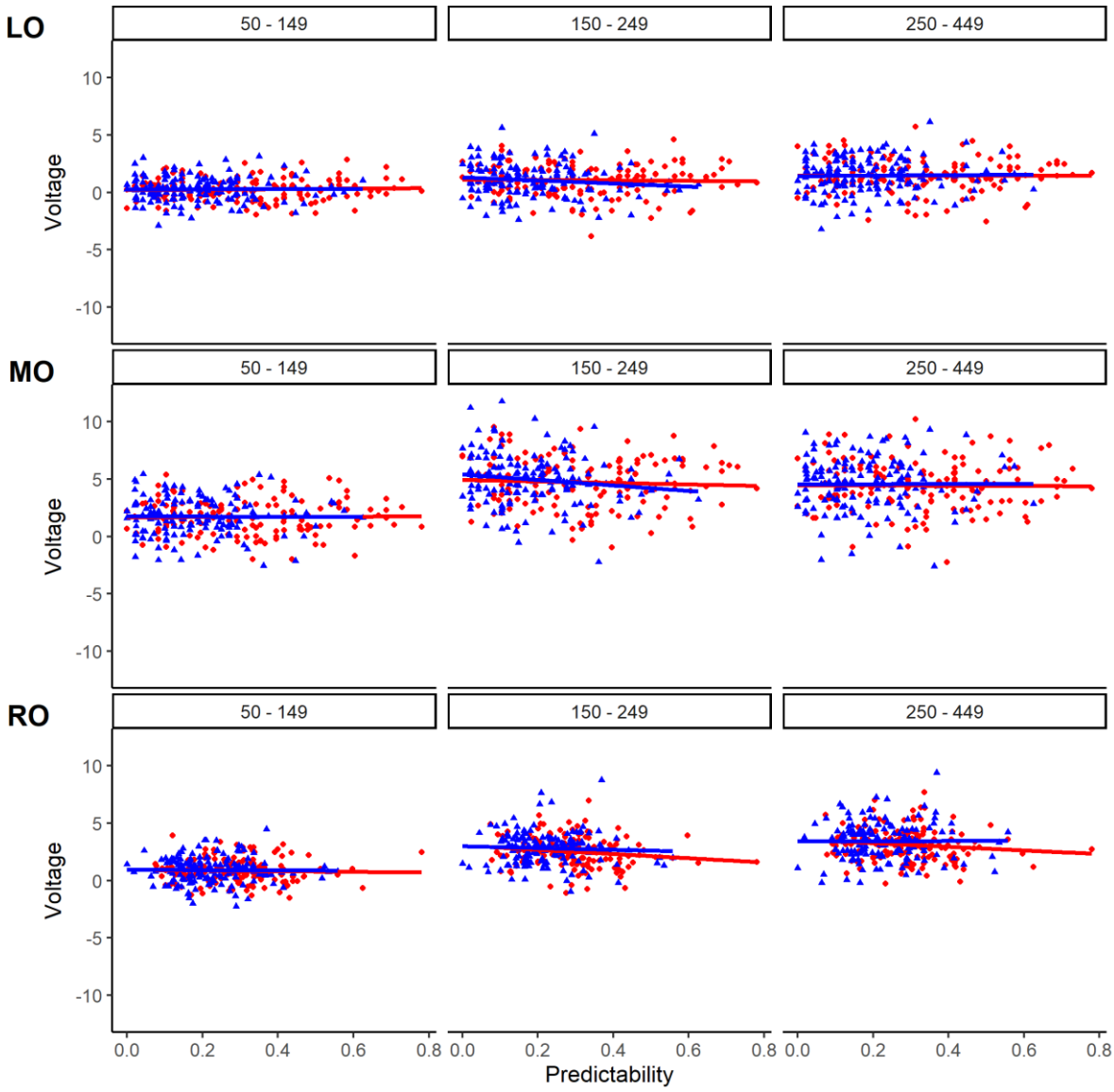


Figure 26. Exp 2: Scatterplots between ***image predictability*** and voltage at LO) left, MO) middle, and RO) right ***parietal/occipital*** regions.

Table 10. Exp 2: *Partial correlation coefficients between the log of image similarity and mean amplitude, controlling for image predictability. Partial correlation coefficients between image predictability and the mean amplitudes within each of the time windows (50-149), (150-249), and (250-449) controlling for the effect of log of image similarity.*

Region	Spatiotemporal Coherence	50-149		150-249		250-449	
		Log Sim	Pred	Log Sim	Pred	Log Sim	Pred
Left Frontal	Randomized	0.11	-0.10	0.00	0.01	0.07	0.07
	Coherent	0.02	0.10	0.08	0.47*	0.17*	0.44*
Middle Frontal	Randomized	0.08	-0.09	0.05	0.04	0.08	0.13
	Coherent	-0.02	0.11	0.06	0.33*	0.13	0.37*
Right Frontal	Randomized	0.07	0.07	0.04	0.07	0.07	0.14*
	Coherent	0.01	0.09	0.12*	0.36*	0.21*	0.38*
Left Central	Randomized	0.02	-0.08	0.01	-0.06	-0.03	0.03
	Coherent	0.01	0.07	0.02	0.23*	0.07	0.27*
Right Central	Randomized	0.02	-0.07	-0.01	0.03	0.04	0.13
	Coherent	0.02	0.06	0.09	0.24*	0.18*	0.27*
Left Parietal/ Occipital	Randomized	0.00	0.01	-0.08	-0.11	-0.07	0.04
	Coherent	0.05	0.03	-0.11*	0.00	-0.07	0.02

Middle Parietal/ Occipital	Randomized	-0.02	0.00	-0.11*	-0.10	-0.11*	0.04
	Coherent	-0.06	0.03	-0.23*	0.01	-0.18*	0.03
Right Parietal/ Occipital	Randomized	-0.01	-0.01	-0.04	-0.03	-0.03	0.09
	Coherent	-0.06	0.12	-0.15*	0.02	-0.13*	0.05

Note: Log Sim = Log Similarity between the target image and its immediately preceding prime. Pred = Image Predictability from Experiment 1. * denotes $p < .05$.

We first assessed if targets shown in coherent sequences shared more features with their primes than targets shown in randomized sequences. As we expected, the log of image similarity was larger in coherent ($M = 1.86$, $SE = 0.02$) than randomized ($M = 1.44$, $SE = 0.02$) sequences, $t(1295) = 16.74$, $p < .0001$, BF > 1,000; therefore, the benefit we observed in coherent sequences may have been due to the fact that those targets had more similar looking primes than scenes shown in randomized sequences.

Scatterplots are shown in Figures 23-26, and partial correlations are provided in Table 10. Partial correlations in Table 10 show that image similarity began to correlate significantly with amplitudes, after controlling for the influence of image predictability, in the left frontal region starting at 50-149 milliseconds and it remained significant into the N400 (See Figure 23). Thus, image similarity could at least partially account for spatiotemporal coherence facilitation over frontal regions. Interestingly, as evident in Figure 24, image similarity negatively correlated with amplitudes in the left, middle, and right parietal/occipital regions in the P200. This result is consistent with the results reported by McLean et al. (2021) and our examination of changes in voltage across time (See Figure 22). Partial correlations were also similar in both the coherent

and randomized sequences, though voltage in right parietal/occipital electrodes were the notable exception. Thus, the visual similarity between target and prime facilitated scene processing above and beyond influences originating from the event model.

However, also as evident in Table 10, image similarity alone was unable to explain why amplitudes were more positive beginning around 150 milliseconds in frontal and central regions. Specifically, image predictability determined from participants in Experiment 1, controlling for the influence of image similarity, correlated with voltage in all 5 of the frontal and central regions (See Figure 25), but not in any of the parietal/occipital regions (See Figure 26). Thus, predictions made prior to viewing a scene facilitates scene gist perception above and beyond the visual similarity between prime and target. This effect is consistent with hypotheses generated from SPECT, namely that processes involved in the back-end event model feedback to influence front-end rapid scene categorization. Specifically, predictions facilitated processing in both the 50-149 ms window and in the window of the N400. Predictions may facilitate processes involved in matching the structural description to representations stored in memory and those involved in mapping the current information with the contents of the event model.

In addition, the partial correlations were larger in the more predictable, coherent sequences (See Figure 25). The reason why image predictability correlated with voltage in the coherent, but not the randomized sequences is unclear. Predictability was only slightly above chance in the randomized condition; however, some images had very high values of predictability (see Figure 4). Scenes in the coherent condition were both globally predictable (e.g., predicting a hallway after multiple views of an office) within each sequence as well as locally predictable (e.g., predicting a hallway after a single view of an office); whereas scenes in the randomized condition could only be locally predictable, possibly due to educated guessing

(e.g., knowing there are always 10 scenes in a sequence, from 5 categories, and counting how many of each category have been shown, to guess the probability of a given category appearing next). Scenes that were locally predictable in a randomized sequence still violated the sequence's global coherence; therefore, amplitudes were larger (more negative) than when scenes were both globally and locally predictable.

Neural Decoding of Image Categories

We also explored the temporal dynamics of scene gist categorization and how emerging categorical representations contributed to behavioral responses. Although univariate analyses of event related potentials are useful in identifying the time course of facilitation of the event model on scene processing, univariate techniques do not provide information about *how* scene category representations emerge over time. Univariate comparisons between waveforms time locked to scenes presented in coherent and randomized sequences do not tell us if voltage at a given point in time contains information that can discriminate between different scene categories. As such, we examined the amount of category-relevant information in the vERPs using a time-resolved decoding procedure on the vERPs (see the Neural Decoding Section of this document for details about how we conducted the neural decoding analysis).

Research has found that categorical representations can be decoded above chance level performance from neural signals within 40 milliseconds, and the first peak in decoding accuracy typically emerges between 100 and 250 milliseconds after scene onset (Cichy et al., 2017; Greene & Hansen, 2020; Ramkumar et al., 2016). Our results were consistent with these previous findings. As evident in Figure 27, the average decoding accuracy was essentially at chance level performance prior to the onset of the target scene in all 3 regions (Occipital: 12.72%; Central: 12.71%; Frontal: 12.65%). Accuracy rose significantly above chance for both

the coherent and randomized sequences in all three regions around 50 milliseconds, as confirmed from a one sample t-test at each time point in the epoch, and decoding accuracy peaked earlier for parietal/occipital regions (121 milliseconds) than central (144 milliseconds) and frontal (152 milliseconds) regions. Thus, information about the scene category being viewed becomes available approximately 50 milliseconds after scene onset. This is consistent with prior work that used temporally resolved decoding (Cichy et al., 2017; Greene & Hansen, 2020; Ramkumar et al., 2016). Interestingly, peaks were at very similar time points between the coherent and randomized sequences; however, the accuracy of the neural decoders were significantly greater when the sequence was coherent than when it was randomized after the initial peak in decoding accuracy. As shown in Figure 27, this was supported by Bayes Factors greater than 3 around 200 milliseconds in the epoch at frontal, central, and at parietal/occipital electrodes. Thus, consistent with hypotheses generated from SPECT, the visual system represents scenes more effectively when they are shown in coherent sequences.

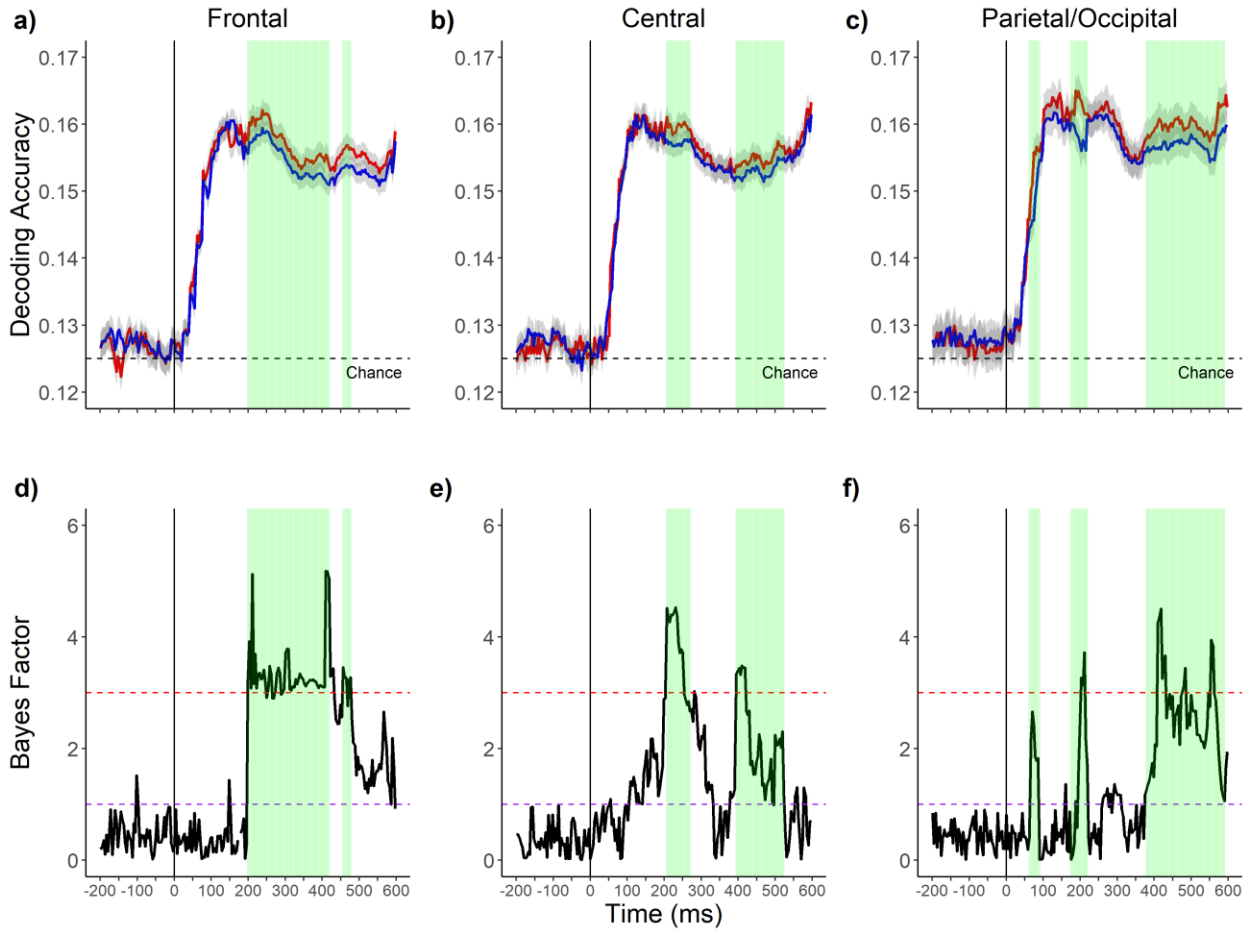


Figure 27. Exp 2: ***Decoding accuracy*** as a function of time in the epoch for the a) Frontal, b) Central, and c) Parietal/Occipital regions. Bayes Factors for each of the paired sample t tests within the epoch for d) Frontal e) Central, and f) Parietal/Occipital electrodes are provided in the bottom row. Green patches represent clusters of statistically significant comparisons. Red dashed lines in the Bayes Factors plots represent a Bayes Factor of 3 and purple lines represent a Bayes Factor of 1 and -1 respectively. Error ribbons correspond to 95% confidence intervals around the means.

To assess this possibility statistically, we ran a linear mixed effects model to determine if decoding accuracy was greater in coherent than randomized sequences. Such an effect could suggest that participants represented the scenes shown in coherent sequences more effectively.

Prior to running the analysis, we averaged decoding accuracy after the onset of the scene (0-600 ms) for each participant. Models included the fixed effects of the channel region (left, middle, and right frontal and parietal/occipital regions, and left and right central regions), the location where the images were photographed (on-campus vs. off-campus), the effect of spatiotemporal coherence (coherent vs. randomized), and all their interactions. Like the analyses reported above, models contained the participant intercepts as well as the random slopes of spatiotemporal coherence, the image location, the electrode regions, and all of their interactions as random effects (i.e., the maximal model).

Consistent with previous explorations of the spatiotemporal dynamics of scene categorization (Greene & Hansen, 2020), we found a significant main effect for region, $F(7, 696) = 3.57$, $p < .001$, BF > 1,000; such that decoding accuracy was significantly greater in the more posterior parietal/occipital electrodes [Left Occipital ($M = 0.16$, $SE = .001$); Middle Occipital ($M = 0.15$, $SE = .001$); Right Occipital ($M = .15$, $SE = .001$)] than anterior regions [Left Central ($M = 0.15$, $SE = .001$); Right Central ($M = 0.15$, $SE = .001$); Left Frontal ($M = 0.14$, $SE = .001$); Middle Frontal ($M = 0.14$, $SE = .001$); Right Frontal ($M = .14$, $SE = .001$)]. Importantly, we also found that decoding accuracy was significantly greater in the coherent ($M = .16$, $SE = .001$) than the randomized ($M = .15$, $SE = .001$) sequences, $F(1,24) = 77.33$, $p < .0001$, BF = 263.75. Even though this was a small difference numerically, it was supported by a large Bayes Factor in favor of the alternative hypothesis. As shown in Figure 28, almost all of the participants showed the benefit. This result is important because it suggests that participants had a more detailed representation of scenes shown in coherent sequences. Importantly, this result is also inconsistent with feed-forward models of scene perception (Serre et al., 2007; VanRullen, 2007; VanRullen & Thorpe, 2002), which would predict that there would be no difference between scenes shown

in coherent and randomized sequences. In addition, the effect of spatiotemporal coherence interacted with the location where the images were photographed, $F(1,696) = 14.71$, $p < .001$, BF = 7.48. Decoding accuracy was significantly better in the coherent than the randomized sequences for both locations; however, the difference was larger in the off-campus, $\beta = 0.01$, $SE = 0.001$, $t = 9.55$, $p < .001$ than the on-campus sequences, $\beta = 0.007$, $SE = 0.001$, $t = 5.82$, $p < .001$. This is consistent with the general finding that top-down facilitation of perception tends to be greater when perception is more difficult (Gregory, 1990; Summerfield & De Lange, 2014). None of the remaining interactions in the analyses were statistically significant.

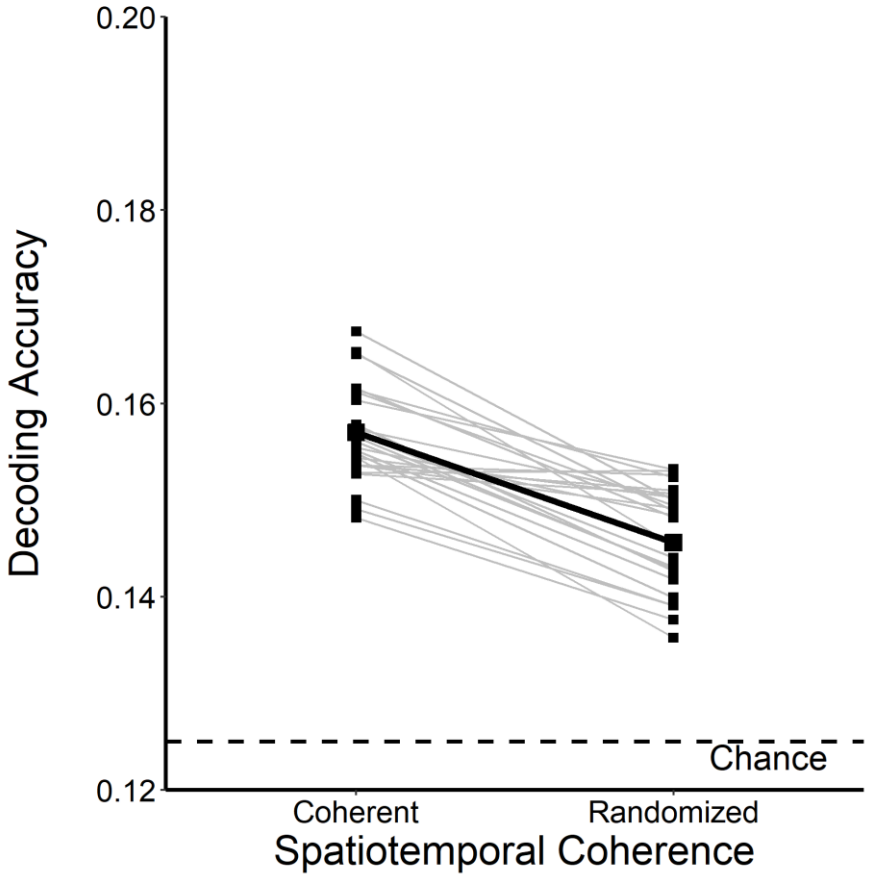


Figure 28. Exp 2: Decoding accuracy after the onset of the images as a function of the spatiotemporal coherence of the image sequences. Decoding accuracy for individual participants are represented by the lines. Least square means generated from the estimated regression

equation are represented by the thick black line and dots. The dashed line at .125 represents chance level performance.

Simultaneity of visual representation and behavioral categorization

We also examined when in the time course of scene processing decoding accuracy diverged between the coherent and randomized sequences. Early differences, before 150 milliseconds, in decoding accuracy between coherent and randomized sequences would support the hypothesis that facilitation arises before the visual system begins to activate higher-level representations, possibly during the construction of the structural description of the scene. Differences in decoding accuracy at or after 150 milliseconds would support the hypothesis that the benefit was due to facilitation occurring in the matching stage.

We averaged decoding accuracy at each time point (i.e., ms), across the image location, and hemisphere, within each region, prior to running the analyses, since the benefit for the coherent sequences was found in both locations and at each of the 8 regions. We then conducted a paired samples t-test using decoding accuracy as the dependent measure at each time point within the epoch. We also conducted the same simulation that we did when we evaluated when in the time course of scene processing vERPs diverged between coherent and randomized sequences. The simulation revealed that a run length of 10 or greater occurred in 5% of the simulations in the frontal sites, 12 or greater occurred in 5% of the simulations in the central sites, and 7 or greater occurred in 5% of the simulations in the parietal/occipital sites. Results are shown in Figure 27. We report Bayes factors for each statistical comparison since they are not influenced by the family wise error rate (Dienes, 2016). Decoding accuracy was significantly greater in the coherent sequences in the frontal regions starting from 199 milliseconds post scene onset, and the difference lasted until 417 milliseconds. The difference became statistically

significant again at 457 milliseconds and it remained significant until 476 milliseconds. These differences were supported by Bayes factors in favor of the alternative hypothesis greater than 3 at each time point where we found a statistically significant difference. We observed consistent results at the central sites though the number of consecutive statistically significant differences were much smaller in the central region. Decoding accuracy was significantly greater in the coherent than the randomized sequences from 207 to 269 milliseconds, and then again from 394 to 523 milliseconds. Bayes factors were greater than 3 between 207 and 269 milliseconds and from 394 to 417 milliseconds. Together, these results suggest that facilitation may arise in the matching and event model integration stages of scene processing.

Remarkably, differences in decoding accuracy between the coherent and randomized sequences in parietal/occipital regions peaked even earlier, consistent with early facilitation accounts (Aitken et al., 2020; Biederman et al., 1982; Edwards et al., 2017; Muckli et al., 2015; Palmer, 1975b). Decoding accuracy for scenes shown in coherent sequences first became significant at 62 milliseconds, but it only lasted until 89 milliseconds; however, Bayes factors in favor of the alternative hypothesis within this cluster of consecutive statistically significant differences peaked at 70 milliseconds with a Bayes factor of 2.66, suggesting weak evidence in favor of the alternative hypothesis. This was the first evidence we found to suggest that the event model influences scene gist perception very early in the time course of scene processing. Decoding accuracy was also significantly greater in coherent sequences between 175 and 218 milliseconds, and again from 378 to 589 milliseconds. Bayes factors in favor of the alternative were greater than 3 between 203 and 210 milliseconds, and again from 410 to 429, 476 to 484, and from 554 to 558 milliseconds. Together, these analyses suggest that the event model not only influences amplitudes in response to scenes presented in coherent sequences, but the event model

also facilitates the perception of scene category representations as they emerge over time, and it may do so as early as 70 milliseconds after the onset of a scene, consistent with early facilitation accounts, but inconsistent with pure feed-forward accounts.

We also measured the similarity of the underlying categorical representations by correlating decoding patterns from each time point in the epoch with the behavioral responses from the behavioral categorization task. We correlated the vector consisting of the 64 entries (8 categories X 8 responses) of the confusion matrix from the neural decoder with the vector of 64 entries of the confusion matrix for each participant at each time point (i.e., each ms) within the epoch for both on- and off-campus sequences separately.

Behavioral confusion matrices averaged across all of the participants for the coherent and randomized on- and off-campus image sequences are represented in Figure 29. As mentioned previously, participants were very accurate at categorizing the scenes, so they did not make very many errors. However, the errors they made were systematic as one would expect given prior research (Choo & Walther, 2016; Loschky et al., 2015). As evident in Figure 29, confusions were made among the indoor categories and among the outdoor categories (e.g., lawns were often misclassified as sidewalks or courtyards), but rarely did an indoor category get confused with an outdoor category or an outdoor category get confused with an indoor category. The notable exception is that some of the stairwells in the off-campus randomized condition (See Figure 29d)) were occasionally identified as a sidewalk.

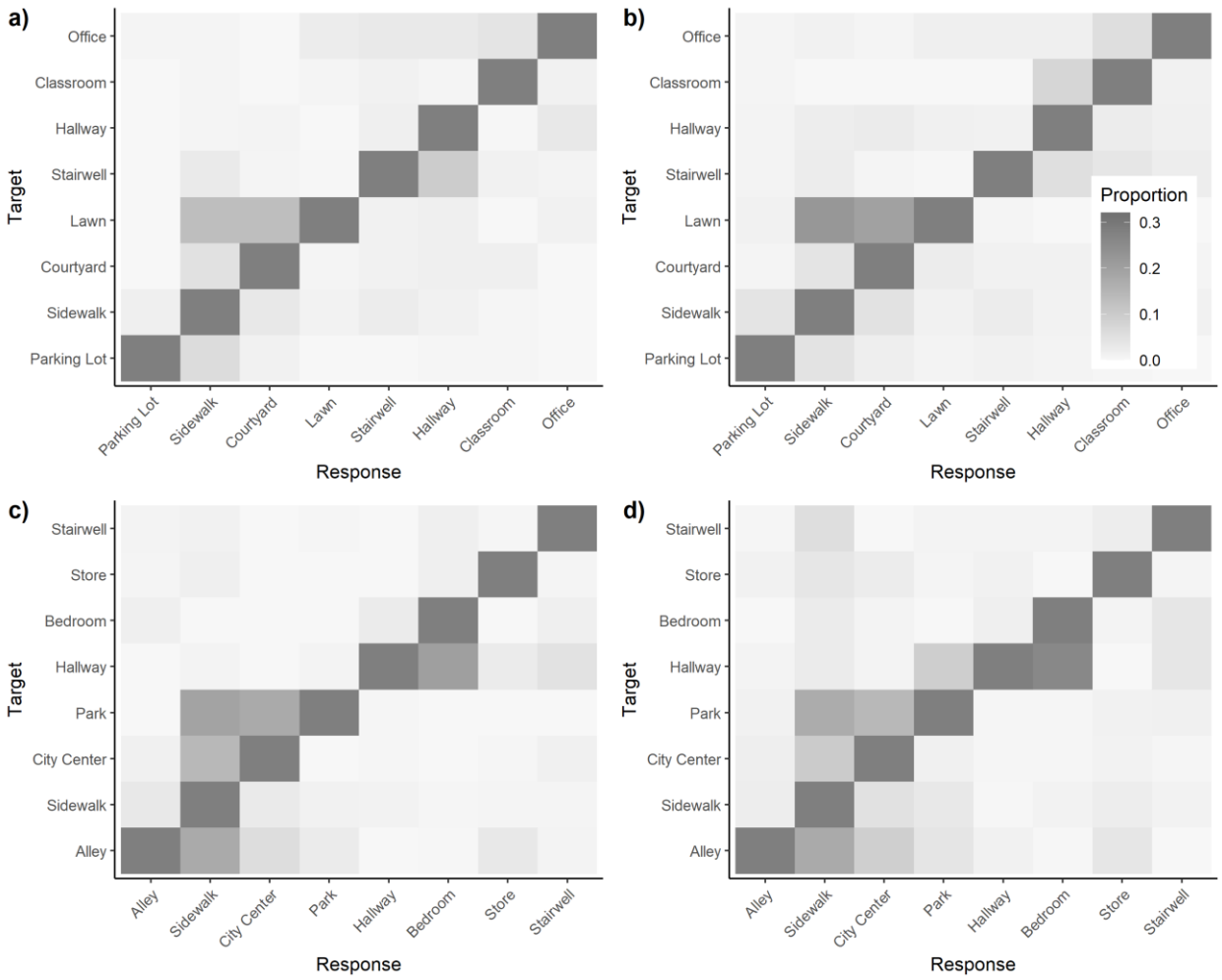


Figure 29. Exp 2: Confusion matrices for coherent and randomized image sequences for on- (top row) and off-campus (bottom row) images. Confusions in coherent sequences across participants are represented in a) and c). Confusions in randomized sequences across participants are represented in b) and d). Rows represent the target image category, and columns represent the average responses made for each response category. Thus, responses on the main diagonal are correct responses. Images belonging to indoor categories were often confused with other indoor categories, and images belonging to outdoor categories were often confused with other outdoor categories.

The unique variance in behavioral responses accounted for by the neural decoder responses at each time point in the epoch are shown in Figure 30. Correlations followed a similar pattern to what we observed for decoding accuracy. Correlations between the responses made by participants and the support vector machine increased monotonically for both the coherent and randomized sequences in all three regions around 50 milliseconds and peaked earlier for parietal/occipital regions (144 milliseconds) than central (191 milliseconds) and frontal (152 milliseconds) regions. These results were consistent with what we found for decoding accuracy. The notable exception is that decoding accuracy increased approximately 350 milliseconds after the onset of the scenes, but correlations between responses from participants and responses from the support vector machine decreased. The reason for this discrepancy between decoding accuracy and the correlations is not clear. Results suggest the model became more accurate as processing time increased (i.e., the values on the main diagonal of the confusion matrices increased), but the errors that were made disagreed with confusions made by participants.

We reran the analysis comparing changes in decoding accuracy at each time point in the epoch between the coherent and randomized sequences with the correlations between behavioral responses and responses made by the SVM as the dependent variable. In frontal regions, correlations were significantly larger in the coherent condition from 242 to 417 milliseconds. Bayes factors were greater than three between 265 to 300 milliseconds and again from 355 to 417 milliseconds. Correlations in central regions were significantly greater in the coherent sequences between 250 to 328 milliseconds and again from 488 to 539 milliseconds. These significant differences were supported by Bayes factors greater than 3 from 312 to 316 milliseconds and again from 500 to 527 milliseconds. Together, these results suggest that the event model facilitated matching processes.

Consistent with decoding accuracy, correlations were significantly greater in the coherent condition at parietal/occipital regions from 70 to 93 milliseconds, from 183 to 226 milliseconds and from 441 to 589 milliseconds. Unlike decoding accuracy, Bayes factors were greater than 3 in parietal/occipital regions from 74 to 82 milliseconds, from 207 to 214, from 492 to 496, and from 562 to 585 milliseconds. Thus, this is clear evidence that the event model facilitates rapid scene categorization very early in perceptual processing as well as during the matching stage. This is consistent with predictions of SPECT that the event model generates predictions before the next image is presented, and that this facilitates perception from low to high levels. It is inconsistent with pure feed-forward accounts of rapid scene categorization.

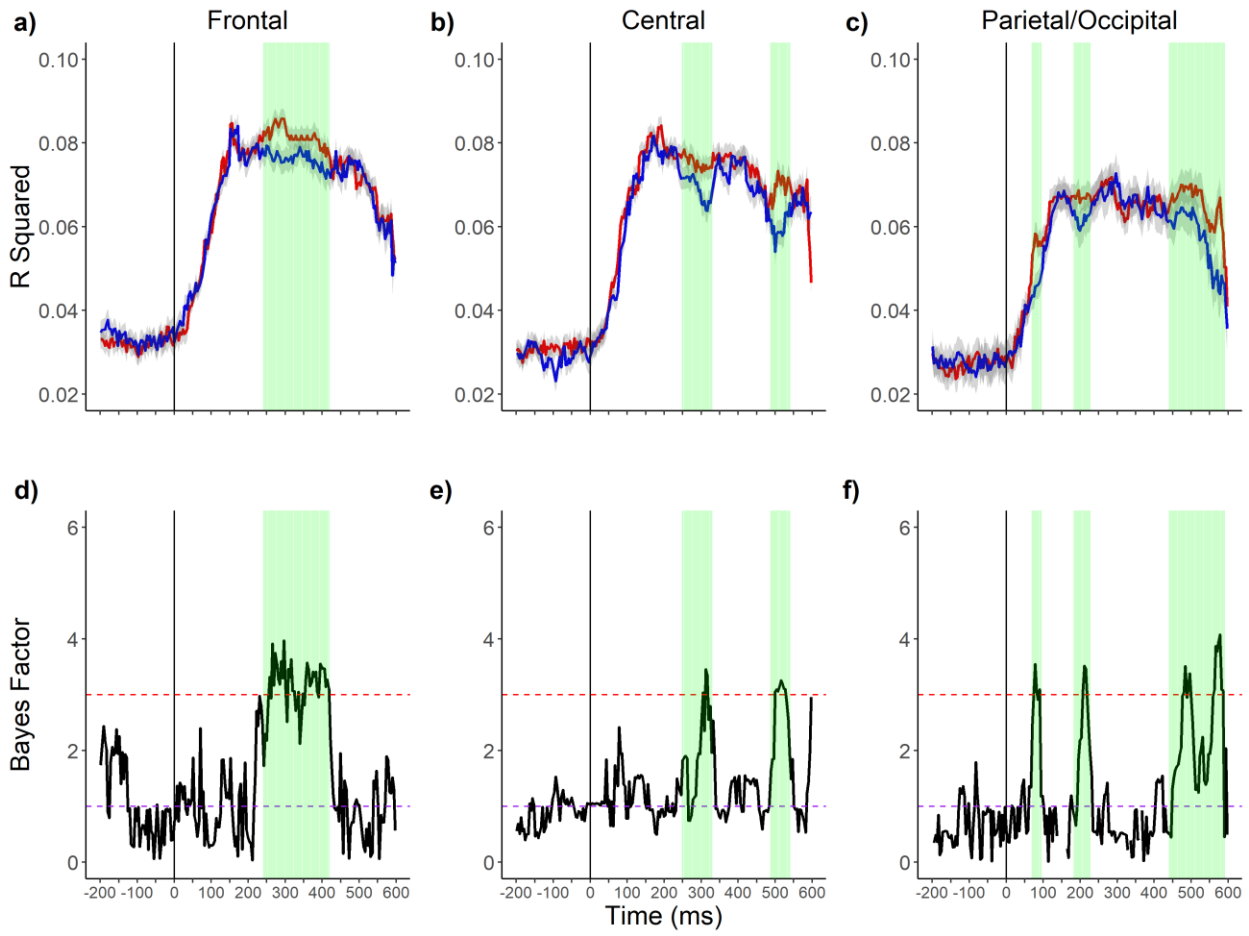


Figure 30. Exp 2: ***Unique variance*** in the behavioral confusion matrices explained by confusions made by the neural decoders over time. Error bars represent between subject 95% confidence intervals at each time point.

Discussion

According to data-driven feed-forward processing accounts of rapid scene categorization, expectations for upcoming scenes should not influence scene recognition (Fabre-Thorpe et al., 2001; Riesenhuber & Poggio, 1999; Serre et al., 2007; VanRullen, 2007; VanRullen & Thorpe, 2002). Nevertheless, the current experiment demonstrated that scene categorization both in terms of rapid scene categorization ability and in the ability of machine learning techniques to decode

scene-category information is better when scenes are shown in more ecologically valid sequences than randomized sequences.

Participants saw scenes in spatiotemporally coherent sequences, organized to represent an approach to a destination (e.g., walking from an office to a parking lot), or their randomized versions, while we recorded their event-related potentials. Participants categorized one of the scenes on each trial. Importantly, participants saw the same images in the coherent and randomized versions; therefore, participants processed the same feed-forward information in both conditions. In addition, we also ensured that the target was the same scene in the same temporal location in both conditions. For example, if the target was the 4th image in an office to a parking lot sequence in the coherent condition—say, a specific hallway image--then it was the same hallway image as the 4th image in that sequence's randomized version. Therefore, we controlled all aspects of the trials, including the target image itself, its ordinal position, the other images in the sequence, and the correct behavioral response in both coherence conditions. Sequences only differed in their spatiotemporal coherence. Importantly, we found that rapid scene categorization performance was greater when the sequence was coherent than when it was randomized, replicating prior work (McLean et al., 2021; Smith & Loschky, 2019).

Facilitation of the P200

We also extended previous work by examining the time course of sequential expectations on scene perception using EEG. We were interested in two different neural components: the P200 and N400. The P200 is a scene-selective component observed over parietal/occipital regions (Harel et al., 2016). Some work has demonstrated that unexpected scenes elicit a larger

P200 than expected scenes (McLean et al., 2021)[14]. In contrast, other work has shown that it cannot be influenced by top-down factors since changing a viewer's goals while viewing a scene does not modulate its sensitivity to global scene properties (e.g., a scene's openness, naturalness, etc.) (Hansen et al., 2018). Consistent with this work, we found little evidence to suggest that predictions made prior to viewing a scene could influence amplitudes of the P200 in parietal/occipital region. As shown in Figure 13, in the parietal/occipital region, average amplitude between 150-249 milliseconds did not significantly differ between coherent and randomized sequences in response to the target scene. Similarly, Figure 16 shows that in the parietal/occipital region, averaged amplitudes across all of the scenes did not differ between coherent and randomized sequences, regardless of their target status. Likewise, as shown in Figure 17, we found no evidence in the parietal/occipital region of facilitation of the P200 in our time point-by-point analysis. Figure 22 shows that it was only when we looked at changes in amplitude as a function of the ordinal position of the scenes on each trial, that we found small differences between the coherent and randomized sequences in the P200 in the parietal/occipital regions. Interestingly however, as shown in Figures 24 and 26, image similarity between successive images modulated the P200 in the middle parietal/occipital region (Figure 24), but image predictability did not (Figure 26). These last two results suggest that our results are more consistent with work demonstrating that the P200 recorded at parietal/occipital regions cannot be influenced by top-down observer-based factors (Hansen et al., 2018) than with work showing that unexpected scene categories elicit a larger P200 because they are unexpected in

[14] We replicated some, but not all of the effects reported by McLean et al. (2020). We did not support the hypothesis that the P200 was smaller for scenes shown in coherent sequences. Consistent with their results, we found that the N400 was more negative in response to scenes shown in randomized sequences. McLean et al. (2020) did not use a multivariate pattern classifier to decode scene categories.

parietal/occipital electrodes (McLean et al., 2021). However, the P200 can be modulated by repeating visual features between successive stimuli.

The reason for the discrepancy between our results and the results of McLean et al. (2021) are unknown, but they may be due to differences in the experimental designs between our and their studies. McLean et al. (2021) showed participants a series of 5 scenes followed by a target. Importantly, scenes were organized to represent an approach to a destination (e.g., the first-person view of someone walking toward the door of a house from outside the house) as we did here. Participants were always asked to identify the 6^{th} scene on each trial. Critically, the target on each trial was either congruent with one's predictions (e.g., after multiple views from outside a house leading up to the door, a living room) or incongruent (e.g., the inside of a parking garage). Thus, in the two predictability conditions, the priming scenes were the same, but the targets differed, though they did ensure that the expected and unexpected scenes belonged to the same superordinate level (indoor vs. outdoor) category. In contrast, we showed participants exactly the same scenes, and only manipulated their order so that all scenes in the sequence were either predictable or not (except the first scene, which was never predictable). Thus, it is possible that McLean et al. (2021) may have observed increased amplitudes in the P200 for unexpected scene images because they used different scenes in the two conditions. This is plausible given that the P200 has been shown to be sensitive to global scene properties both at the image and categorical levels (Harel et al., 2016; Harel et al., 2020; Lowe et al., 2018) regardless of whether participants are asked to attend to the scene category (Hansen et al., 2018). Any systematic difference in the global scene properties between the expected and unexpected scenes used in McLean et al. (2021) may have resulted in differences in the P200.

The effect of image similarity and predictability on scene processing

It is also possible that McLean et al. (2021) observed differences in the P200 because expected scenes shared more visual features with their primes than the unexpected scenes they used. We found that image similarity, as quantified from the shared spectral information between sequential prime and target pairs, but not image predictability from an independent sample of participants, correlated uniquely with amplitudes in the P200 primarily in the middle parietal/occipital region. Therefore, the facilitation effects on the P200 in parietal/occipital regions observed by McLean et al. (2021) may have been the result of processing more versus less similar features between prime and target in the feed-forward sweep, rather than top-down predictions that feed-back and influence scene processing. Our results are correlational, so future research could manipulate the predictability of an image while holding the visual similarity between prime and target constant to examine the unique contribution of each on the P200.

Future research could also evaluate how the shared visual feature overlap between scenes can facilitate scene categorization using a different method for quantifying image similarity. We used the Spatial Envelope Model because it is diagnostic of a scene's global properties (Greene & Oliva, 2009b; Oliva & Torralba, 2001, 2006), and its features are related to multi-scale gabor pyramids that resemble receptive fields in V1 (Ringach, 2002). Features from the spatial envelope have also been found to explain a significant amount of variance in MEG (Harel et al., 2016; Ramkumar et al., 2016) as well as fMRI responses (Park et al., 2011; Watson et al., 2014). However, other scene visual similarity metrics may account for unexplained variance in voltage that shared localized spectral amplitude information does not.

Nevertheless, although we did not observe differences in the amplitudes of the P200 at parietal/occipital electrodes, we did find evidence of facilitation in the same window (150-249

ms) in the frontal and central regions. This effect was surprising because we did not expect differences in frontal and central regions until, perhaps the later N400 component. However, our results are consistent with the results of McLean et al. (2021), who also found such differences in both frontal and central regions. Given that neural activity after 150 milliseconds is usually considered to reflect the formation of perceptual judgements (Johnson & Olshausen, 2002; Johnson & Olshausen, 2005; Schendan, 2019; VanRullen & Thorpe, 2001c), our results suggest that predictions may constrain the scene categories that are considered for matching to the visual stimulus (Bar, 2004; Bar & Ullman, 1996; Friedman, 1979; Palmer, 1975a; Trapp & Bar, 2015).

Neural mechanisms underlying the effect of predictions on scene processing

Source localization of the vERP differences between the spatiotemporally coherent versus randomized sequences in the 150-249 milliseconds window, at frontal and central electrodes, revealed possibly differential processing in the precuneus and poster cingulate cortex. These two brain regions are part of the posterior medial and default mode networks (Buckner et al., 2008; Inhoff & Ranganath, 2017; Leech & Sharp, 2014; Ranganath & Ritchey, 2012). The posterior medial network has a central role in the construction of the event model (Ranganath & Ritchey, 2012) and both have projections to the parahippocamal cortex and the retrosplenial cortex (Libby et al., 2012; Ritchey et al., 2014), which are involved in scene perception (Epstein, 2005; Kahn et al., 2008). Specifically, the parahippocampal cortex, anterior lingual, and medial fusiform gyri make up a region collectively known as the parahippocampal place area (Epstein & Kanwisher, 1998). Functional differences between the parahippocampal place area and the retrosplenial cortex are currently under investigation, but the general consensus is that the parahippocampal place area is involved in perceptual processing of a scene's spatial layout and category (Bilalić et al., 2019; Epstein & Kanwisher, 1998; Epstein, 2005; Harel et al., 2013;

Walther et al., 2009); whereas the retrosplenial cortex is more involved in integrating local scene information into a broader representation and likely represents navigational information (Aminoff et al., 2013; Cho & Sharp, 2001; Dilks et al., 2013; Park et al., 2007; Persichetti et al., 2016). Unlike the parahippocampal place area, the retrosplenial cortex contains head direction cells instead of place cells (Cho & Sharp, 2001; Vann et al., 2009). These cells respond whenever the head is pointing in a particular direction, providing information about the orientation of the individual in space and self-motion. Lesions of the retrosplenial cortex produce topographical disorientation, a disorder that leaves patients with the inability to update their egocentric heading and find one's way in the environment (Aguirre & D'Esposito, 1999; Hashimoto et al., 2010).

When viewing a coherent sequence, the parahippocampal place area may represent the spatial layout and category of a particular scene. The retrosplenial cortex may integrate information across successive scenes, and update representations of one's current position (Cooper & Mizumori, 1999). Integrating inputs from the parahippocampal place area and the retrosplenial cortex, the posterior medial network, which consists of the posterior cingulate and precuneus, may construct an event model of the sequence. The event model may then be used to orient the individual and generate predictions about the visual input that would be expected from apparent motion through the environment. The posterior medial network may then send feed-back signals to the parahippocampal place area and the retrosplenial cortex (Kahn et al., 2008; Libby et al., 2012) , facilitating matching processes. Our vERP results suggest that feedforward processing may underlie information extraction until facilitation in the matching stage. Such claims are also consistent with the Posterior Medial- Anterior Temporal (PM-AT) framework (Inhoff & Ranganath, 2017; Ranganath & Ritchey, 2012).

We note that the method we used to perform the source localization of the difference in vERPs did not find evidence that the differences were localized in the parahippocampal place area or the retrosplenial cortex. However, that may be due to the fact that these two brain regions are deep within the brain in the interior junction of the temporal and occipital lobes. The technique we used to localize the differences in the vERPs is known to be less able to localize dipoles within deeper brain structures further from the skull (Acar & Makeig, 2013). Future research could examine the functional connections between the posterior medial network and these scene selective brain regions with fMRI or MEG when participants view scenes in coherent versus randomized sequences. Future work could also investigate if preparatory activity in the parahippocampal place area, informed by signals sent from the posterior medial network, facilitates scene processing.

Facilitation of the N400

As hypothesized, we also found differences in a later vERP component, the N400. The N400 is a negative going waveform localized over central and frontal electrodes, though it has a more frontal topography, and has a longer duration (lower frequency) in response to visual images than auditory stimuli or text (Ganis et al., 1996; Holcomb & Mcpherson, 1994). The N400 is thought to be a domain-general index of semantic processing, possibly reflecting semantic access (see Kutas & Federmeier, 2011 for a review) or processes involved in integrating the meaning of an item into the event model (Cohn & Kutas, 2015; Hagoort et al., 2009; Kutas & Federmeier, 2000). Though it is modulated by the degree to which a stimulus is predictable and consistent with preceding information (Coderre et al., 2020; DeLong et al., 2005; Van Petten & Luka, 2012), it is not simply the brain's response to anomalies. Instead, it reflects the ease of integrating information into working memory, whereby its amplitude is more positive

when it is easier to integrate information into working memory (Kutas & Federmeier, 2000). According to SPECT, the first image within a sequence should lay the foundation of the event model. Subsequent scenes map onto the event model to the extent that they cohere with previous scenes. The N400 may serve as an index of this mapping process (Cohn & Kutas, 2015; Gernsbacher, 1990; Loschky et al., 2020). Specifically, it is reduced when scenes cohere.

In addition, we also found that the amplitudes of the N400 changed as a function of the amount of visual similarity between the prime and target scene in terms of their shared spectral information, and it did so in both the coherent and randomized sequences. This could be expected given current theories of the N400. Two visually similar scenes, all else being equal, will likely be semantically more coherent than two scenes that are less visually similar. This was shown in our finding that visual similarity was significantly greater in the coherent sequences than the randomized sequences, even though all such pairs of coherent versus randomized sequences contained the same sets of images. Thus, integrating visually similar images into the event model should be easier when the sequences are coherent.

Importantly, we also found that image predictability, as determined from participants in Experiment 1, independently correlated with amplitudes of the N400 after controlling for the influence of image similarity, and this effect was stronger in the coherent than in the randomized sequences. The mechanism underlying this effect is unknown. A simple explanation is in terms of the greater range in predictability values in the coherent condition, given that the predictability in the randomized condition was very nearly at chance. An alternative explanation is that predictability may only influence the N400 when scenes are both locally and globally coherent. Namely, randomized sequences inevitably have sequential pairs of images that are locally coherent, but they are never globally coherent. Previous researchers have found evidence

consistent with this idea (Coderre et al., 2020; DeLong et al., 2005; Van Petten & Luka, 2012). Anomalous words that are unexpected in sentences tend to elicit a larger N400 than predictable words (e.g., the cat picked up a *hammer vs.* the cat picked up a *toy*). However, the larger N400 effect that is usually associated with processing such anomalies can be eliminated if sentences are embedded within a story that is familiar to the reader and provides context for the critical word (e.g., in the context of a Tom and Jerry cartoon) (Filik & Leuthold, 2008). That is, the N400 can be reduced if a sentence that is typically locally incoherent becomes globally coherent. In addition, words that are locally coherent within sentences that would typically fit within a sentence can elicit a larger (i.e., more negative) N400 if prior context enables one to make a different prediction based on the sentence's global coherence (Van Berkum et al., 1999). For instance, reading the word *slow* in the following sentence may elicit a larger N400 than the word *quick* if prior information in a story established that the character was *quick*: "Jane told her brother that he was exceptionally *slow/quick* at reading". Thus, scenes that were highly predictable in the randomized sequences due to their local coherence may have elicited similar N400s to scenes that were not predictable because they lacked global coherence. As such, their amplitudes remained negative.

On the other hand, scenes that were highly predictable in the coherent sequences should have been both locally and globally coherent; therefore, amplitudes in response to those predictable scenes were reduced (i.e., more positive). This finding is consistent with hypotheses generated from SPECT that the extent to which facilitation will be found depends upon the degree of coherence between the incoming scenes and the event model (Loschky et al., 2020).

Predictable scenes elicit clearer neural signals

We also observed a benefit in the ability to decode the image sequences from the neural activity over time. To do this, we assessed decoding accuracy between scenes shown in coherent and randomized sequences using a multivariate pattern classifier, trained to discriminate among different scene categories from participants' neural activity. Models were separately trained on scenes shown in coherent and randomized sequences. We found that decoding accuracy for scenes shown in coherent sequences was better than decoding accuracy for scenes shown in randomized sequences (See Figures 27 and 28). Importantly, decoding accuracy between coherent and randomized sequences began to diverge around 62 milliseconds post-stimulus in parietal/occipital electrodes. If this divergence can tell us the time course for when the event model begins to facilitate rapid scene categorization, then this is the first evidence we have to suggest that top-down predictions from the event model may facilitate early perceptual analysis of scene categories (Aitken et al., 2020; Biederman et al., 1982; Muckli et al., 2015; Palmer, 1975b) in parietal/occipital regions, though this early difference was associated with a weak Bayes factor in favor of the alternative hypothesis. Finding support for early facilitation in the neural decoding analysis, but not in the analysis of vERPs could be expected given that vERPs reflect changes in activation between the coherent and randomized condition over time; whereas, accuracy of the neural decoders reflect changes in the pattern of activation between categories over time (Hebart & Baker, 2018). A remaining question is whether this early decoding advantage for scenes in coherent sequences is better predicted by prime-to-target visual similarity, or predictability of the target. Given our earlier results showing that P200 amplitudes in the parietal/occipital region were better predicted by visual similarity than predictability, we

might expect to find the same to be true for this early coherence facilitation of decoding accuracy in the parietal/occipital region.

We also correlated the distribution of responses of the neural decoder with the distribution of behavioral responses of participants, and we found higher correlations for scenes presented in coherent sequences (see Figure 30). Together, these results provide evidence that vERPs not only contain information relevant for scene categorization, but activity patterns mirror behavioral categorization performance (e.g., a better ability to categorize scenes shown in coherent sequences).

Prior work has found that decoding scene representations using fMRI, MEG, and EEG mirror different behavioral measures. Specifically, decoding in brain regions that are associated with scene processing correlates with the distribution of behavioral scene categorization errors made by subjects (Ramkumar et al., 2016; Torralbo et al., 2013; Walther et al., 2009). In addition, decoding accuracy decreases when participants view inverted scenes or poor exemplars of scene categories in comparison to upright scenes or representative exemplars of scene categories (Torralbo et al., 2013; Walther et al., 2009). Further, decoding errors as a function of processing time correlates with behavioral errors made by subjects performing scene categorization tasks similar to what we used (Ramkumar et al., 2016).

We extended prior work by demonstrating that the relationship between decoding errors and behavioral errors is strengthened when scenes are shown in coherent sequences, and they begin to become stronger approximately 70 milliseconds after scene onset in parietal/occipital electrodes and approximately 200 milliseconds after scene onset in all of the regions. The early difference could suggest that predictions facilitate early perceptual processing (Aitken et al., 2020; Biederman et al., 1982; Muckli et al., 2015; Palmer, 1975b), and differences at the later

time point (~150 milliseconds) may reflect facilitation in matching the structural description to a representation stored in semantic memory (Bar, 2004; Bar & Ullman, 1996; Friedman, 1979; Leroy et al., 2020; Mudrik et al., 2010; Palmer, 1975a; Schendan, 2019; Trapp & Bar, 2015). Future research could evaluate if correlations between responses made by neural decoders and humans are better in both scene-selective brain regions and event model-selective regions using MEG or fMRI when scenes are shown in a coherent sequence. This could help identify which brain regions are involved in producing the stronger correlations we observed.

Chapter 4 - Experiment 3

It is worth noting that decoding accuracy in Experiment 2 was very poor, though it was significantly greater than chance, which is really all that is needed to get information from the decoder (Hebart & Baker, 2018). Furthermore, correlations between responses made by neural decoders and human observers were small, though poor performance of the neural decoders could be expected given prior work (Ramkumar et al., 2016; Greene & Hansen, 2020). In fact, Ramkumar et al. (2016) found that decoding accuracy for different scene categories capped at 20% (chance was 16.67%), and R^2 values ranged from .001 to .04. On the other hand, the above-chance performance of the neural decoders can be considered to be quite remarkable given the great amount of variability between exemplars from each category. For instance, exemplars that belonged to a given category (e.g., office) can have different colors, textures, spatial layouts, and individual object components. Thus, decoders must extract very abstract information that distinguishes an office from a classroom in order to correctly categorize a given scene from a messy signal.

The small correlations between responses provided by neural decoders and human behavior could have also been expected given that categorization performance was very high (Raw $M_{\text{Accuracy}} = 83.35\%$, $SE_{\text{Accuracy}} = 0.004$) in Experiment 2. As shown in Figure 29, participants made very few errors, though the errors they made were very systematic. If all entries in the confusion matrices of participants were along the main diagonal, then we would get no information from them. In Experiment 3, we sought to reduce rapid scene categorization performance by decreasing the duration for which images were shown, and by immediately following the target scene on each trial with a noise mask (Enns & Di Lollo, 2000; Loftus & Mclean, 1999; Massaro & Loftus, 1996). The harder the scene is to perceive, the more errors

both the humans and the neural decoder should make, resulting in more decoding mistakes. Thus, decoding accuracy should be lower in Experiment 3 than in Experiment 2. Subsequently, this should also increase the off-diagonal entries in both the decoder and the human confusion matrices. If we assume that errors are systematically related to the categorization process (Choo & Walther, 2016; Loschky et al., 2015; Walther & Shen, 2014), then correlations between responses made by neural decoders and humans should be greater in Experiment 3 than in Experiment 2 since there should be greater variability to capture in the human responses. This should then lead to greater insights into the timing of the facilitation of scene categories by the event model. Thus, Experiment 3 was carried out to both replicate Experiment 2, and to gain greater insight from the neural decoding.

Method

Participants

Twenty-four (N = 12 females, N = 12 males) participants participated in Experiment 3. None of them participated in Experiments 1 or 2, and 22 of 24 were right-handed. Sixteen participants were compensated with $15 an hour for their participation, and the remaining participants were compensated with course credit. Age of the participants ranged from 18 to 45 (M age = 23.08). We removed EEG data from 3 participants for excessive movement artifacts. Their behavioral data were retained in the final dataset. Participants began the experiment by signing an electronic informed consent form, which was authorized from Kansas State University's Institutional Review Board. Participants had normal or corrected to normal visual acuity (< 20/30 Snellen acuity) as measured using the Freiburg Visual Acuity and Contrast Test (Bach, 2006). None of the participants were aware of the purpose of the experiment and were

asked not to discuss the purpose of the experiment with others. The experiment lasted approximately 2.5 hours. We encouraged participants to take a break after the first block of trials.

EEG Data Acquisition and Preprocessing

We recorded EEG signals using the same system as in Experiment 2. Data were preprocessed and cleaned in the same way. EEG data for 9% of the images was removed from the total dataset due to artifacts (Range = 0% - 15%). The remaining data were submitted to ICA. Components identified to be related to eye movements, EKG, channel noise, and muscle movements were removed from the dataset following the same protocol as in Experiment 2. We removed an average of 15.36, *SD* = 5.12, components across participants. (Range = 7 - 30).

All of the continuous EEG data were then epoched for 800 ms, from 200 ms prior to the onset of the images to 600 ms after onset of the images as before. Again, we used the mean voltage in the 200 milliseconds prior to stimulus onset to do the baseline correction on the voltage in the entire epoch.

Procedure

The procedure was the same as Experiment 2, except images were flashed for half as long, at 48 milliseconds each, and target images were masked. Masking the target could be a potential problem with the neural decoding analysis since neural decoders were trained on scenes without a mask, but tested on scenes that were masked. Each image, except for the target, was interleaved by a 752-millisecond neutral gray screen (800 millisecond SOA). Unlike the other images, the target on each trial was immediately followed by a colored random 1/f noise mask for 96 milliseconds. This would make the target, but not the primes, more difficult to categorize. Specifically, in Experiment 2, which did not include a visual backward mask, once the target image was removed from the screen after 96 ms, we can assume that the visual system continued

to extract information for at least 200 ms longer (Loschky et al., 2010; Loschky, Sethi, et al., 2007). However, in Experiment 3, presenting the visual noise mask immediately after the target would curtail further accumulation of information from the target image, and these effects would be shown throughout the feed-forward sweep of neural activity (Kovacs et al., 1995; Loftus & Mclean, 1999; Rieger et al., 2005; Rolls et al., 1999).

We asked participants the same 4 questions from Experiment 2 after Experiment 3. As in Experiment 2, two raters independently judged from the participants' responses whether they reported anything about the coherence of the image sequences. Raters produced strong reliability (Cohen's $k = 0.86$). We resolved discrepancies between raters through thoughtful discussion to produce the final coding of the responses. Few participants reported that they noticed the manipulation on the second question (10%). Many more reported that they noticed the coherence of some of the image sequences on the third question (43%), and almost all of the participants reported that they noticed the coherence by the final question (81%). We did not evaluate whether noticing the coherence contributed to the benefit we observed in rapid scene categorization performance or in decoding accuracy for the coherent sequences since 4 participants failed to notice the manipulation.

Results

For completeness, we conducted all of the same analyses reported in Experiment 2 in an attempt to replicate the new effects. As in Experiment 2, we will begin by first describing the behavioral results. These will be followed by the vERP results time locked to the target on each trial and then time locked to all of the images within the trial, regardless of whether the image was a target or not. We will then report results investigating how each of components of interest changed across time within a trial. Finally, we will present the neural decoding analyses, which

were the reason for running Experiment 3 with reduced image durations and masking of target images. By doing so, we hoped to increase variance in the behavioral confusion matrices used in our decoding analyses.

Behavioral Results

We started by using a logistic mixed effects model to predict the probability of correctly categorizing the target scene from the fixed effects of spatiotemporal coherence (coherent vs. randomized), the location the images were photographed (on-campus vs. off-campus), and their interaction. As in Experiment 1, the participant intercept was allowed to vary as a random effect, and the main effects of spatiotemporal coherence, location, and their interaction were allowed to vary as random effects (by-participant intercept and by-participant slope random effects). Spatiotemporal coherence (Coherent = 0, Randomized = 1) and image location (off-campus = 0, on-campus = 1) were both dummy coded prior to entry into the model.

As shown in Figure 31, masking the scenes decreased overall accuracy as we hypothesized, though not by very much. Accuracy in the coherent (M = 0.86, SE = 0.02) sequences was similar to what it was in Experiment 2, but accuracy in the randomized sequences decreased more than in Experiment 2 (M = 0.81, SE = 0.02). Importantly, consistent with Experiment 2 and the results of Smith and Loschky (2019), we found that the difference in categorization accuracy between the coherent and randomized conditions was significant, β = -0.38, SE = 0.11, z = -3.46, p < .001, BF = 12.41; and this difference was notably larger in Experiment 3 than in Experiment 2 (β = -0.23, SE = 0.11, z = -2.09, p = .03, BF = 3.97). Further, the evidence in favor of the alternative hypothesis was larger in Experiment 3. This is consistent with the general principle that predictions or priors play a greater role in perception when the sensory signal is weaker, ambiguous, or imprecise (Gregory, 1990; Summerfield & De Lange,

2014). In addition, it is clear from Figure 31 that most of the participants showed the advantage in categorization performance. This result is consistent with the hypothesis that back-end processes involved in construction of the event model feeds back to influence front-end processes involved in information extraction. The logistic model was able to successfully discriminate correct from incorrect trials, AUC = 0.81.

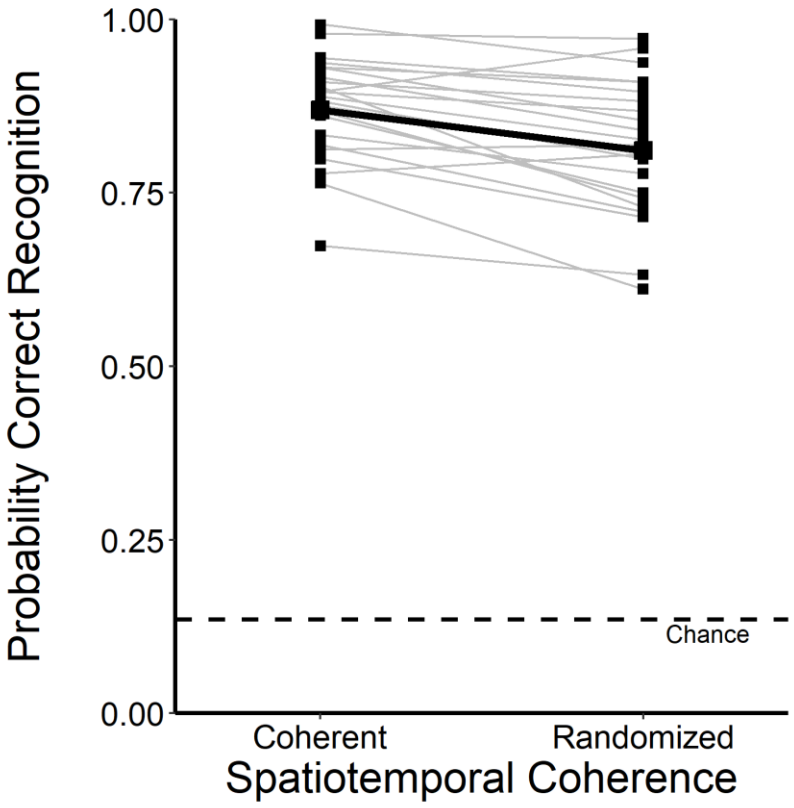


Figure 31. Exp 3: Rapid scene gist categorization performance as a function of the spatiotemporal coherence of the image sequences. The proportion of times each participant accurately categorized the target scenes is represented by the lines. Least square means generated from the estimated regression equation are represented by the thick black line and dots.

As in Experiments 1 and 2, we also examined how scene categorization changed as a function of the ordinal position of the target on each trial. Due to issues with model convergence with the maximal random effect structure (Barr et al., 2013), we specified the probit, rather than the canonical logit function, as the link for the generalized linear mixed effects model (i.e., we ran a probit mixed effects model as opposed to a logistic mixed effects model) using binary accuracy (correct = 1, incorrect = 0) as the dependent variable. Results are shown in Figure 32. As in Experiment 2, we observed a significant effect for spatiotemporal coherence, β = -0.20, *SE* = 0.06, z = -3.31, $p < .001$, BF = 12.41. Importantly, we also observed a significant effect for the ordinal position of the target on the trials, β = 0.49, *SE* = 0.02, z = 3.15, $p < .001$, BF = 69.95, consistent with hypotheses generated from SPECT that the extent of facilitation depends upon the degree of spatiotemporal coherence between the event model and the incoming scene information (Loschky et al., 2020). The slope of this effect was also steeper than what it was in Experiment 2 (β = 0.03, *SE* = 0.005, z = 6.69, $p < .001$, BF = 3.45). Unlike Experiment 2, the facilitation of rapid scene categorization performance accuracy increased at the same rate for coherent and randomized sequences as evident from a non-significant interaction between spatiotemporal coherence and the ordinal position of the scene on a trial, β = 0.01, *SE* = 0.02, z = 0.31, p = 0.75, BF = 1.62. The reason for the difference between the experiments is unclear, though the statistically significant interaction between spatiotemporal coherence and the ordinal position of the scene in Experiment 2 was associated with a small Bayes factor in favor of the alternative hypothesis (Experiment 2: BF = 3.45, Experiment 3: BF = 1.62). None of the remaining interactions were significant.

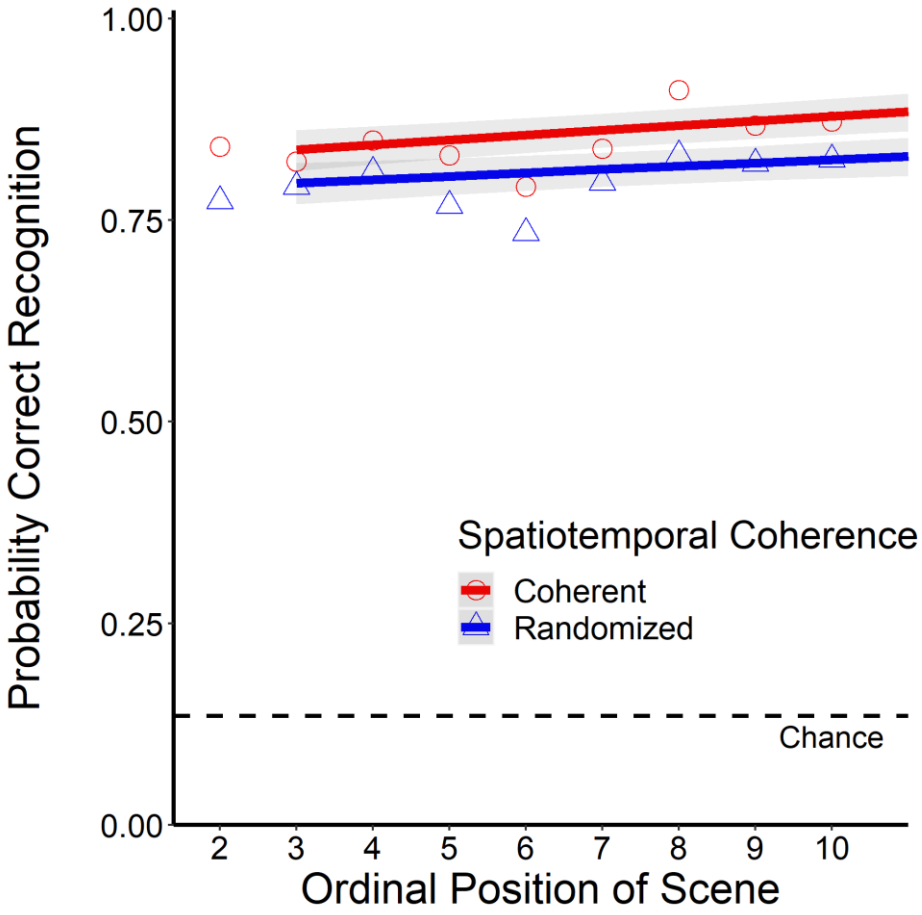


Figure 32. Exp 3: Rapid scene gist categorization accuracy as a function of the ordinal position (2-10) of the target scene on each trial, the spatiotemporal coherence of the image sequences, and the location the images were photographed. The proportion of instances when the target image was correctly categorized is represented by dots in the figure. The lines reflect the least square means calculated from the estimated regression equation.

vERPs to the Target Image

As in Experiment 2, we next ran a series of analyses to examine differences in vERPs to target scenes presented in coherent and randomized sequences. Comparing vERPs time locked to the onset of the target scene is important because the target was the same image, at the same

ordinal position (2-10) in both the coherent and randomized conditions. Participants extract the same information in the feed-forward sweep in both versions of the sequence; however, targets were predictable when the sequence was coherent and not when it was randomized.

In Experiment 2, we found that amplitudes between coherent and randomized sequences differed in the 150-249, and N400 windows, but not in the early component (50-149 millisecond window). To foreshadow the results, we replicated the primary effects we observed in Experiment 2.

Grand averages are reported in Figures 33 and 34 time locked to the onset of the target for the frontal and central as well as the parietal/occipital regions, respectively. The waveforms shown in Figures 33 and 34 are very different from the waveforms reported in Experiment 2 (See Figures 10 and 11). Amplitudes in the frontal and central sites are much more negative than they were in Experiment 2, and amplitudes in the parietal/occipital regions are reduced. Differences between experiments could be due to either the decreased duration of the target or the onset of the perceptual mask that immediately followed it, though, variations in target duration are generally less important to perception and related EEG than variations in the stimulus onset asynchrony of a backward mask (Loschky, Sethi, et al., 2007; Robinson et al., 2019). Further, these waveforms contain both components in response to the target image, but also components in response to the onset of the perceptual mask. Differences in the waveforms may have also been due to differences in the cortical folds between the group of participants. Nonetheless, overall differences as a result of the spatiotemporal coherence manipulation are similar to what we found in Experiment 2. Amplitudes were more positive in the coherent sequences than in the randomized sequences in the frontal electrodes at each time point after approximately 150-249 milliseconds, consistent with matching accounts of facilitation, and waveforms in response to on-

campus coherent sequences were more positive than both off-campus coherent and the randomized sequences. In addition, the difference in the waveforms between conditions time locked to the target scene in the parietal/occipital electrodes do not appear to differ significantly.

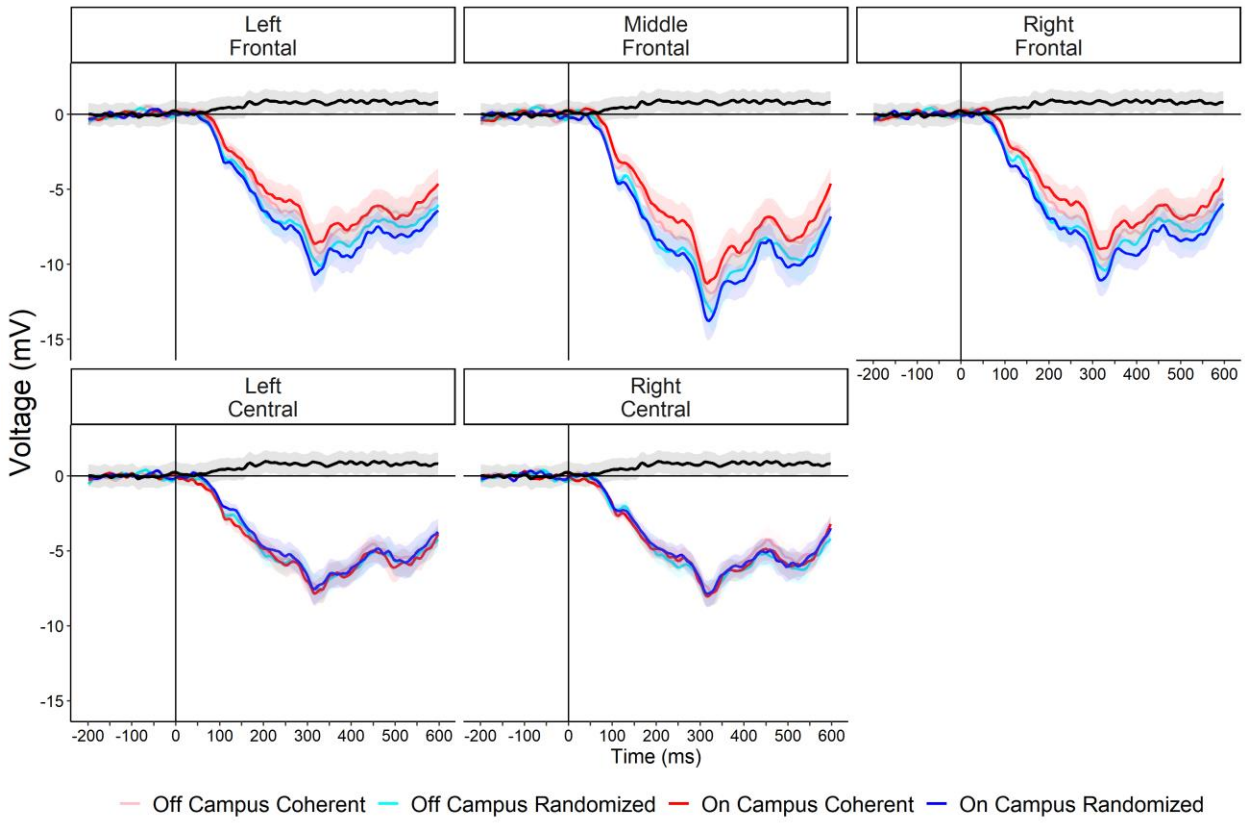


Figure 33. Exp 3: Grand average vERP waveforms time locked to the ***target*** image for the Frontal/Central electrodes. Responses to target images in the coherent condition are in red and responses to the images in the randomized condition are in blue. The difference between the coherent and randomized lines are represented by the black line in the figure.

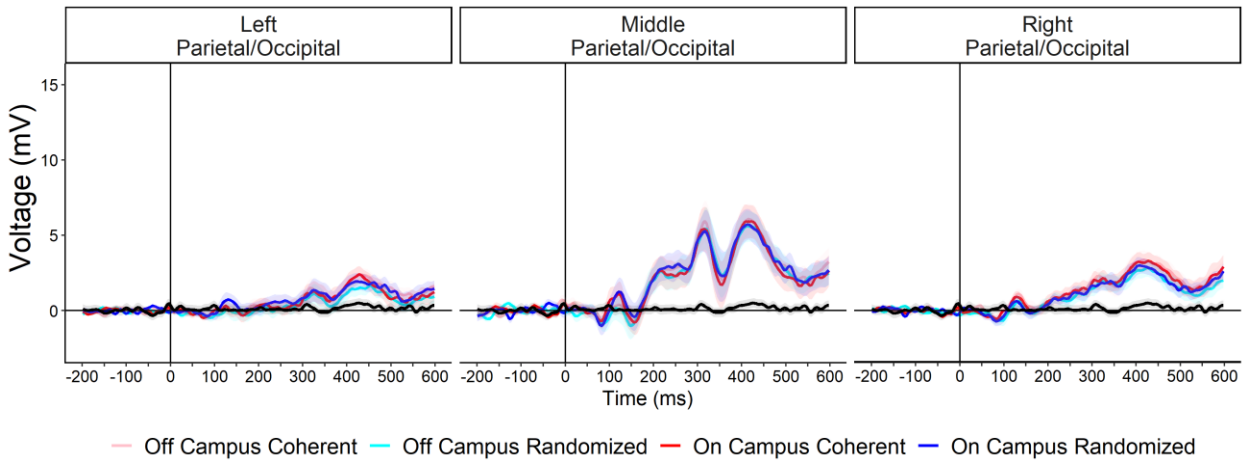


Figure 34. Exp 3: Grand average vERP waveforms time locked to the ***target*** image for the Parietal/Occipital electrodes. Responses to target images in the coherent condition are in red and responses to the images in the randomized condition are in blue.

Frontal/Central Electrode Sites.

We averaged amplitudes at each time point in the epoch across trials for each of the electrodes, and then again across the regions. We then averaged amplitudes within each of the windows (50-149, 150-249, 250-449) after the scene onset and submitted the averages to 3 different linear mixed effects models (differing only in the values of the time bins). Each model included the same fixed and random effects used in Experiment 2. Least square means of amplitude at each window are shown in Figure 35 and results from each model are provided in Table 11. Early facilitation accounts propose that predictions facilitate the integration of local features that form scenes. As such, predictions should influence scene processing very early (50-150 ms). Matching accounts predict that differences appear later (150-249 ms). Feed-forward models predict that vERPs would not differ, or that they only would very late in the epoch, possibly during semantic integration processes indexed by the N400 (Ganis & Kutas, 2003;

Hagoort et al., 2009). To foreshadow the results, analysis of vERPs to the target supported matching accounts of facilitation as they did in Experiment 2, and the hypothesis that scenes in coherent sequences are easier to integrate into the event model.

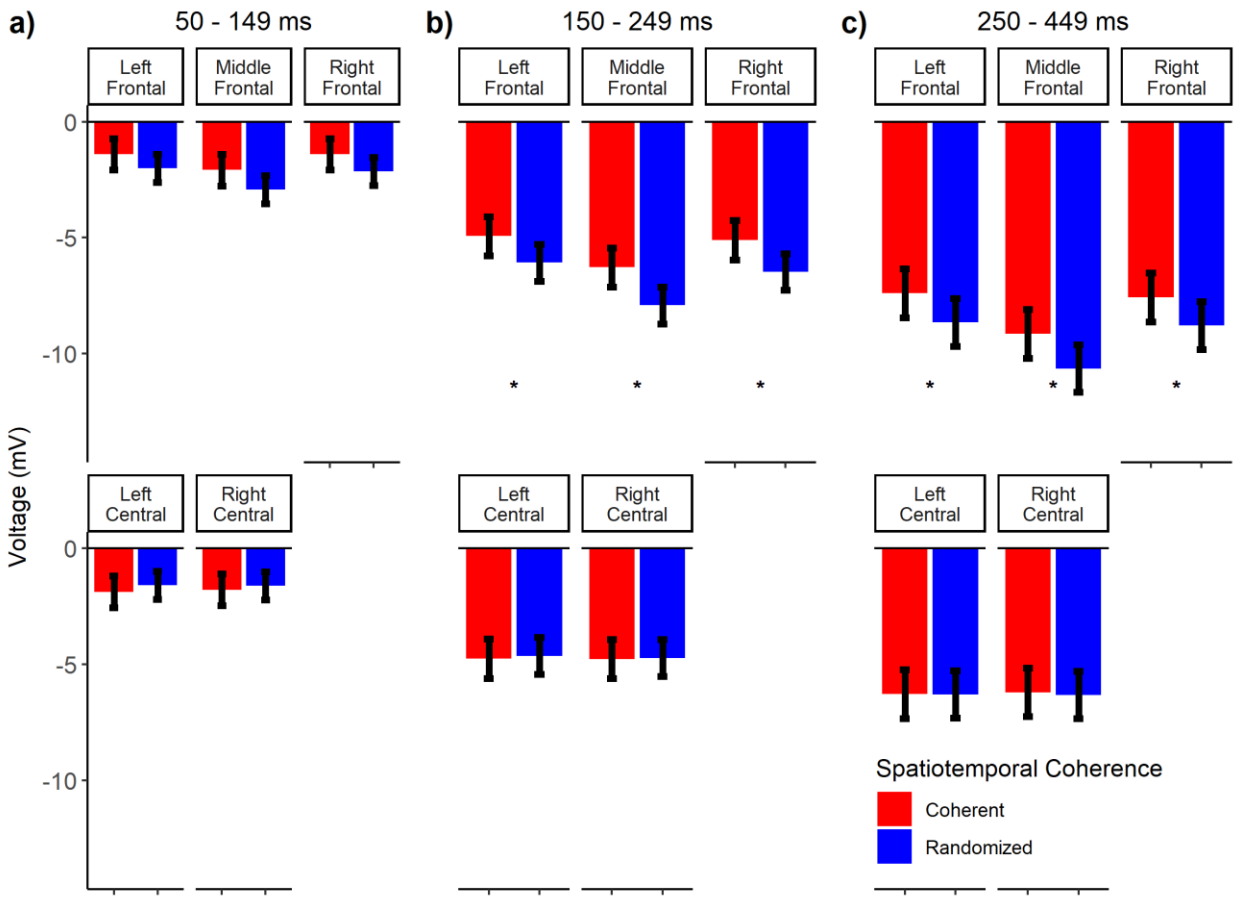


Figure 35. Exp 3: Least square means of amplitudes in response to the ***target*** image at the frontal and central sites. Amplitudes are reported for the a) 50-149, b) 150-249, and c) 250-449 windows.

Table 11. Exp 3: *Summary of the results for the frontal/central electrodes. Amplitudes were time locked to the target image.*

Window	Factor	*df*	*F*	*β*	*t*	*p*
50-149 ms	Region	4,384	2.78			.03*
	SC	1,24	0.59	-0.07	-0.33	.45
	Location	1,24	0.24	-0.24	-0.95	.63
	Region*SC	4,384	1.57			.18
	Region*Location	4,384	0.03			.99
	SC*Location	1,24	0.06			.82
	Region*SC*Location	4,384	2.06			.09
150-249 ms	Region	4,384	46.97			<.001*
	SC	1,24	24.78	-0.97	-4.41	<.0001*
	Location	1,24	0.8	-0.14	-1.29	.38
	Region*SC	4,384	8.23			<.0001*
	Paired t-tests (for Region*SC)					
	Left Frontal			1.15	3.8	<.001*
	Middle Frontal			1.64	5.43	<.0001*
	Right Frontal			1.37	4.52	<.0001*
	Left Central			-0.11	-0.38	.99
	Right Central			-0.04	-0.14	.99
	Region*Location	4,384	0.12			.98
	SC*Location	1,24	0.76			.39
	Region*SC*Location	4,384	1.37			.24
250-449 ms	Region	4,384	63.00			<.001*
	SC	1,24	18.36	-0.23	4.29	<.001*
	Location	1,24	0.00	-0.27	0.04	.97
	Region*SC	4,384	3.33			.01*

Paired t-tests (for Region*SC)				
Left Frontal			1.26	.01*
Middle Frontal			1.5	.001*
Right Frontal			1.22	.01*
Left Central			0.02	.99
Right Central			0.11	.99
Region*Location	4,384	0.02		.99
SC*Location	1,24	1.91		.17
Region*SC*Location	4,384	1.28		.28

Note: SC = Spatiotemporal coherence of scene sequences. * denotes p < .05.

50-149 ms window.

See Figure 35a). Consistent with what we observed in Experiment 2, amplitudes were significantly more positive at the central than the frontal regions [Right Frontal (*M* = -1.77, *SE* = 0.30); Middle Frontal (*M* = -2.50, *SE* = 0.30); Left Frontal (*M* = -1.70, *SE* = 0.30); Right Central (*M* = -1.70, *SE* = 0.30); Left Central (*M* = -1.73, *SE* = 0.30)], $F(4,384) = 2.78$, $p = .03$, BF = 8.79. As evident in Figure 36a), responses to targets in the coherent (*M* = -1.71 *SE* = 0.31) and randomized (*M* = -2.05, *SE* = 0.28) sequences did not significantly differ, inconsistent with early facilitation accounts, but consistent with feed-forward models of scene perception. Further, this lack of a significant difference was supported by a small Bayes factor in favor of the alternative hypothesis, $F(1,24) = 0.59$, $p = .45$, BF = .55, and a small Bayes factor in favor of the null, BF = 1.81. None of the remaining effects were statistically significant. See Table 11 for details.

150-249 ms window.

As shown in Figure 35b), the pattern of differences in the amplitudes were similar, though not identical, to what we observed in Experiment 2. We again found a main effect of region, $F(4,384) = 46.97$, $p < .001$, BF > 1,000. We found in Experiment 2 that responses were

more positive in coherent than randomized sequences consistent with matching accounts of facilitation. Importantly, we replicated this effect in Experiment 3. Amplitudes in coherent (M = -5.17, SE = 0.83) sequences were more positive than amplitudes in randomized (M = -5.97, SE = 0.77) sequences, $F(1,24) = 24.78$, $p < .0001$, BF = 98.97. This effect was larger than what we observed in Experiment 2 ($F(1,24) = 31.84$, $p < .001$, BF = 36.48). The larger difference in rapid scene categorization accuracy may have been due to a larger difference in the waveforms between coherent and randomized sequences. We also found a significant interaction, which we did not observe in Experiment 2, between spatiotemporal coherence and the region, $F(4,384) = 8.23$, $p < .001$, BF = 3.48. The difference in amplitudes between coherent and randomized sequences was larger in the frontal regions [Right Frontal, $\beta = 1.37$, $SE = 0.30$, $t = 4.52$, $p = .0001$; Middle Frontal, $\beta = 1.64$, $SE = 0.30$, $t = 5.43$, $p < .001$; Left Frontal, $\beta = 1.15$, $SE = 0.30$, $t = 3.80$, $p = .0003$] than at central regions [Right Central, $\beta = -0.01$, $SE = 0.30$, $t = -0.14$, $p = .99$; Left Central, $\beta = -0.11$, $SE = 0.30$, $t = -0.38$, $p = .99$]. We can only assume that the inclusion of backward masking in Experiment 3 led to the diminished effect of spatiotemporal coherence in the central region relative to the frontal region, in comparison to Experiment 2 which did not include masking. None of the remaining interactions were significant. See Table 11 for details.

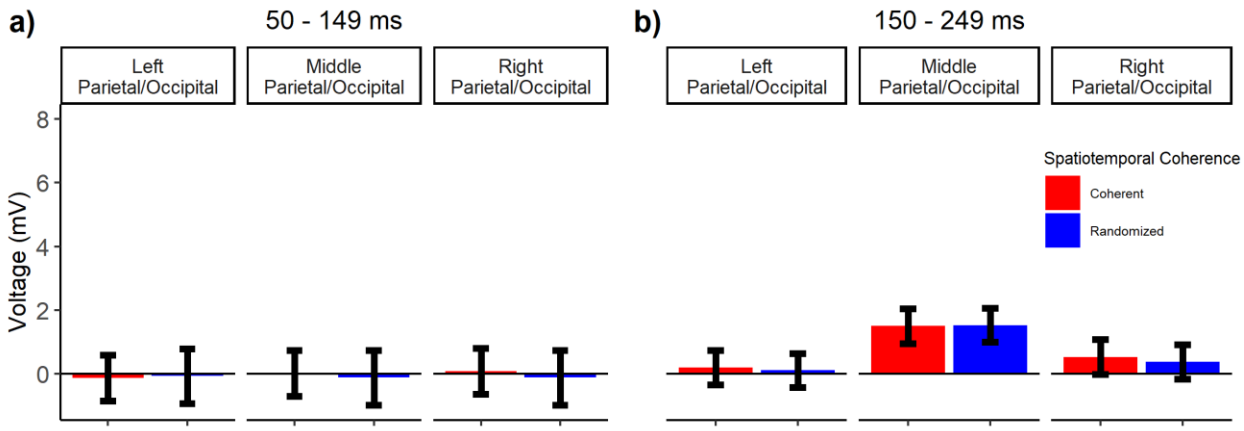
250-449 ms window.

As shown in Figure 35c), we observed analogous effects to what we observed in the previous time window, suggesting that it was easier to integrate information from the current scene into the event model when the sequences were coherent. We observed a significant effect of region, $F(4,384) = 63.00$, $p < .0001$, BF > 1,000; spatiotemporal coherence, $F(1,24) = 18.36$, $p < .001$, BF = 14.81; and an interaction between region and spatiotemporal coherence, $F(4,384) = 3.33$, $p = .04$, BF = 2.18. Interestingly, this effect was notably smaller than what we observed in

Experiment 2 ($F(1,24) = 24.40$, $p < .001$, BF = 37.64). Together with the results from the 150-249 ms window, this could suggest that the larger difference in rapid scene categorization between the coherent and randomized conditions in Experiment 3 may have been driven by more facilitation occurring when matching the structural description to a representation in semantic memory. Again, the difference in the average amplitudes of coherent (M = -7.32, SE = 1.03) and randomized (M = -8.14, SE = 1.00) sequences was significant in the frontal regions [Right Frontal, β = 1.22, SE = 0.39, t =3.11, p = .01; Middle Frontal, β = 1.50, SE = 0.39, t = 3.81, p = .001; Left Frontal, β = 1.27, SE = 0.39, t =3.23, p = .01] but not in the central regions [Right Central, β = 0.11, SE = 0.39, t = 0.30, p = .99; Left Central, β = 0.02, SE = 0.39, t = 0.01, p = .99]**.** The interaction was consistent with what we found in Experiment 2. None of the remaining effects were statistically significant. See Table 11 for details.

Parietal/Occipital Electrode Sites.

As in Experiment 2, linear mixed effects models for the early component analysis (50-149 ms) and the analysis of the P200 (150-249 ms) included the same fixed and random effects that were used in Experiment 2. Least square means of amplitude for the parietal/occipital electrode sites are reported in Figure 36, and results of each of the models are reported in Table 12. We found no difference in vERPs in either the early window or the P200 in Experiment 2 consistent with feed-forward accounts. We replicated the null effects in Experiment 3.



1966

1967 *Figure 36.* Exp 3: Least square means of amplitudes in response to the ***target*** scene at the

1968 parietal/occipital regions. Amplitudes are reported for the a) 50-149, and b) 150-249 windows.

1969 Error bars correspond to 1 standard error around the estimated means.

1970

1971 Table 12. Exp 3: *Summary of the results for the parietal/occipital electrodes. Amplitudes were*

1972 *time locked to the target image.*

Window	Factor	*df*	*F*	*β*	*t*	*p*
50-149 ms	Region	2,192	0.11			.90
	SC	1,24	0.26	-0.07	-0.16	.61
	Location	1,24	1.37	0.06	0.17	.25
	Region*SC	2,192	0.29			.75
	Region*Location	2,192	0.35			.71
	SC*Location	1,24	0.20			.66
	Region*SC*Location	2,192	0.05			.95
150-249 ms	Region	2,192	12.90			<.001*
	SC	1,24	0.09	-0.33	-0.57	.76
	Location	1,24	0.02	-0.16	-0.26	.90
	Region*SC	2,192	0.05			.95
	Region*Location	2,192	0.02			.98

SC*Location	1,24	0.30	.59
Region*SC*Location	2,192	0.08	.92

50-149 ms window.

See Figure 36a). Qualitatively, looking at the figure, one can see that amplitudes did not differ between the coherent and randomized conditions consistent with feed-forward models of scene perception, nor did there appear to be differences between the different regions. This lack of a significant difference was confirmed by a nonsignificant effect of region, $F(2,192) = 0.11$, $p = .90$, BF = 0.0008, and spatiotemporal coherence, $F(1,24) = 0.26$, $p = .61$, BF = 0.03. These results are inconsistent with early facilitation accounts, but consistent with feed-forward accounts and with the results reported in Experiment 2. None of the remaining effects were significant. See Table 12.

150-249 ms window.

We also evaluated if the P200 was more positive in response to scenes shown in the randomized condition. McLean et al. (2021) found that the P200 was more positive for unexpected than expected scenes; however, we failed to find differences in the P200 time-locked to the onset of the target in Experiment 2. As mentioned previously, this may be due to the fact that McLean et al. (2021) showed participants different scenes in their expected versus unexpected conditions, and the P200 is sensitive to layout information (Harel et al., 2016). Consistent with Experiment 2, we found an effect for region, $F(2, 192) = 12.90$, $p < .001$, BF > 1,000, but responses to target scenes in coherent ($M = 0.74$, $SE = 0.50$) and randomized ($M = 0.67$, $SE = 0.48$) sequences did not significantly differ, $F(1, 24) = 0.09$, $p = .76$, BF = 0.10. Thus, our results of the analysis of parietal/occipital electrodes was consistent with feed-forward

accounts of scene processing (Serre et al., 2007; VanRullen, 2007). None of the remaining effects were statistically significant. See Table 12.

vERPs to all of the images

As in Experiment 2, we also evaluated the three vERP components of interest time locked to the onset of *all the scenes*, regardless of their target status. We removed the first scene within each trial from these analyses and behaviorally incorrect trials. Scalp maps of voltage differences across the conditions are shown in Figure 37. Scalp maps were similar to what we observed in Experiment 2. Again, the difference in the scalp maps between the coherent and randomized sequences appear to be localized in the frontal and central regions, with the difference being much stronger in the frontal regions, beginning in the 150-249 millisecond window. This difference was confirmed statistically and details of those results are reported in the Appendix. This is consistent with when we would expect to see differences according to matching accounts of facilitation (Mudrik et al., 2010; Schendan, 2019; Smith & Federmeier, 2020; Truman & Mudrik, 2018). The difference is again very small over parietal/occipital regions, and it was not significant when we analyzed those results. See the Appendix for details.

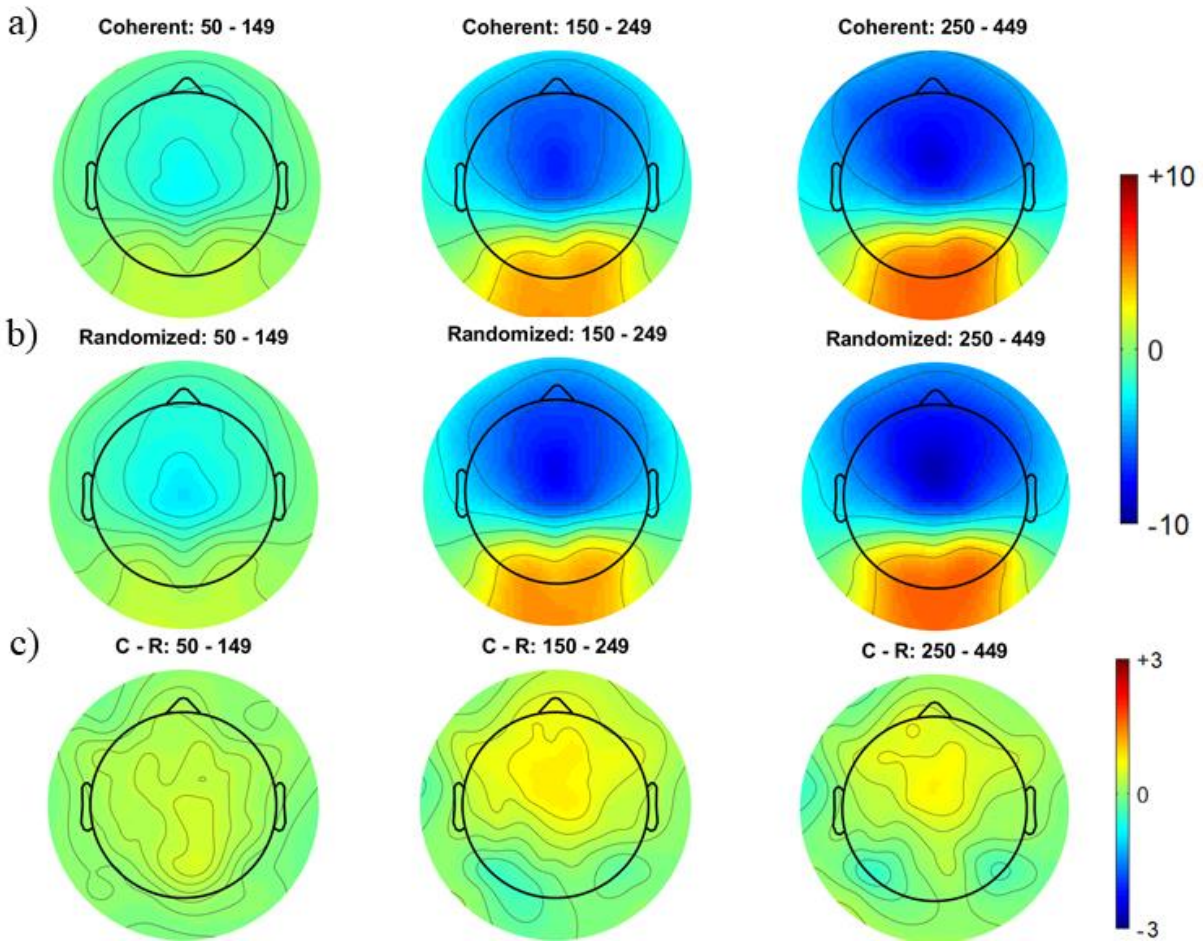


Figure 37. Exp 3: Scalp maps of the mean voltage time locked to the onset of scenes within the a) coherent, b) randomized sequences. The difference between the coherent and randomized conditions are represented in c). Scalp maps do not include behaviorally incorrect trials or responses to the first scene within a trial. Voltage ranged from -10 to +10 microvolts in the coherent and randomized sequences and -5 to +5 in the difference maps.

Analysis of vERP divergence

To be consistent with Experiment 2, we also examined when waveforms in response to scenes presented in coherent and randomized sequences begin to diverge when no prior assumptions were made about the window. Those results are reported in the Appendix. Early

facilitation accounts predict that differences in the vERPs would arise prior to 150 milliseconds (Biederman et al., 1982; Palmer, 1975b). Matching accounts predict that differences would arise between 150 and 250 milliseconds (Bar & Ullman, 1996; Mudrik et al., 2010; Schendan, 2019). Feed-forward accounts predict that the vERPs would either not differ or they would only differ later in the epoch, possibly in components that have previously been associated with semantic integration processes (Ganis & Kutas, 2003). We replicated the differences we observed in Experiment 2 though waveforms diverged a little later than they did in Experiment 2. Waveforms diverged significantly at 152 milliseconds post scene onset over the frontal region, and the effect remained significant until 375 milliseconds. We obtained analogous effects at the central electrode region. Waveforms differed significantly from 195 to 343 milliseconds. Waveforms recorded by parietal/occipital electrodes did not significantly differ. See the Appendix for details.

Exploratory analyses of source localization

As in Experiment 2, we next ran an exploratory analysis to identify the possible neural sources of the difference between vERPs at frontal and central sites. We found that the source of the difference in Experiment 2 could be localized to motor regions, and more importantly, regions that have previously been associated with the construction and maintenance of the event model (Ranganath & Ritchey, 2012; Stawarczyk et al., 2019).

We followed the same steps as we did in Experiment 2 to identify the possible source of the difference between the conditions. We first identified the top 12 independent components that contributed the most to the average scalp distribution between 150 and 249 milliseconds post-stimulus for each participant. We then fit single equivalent current dipoles to each of the 12 components resulting in 252 (21 participants X 12 components) independent components across

participants. We clustered the 12 independent components from each participant using a k-means (k = 12) cluster analysis to identify common independent components across participants using the similarity between each of their vERPs, scalp distributions, and dipole locations. Components that were further than 3 standard deviations from any centroid were categorized into an outlier cluster and omitted from the analyses. We set the value of k to the same value as we used in Experiment 2 to be consistent across experiments. Coordinates of the centroids for each cluster and their labels are available in Table 13. Lastly, to compare vERPs in the coherent and randomized conditions, we back projected data from each of the independent components to the individual channels, and then repeated the analyses reported in the Analysis of vERP divergence sections of this document using the back projected data from each of the components within each cluster.

Table 13. Exp 3: *Montreal Neurological Institute (MNI) coordinates and labels of the centroids of independent component clusters. Cluster are in no particular order.*

Cluster Index	Brodmann Area of Centroid	MNI Coordinates (X, Y, Z)	Number of Components	Number of Participants
1	Brodmann Area 6	(-48, -5, 36)	18	12
2	Brodmann Area 23	(10, -49, 24)	12	11
3	Brodmann Area 7	(-23, -70, 54)	25	16
4	Brodmann Area 31	(10, -43,41)	22	31
5	Brodmann Area 19	(47, -77, -12)	16	13
6	Brodmann Area 6	(11, -6, 60)	27	16
7	Brodmann Area 20	(-53, -1, -41)	4	2
8	Brodmann Area 18	(2, -90, 26)	33	17
9	Brodmann Area 37	(-56, -56, -1)	33	37
10	Brodmann Area 36	(-31, 0, -34)	10	38
11	Brodmann Area 22	(58, -35, 17)	24	22
12	Brodmann Area 10	(-2, 53, 4)	14	10

VERPs from the frontal and central electrodes back projected from each of the 12 clusters of independent components are shown in Figures 38 and 39, respectively. This analysis revealed 4 clusters that contributed to the difference between frontal electrode regions and 2 that contributed to the difference observed in the central regions. Consistent with the results of Experiment 2, waveforms at the frontal electrodes began to diverge significantly in Cluster 2 (localized in Brodmann Area 23) and in a slightly more dorsal region in Cluster 4 (localized in Brodmann Area 31) at 156 milliseconds. Amplitudes remained statistically significant until 394 in Cluster 2 and until 585 milliseconds in Cluster 4. Both of the Brodmann areas make up the posterior cingulate (Strotzer, 2009), which is one of the two locations where we observed significant differences in Experiment 2. Also consistent with the results of Experiment 2, the differences we observed in Clusters 2 and 4 were additionally associated with a significant difference in Cluster 6 (Brodmann Area 6) in both the frontal and central regions. Brodmann area 6 is composed of the primary motor and supplementary motor areas.

As shown in Figure 38, waveforms time locked to the onset of the scenes were additionally associated with a 4th cluster that showed a significant difference between the coherent and randomized sequences starting at 160 milliseconds and lasting until 585 milliseconds. The centroid of Cluster 9 lies within Brodmann area 37. Brodmann area 37 is an inferior occipitotemporal region, which contains the left fusiform gyrus. While the left fusiform gyrus has traditionally been considered to be a face processing region (Blonder et al., 2004; Kanwisher et al., 1997), some work has supported its involvement in categorization and semantic processing (Ardila et al., 2015). For instance, it is involved in accessing the names and meanings of pictures and words (McDermott et al., 2003; Usui et al., 2003). Importantly, some work has found that neurological responses increase in the left fusiform gyrus when one reads coherent

sentences compared to those with words presented in a random order, which suggests that it may play a role in binding meaningful components into one coherent situation (Vandenberghe et al., 2002).

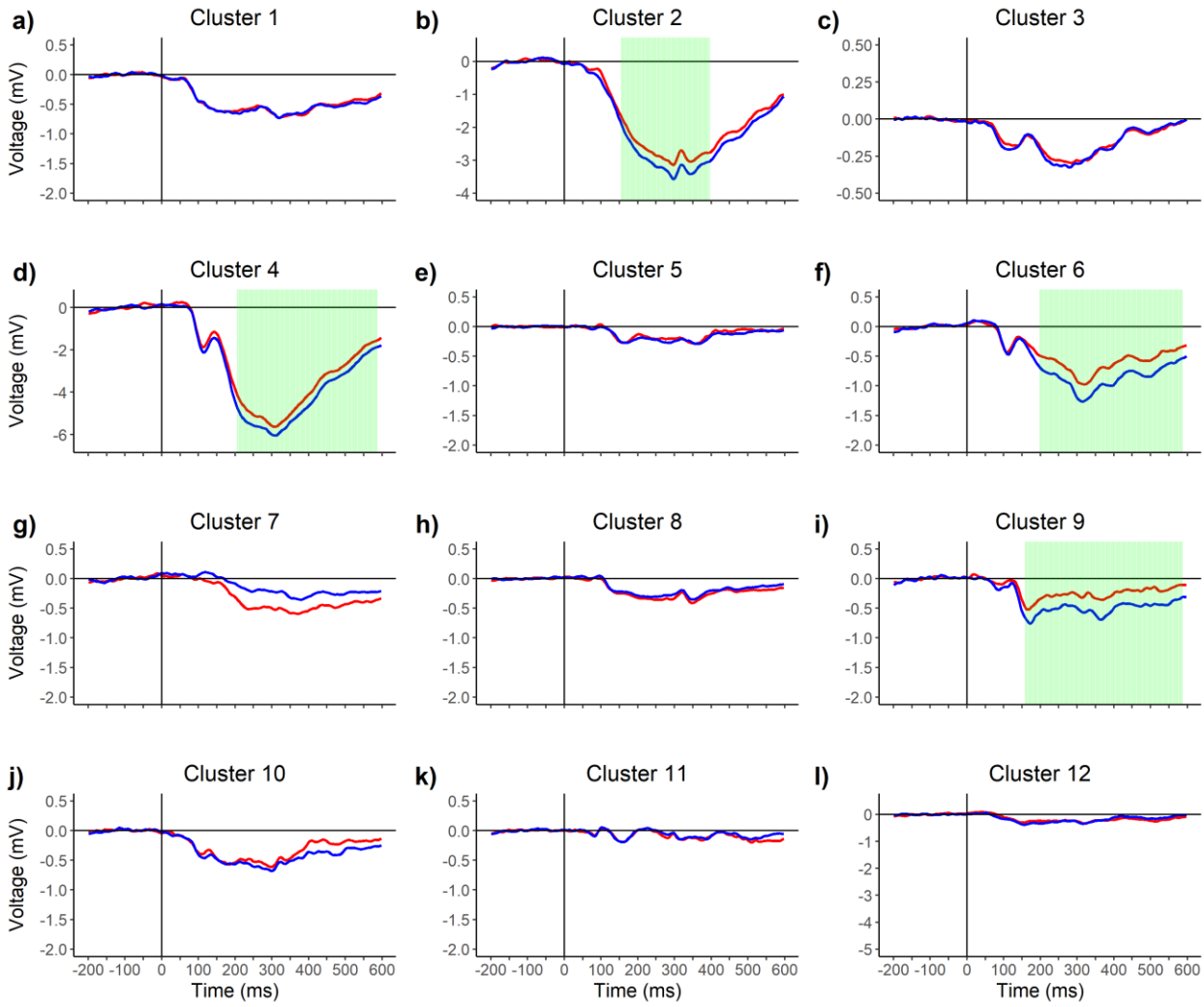


Figure 38. Exp 3: Grand average waveforms at frontal regions time locked to the onset of the scenes back projected from each of the 12 clusters of independent components. Waveforms in response to scenes shown in the coherent sequences are represented in red and waveforms in response to images shown in the randomized sequences are represented in blue. Green patches represent clusters of statistically significant comparisons.

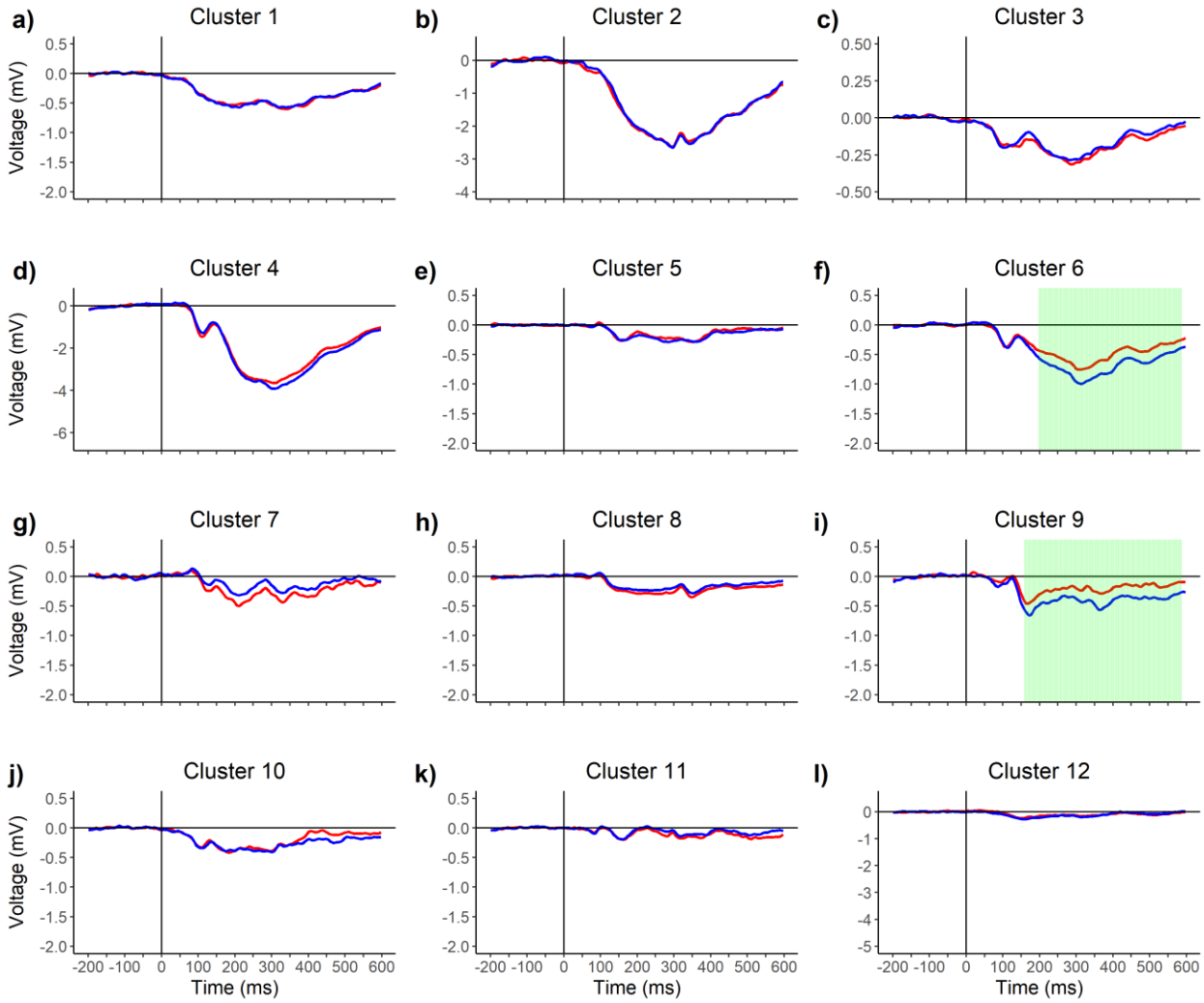


Figure 39. Exp 3: Grand average waveforms at central electrodes time locked to the onset of the images back projected from each of the 12 clusters of independent components. Waveforms in response to scenes shown in the coherent sequences are represented in red and waveforms in response to images shown in the randomized sequences are represented in blue. Green patches represent clusters of significant comparisons.

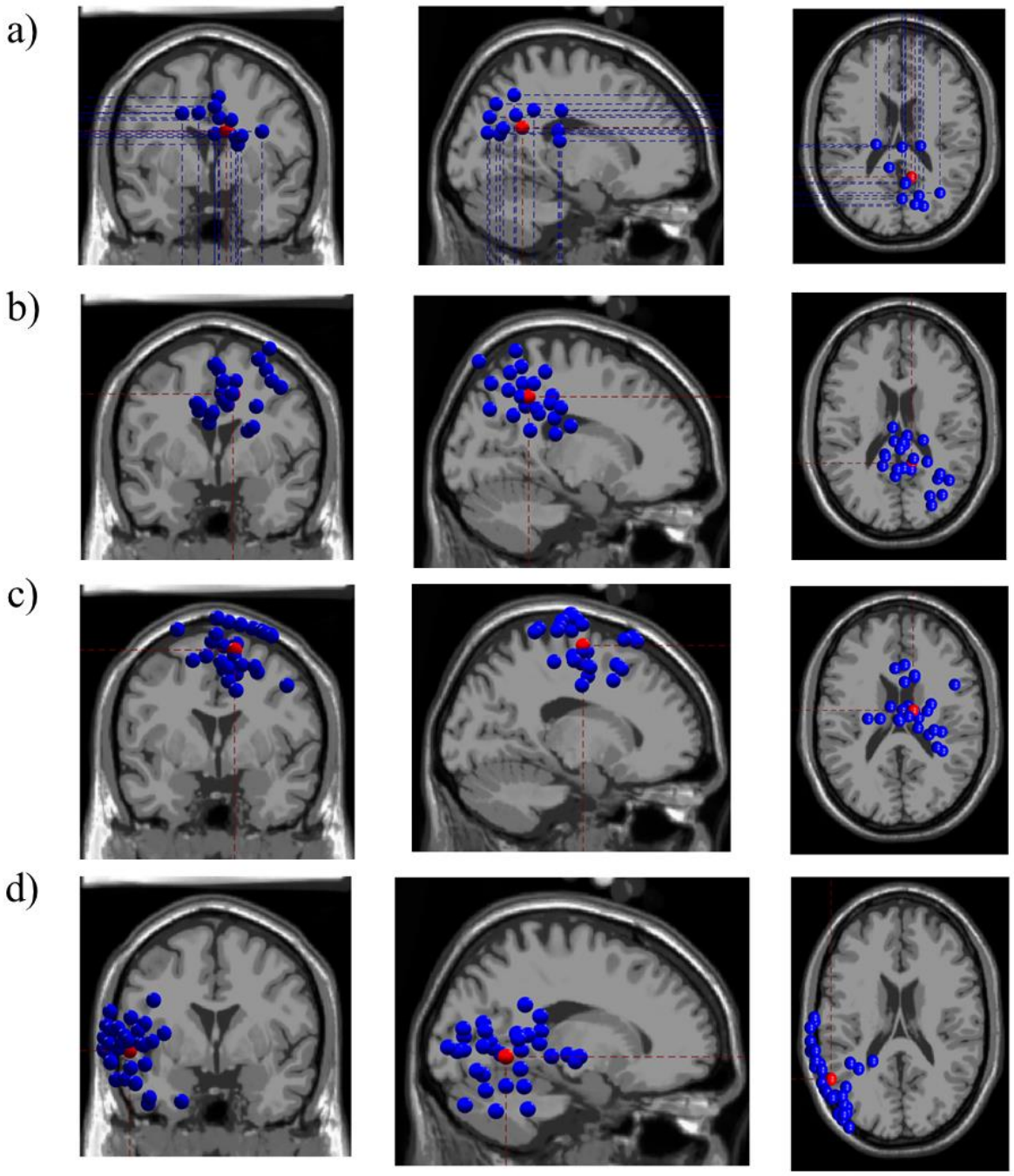


Figure 40. Exp 3: Clusters of sources of independent components for all subjects across trials and conditions for a) Cluster 2 (Brodmann area 23), b) Cluster 4 (Brodmann are 31), c) Cluster 6 (Brodmann area 6), and d) Cluster 9 (Brodmann area 37).

As shown in Figure 39 and consistent with data reported at frontal electrodes, waveforms at the central region, back projected from independent components that made up Cluster 6 (Brodmann area 6) and Cluster 9 (Brodmann area 37) diverged significantly at 199 and 160 milliseconds, respectively. Amplitudes remained significantly different in both Clusters until 585 milliseconds.

Again, we found clusters of components that had centroids localized in early visual areas such as Brodmann area 18 and 19; however, the vERPs from the components within each cluster failed to show differences between the coherent and randomized sequences at the frontal and central regions. Thus, we have no evidence from the vERPs alone that the event model facilitated early perceptual analysis of the scenes.

Changes in vERPs within a trial

We also explored how amplitudes in response to scenes shown in coherent and randomized sequences change over time within a trial. We assume that scene categories become more predictable as the event model is constructed within a trial. In Experiment 3, we found from our behavioral results that the ability to categorize the target scenes (See Figure 32) improved as a function of the ordinal position of the target scene. In addition, amplitudes in response to scenes in Experiment 2 differed after the first image was shown on a trial. Results from Experiment 2 were consistent with SPECT. Namely, that back-end processes involved in construction of the event model feeds back and influences front-end processes involved in information extraction. As such, we also explored changes in vERPs as a function of the ordinal position of the scenes (1-10) in Experiment 3. We again hypothesized that vERPs in response to the first scene on a trial would not differ between the coherent and randomized sequences, but they would afterwards since the first scene on a trial lays the foundation of the event model. We

removed all behaviorally incorrect trials prior to running the analyses. Results are reported in Tables 14 and 15 for each of the 3 windows (50-149, 150-249, 250-449), and least square means of amplitude for each window are shown in Figures 41 and 42. We found in Experiment 2 that the event model did not influence early perceptual analysis of the scenes, but it did influence matching and semantic integration processes.

Frontal and Central Electrodes.

As in Experiment 2, we modeled amplitudes in response to scenes at each ordinal position using linear mixed effects models. Each of the models contained the same fixed and random effects as in Experiment 2. As in Experiment 2, we first ran the analysis by treating the location of the images on each trial as a continuous effect; however, the majority of those models failed to converge with the complex random effect structure that we specified. Thus, as in Experiment 2, we again treated the ordinal position of the scene on each trial as a categorical predictor to keep the analyses consistent across models and across experiments. Model output is shown in Table 14.

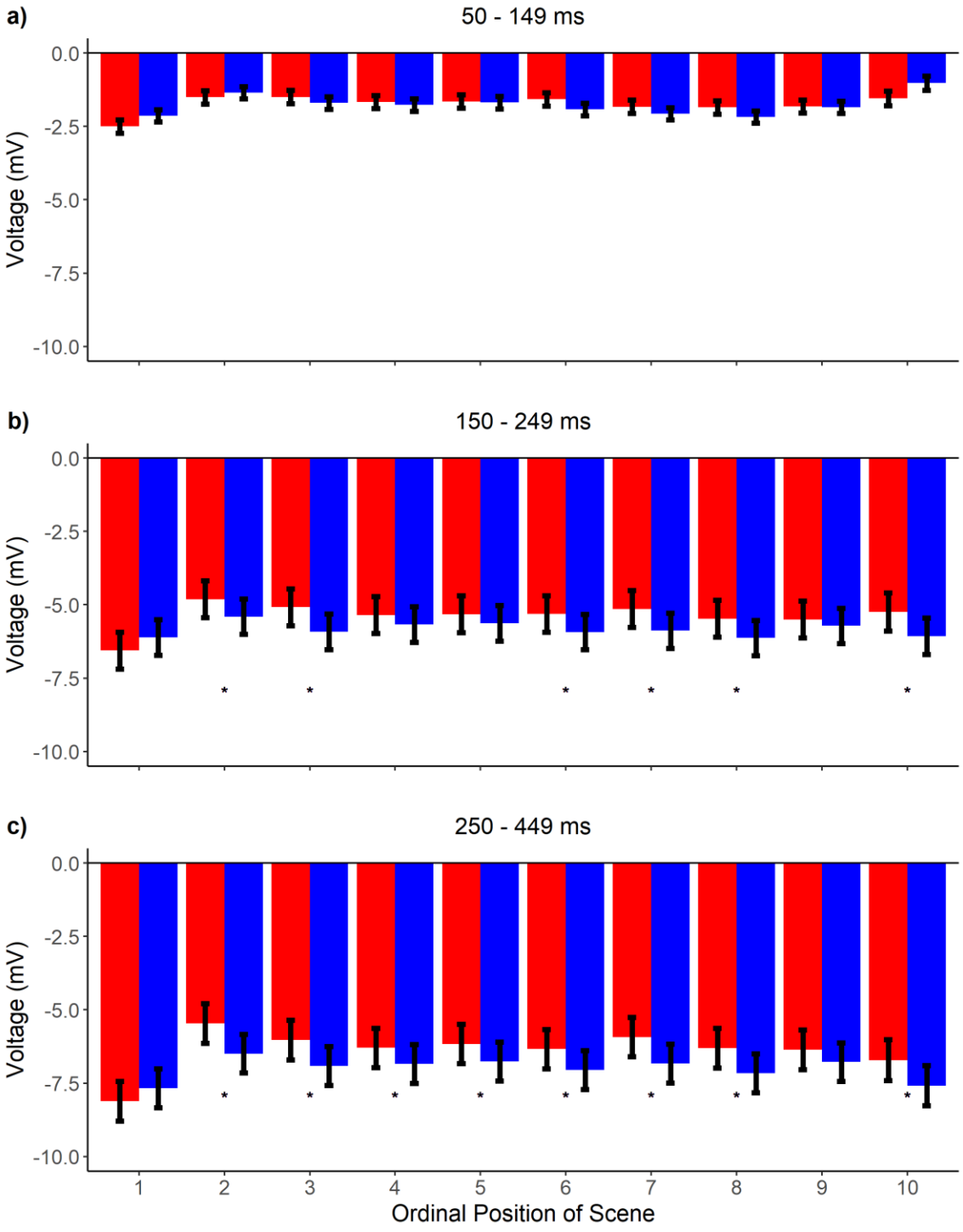


Figure 41. Exp 3: Frontal/Central electrode amplitudes in response to each ordinal position (1-10) of the scenes on a trial, excluding behaviorally incorrect trials. Responses to scenes in coherent sequences are in red, and responses to scenes in randomized sequences are in blue.

Table 14. Exp 3: *Summary of the results for the frontal/central electrodes. Amplitudes were time locked to the onset of the scenes in the experiment in the 1st through the 10th positions.*

Window	Factor	*df*	*F*	*p*
50-149	Region	44,408	50.48	<.001*
	SC	1,24	0.004	.95
	Location	1,25	5.04	.03*
	Ordinal Position	94,408	15.79	<.001*
	Region*SC	44,408	0.31	.87
	Region*Location	44,408	1.24	.29
	SC*Location	14,408	1.54	.22
	Region* Ordinal Position	364,408	1.57	.02*
	SC* Ordinal Position	94,410	1.02	.42
	Location* Ordinal Position	94,394	6.47	<.001*
	Region*SC*Location	44,408	0.19	.94
	Region*SC* Ordinal Position	364,408	0.15	.99
	Region*Location* Ordinal Position	364,408	0.89	.65
	SC*Location* Ordinal Position	94,408	1.09	.37
	Channels*SC*Location* Ordinal Position	364,408	0.16	.99
150-249	Region	4.4408	239.82	<.001*
	SC	1,24	11.08	.003*
	Location	1,25	0.89	.36
	Ordinal Position	94,408	9.90	<.001*
	Region*SC	44,408	0.37	.83
	Region*Location	44,408	3.00	.02*
	SC*Location	14,408	12.67	<.001*
	Region* Ordinal Position	364,408	1.69	.006*
	SC* Ordinal Position	94,410	3.11	<.001*
	Location* Ordinal Position	94,394	8.34	<.001*

	Region*SC*Location	44,408	0.07	.99
	Region*SC* Ordinal Position	364,408	0.20	.99
	Region*Location* Ordinal Position	364,408	1.26	.14
	SC*Location* Ordinal Position	94,408	0.79	.63
	Channels*SC*Location* Ordinal Position	364,408	0.09	.99
250-449	Region	4.4408	239.75	<.001*
	SC	1,24	16.70	<.001*
	Location	1,25	2.54	.13
	Ordinal Position	94,408	19.57	<.001*
	Region*SC	44,408	0.77	.54
	Region*Location	44,408	5.58	<.001*
	SC*Location	14,408	16.29	<.001*
	Region* Ordinal Position	364,408	1.65	.01*
	SC* Ordinal Position	94,410	3.46	<.001*
	Location* Ordinal Position	94,394	11.76	<.001*
	Region*SC*Location	44,408	0.09	.99
	Region*SC* Ordinal Position	364,408	0.23	.99
	Region*Location* Ordinal Position	364,408	1.91	<.001*
	SC*Location* Ordinal Position	94,408	2.14	.02*
	Channels*SC*Location* Ordinal Position	364,408	0.15	.99

50-149 ms window.

See Figure 41a). Consistent with the results of Experiment 2, we observed a significant main effect for region, $F(44,408) = 50.48$, $p < .001$, BF > 1,000, and the ordinal position of the scene within a trial (1-10), $F(94, 408) = 15.79$, $p < .001$, BF > 1,000. We also observed a significant interaction between where the scenes were photographed and the ordinal position of the scene, which we also observed in Experiment 2, $F(94,394) = 6.47$, $p < .001$, BF >1,000.

Amplitudes in on-campus sequences were significant more positive than off-campus sequences when the scene was the 1st , β = 0.59, *SE* = 0.15, *t* = 3.85, *p* = .001 and 10th scene on trial, β = 1.08, *SE* = 0.25, *t* = 4.39, *p* = .001**.** Importantly, we found no evidence to suggest that amplitudes in response to images shown in coherent (*M* = -1.72, *SE* = 0.29) sequences differed from those shown in randomized (*M* = -1.73, *SE* = 0.29) sequences, *F*(1,24) = 0.004, *p* = .95, BF = 0.002 consistent with feed-forward models. Thus, we have no evidence to suggest from the analysis of the vERPs that predictions made prior to viewing a scene facilitates the integration of features that make up scenes as early facilitation accounts hypothesize.

150-249 ms window.

See Figure 41b). Again, we observed a significant main effect for region, *F*(4,408) = 239.82, *p* <.001, BF >1,000 and the ordinal position of the scene on the trials, *F*(94,408) = 9.90, *p* <.001, BF >1,000. Consistent with Experiment 2 and our hypothesis that the event model facilitates processes involved in matching the structural description to representations stored in semantic memory, we also observed a significant main effect of spatiotemporal coherence, *F*(1,24) = 11.08, *p* = .003, BF = 4.09, and an interaction between spatiotemporal coherence and the ordinal position of the scene on each trial, *F*(91,394) = 8.34, *p* <.001, BF = 9.07. Amplitudes did not significantly differ between coherent and randomized sequences for the first scene on a trial, β = -0.20, *SE* = 0.22, *t* = -0.89, *p* = .37, but amplitudes were significantly more positive in the coherent sequences at the 2nd 3rd, 6th, 7th, 8th, and 10th ordinal position (all Bonferroni corrected p values < .05). Amplitudes were numerically more positive in the coherent sequences at the remaining positions. Unlike Experiment 2, we did not observe a significant three-way interaction between spatiotemporal coherence, the location where the image was photographed, and the ordinal position of the scene on a trial, *F*(364,408) = 1.26, *p* = .14, BF = 2.17. The lack

of a three-way interaction is not surprising considering that this effect was previously associated with a small Bayes factor in favor of the alternative hypothesis (Experiment 2 BF = 2.98, Experiment 3 BF = 2.17).

250-449 ms window.

See Figure 41c). Effects were all consistent with what we observed when we analyzed the previous window except that we also observed a statistically significant three-way interaction between spatiotemporal coherence, the ordinal position of the image, and the location where the images were photographed (on-campus vs. off-campus), $F(94,408) = 2.14$, $p = .02$, BF = 34.75. Consistent with the hypothesis that the first scene within a trial lays the foundation of the event model, the N400 did not significantly differ between coherent and randomized sequences for the first image in a sequence in both the on-, $\beta = -0.29$, $SE = 0.35$, $t = -0.83$, $p = .40$ and off-campus sequences, $\beta = -0.60$, $SE = 0.35$, $t = -1.72$, $p = .08$. The remaining scenes were easier to integrate into the event model when they were shown in coherent sequences. The N400 was more positive in response to scenes shown in the coherent sequences at all of the remaining ordinal positions in the on-campus sequences and was more positive in the 2nd, 3rd, and 8th ordinal positions in the off-campus sequences. Amplitudes were numerically more positive in the coherent sequences in the positions that did not show a statistically significant difference. The remaining statistically significant interactions were the same as those observed in the previous window See Table 14.

Parietal/Occipital Electrodes.

Linear mixed effects models for the parietal/occipital electrode sites contained the same fixed and random effects as the models conducted on the frontal and central sites with the exception that the analyses were conducted on the amplitudes from the parietal/occipital regions. Least square means from each of the models are shown in Figure 42, and output is provided in

Table 15. We did not find evidence in Experiment 2 to suggest that scenes shown in coherent and randomized sequences were processed differently in the early component, but the P200 was more positive in the randomized sequences at a few of the ordinal positions consistent with the results of McLean et al. (2021). We found analogous effects in Experiment 3.

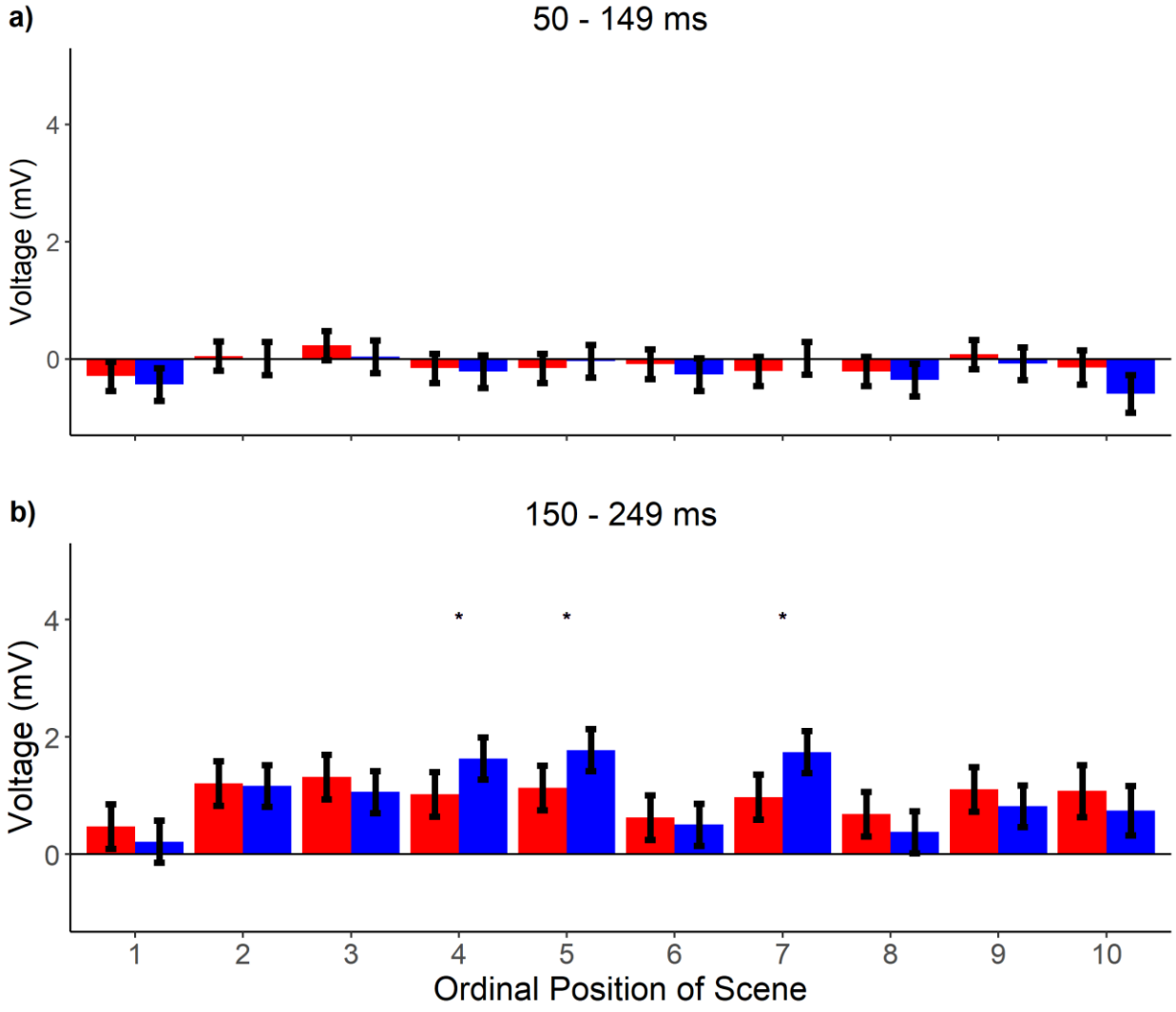


Figure 42. Exp 3: Parietal/Occipital electrodes average amplitudes time locked to the onset of scenes shown at each ordinal position. Behaviorally incorrect trials were removed from this analysis. Responses to images in coherent sequences are in red, and responses to images in randomized sequences are in blue.

Table 15. Exp 3: *Summary of the results for the parietal/occipital electrodes. Amplitudes were time locked to the onset of the images in the experiment in the 1st – 10th position.*

Window	Factor	*df*	*F*	*p*
50-149	Region	22,640	0.09	.91
	SC	1,24	0.74	.70
	Location	1,36	0.12	.73
	Ordinal Position	92,640	2.40	.01*
	Region*SC	22,640	0.32	.72
	Region*Location	22,640	0.59	.55
	SC*Location	12,640	0.89	.35
	Region* Ordinal Position	182,640	0.96	.50
	SC* Ordinal Position	92,643	0.59	.80
	Location* Ordinal Position	92,640	0.75	.66
	Region*SC*Location	22,640	0.01	.99
	Region*SC* Ordinal Position	182,640	0.25	.99
	Region*Location* Ordinal Position	182,640	0.76	.75
	SC*Location* Ordinal Position	92,640	0.42	.93
	Channels*SC*Location* Ordinal Position	182,640	0.14	.99
150-249	Region	22,640	92.40	<.001*
	SC	1,24	0.14	.71
	Location	1,36	0.93	.34
	Ordinal Position	92,640	7.40	<.001*
	Region*SC	22,640	0.26	.77
	Region*Location	22,640	2.81	.06
	SC*Location	12,640	0.05	.82
	Region* Ordinal Position	182,640	0.92	.55

SC* Ordinal Position	92,643	2.31	.01*
Location* Ordinal Position	92,640	0.56	.83
Region*SC*Location	22,640	0.41	.67
Region*SC* Ordinal Position	182,640	0.20	.99
Region*Location* Ordinal Position	182,640	0.70	.81
SC*Location* Ordinal Position	92,640	0.41	.93
Channels*SC*Location* Ordinal Position	182,640	0.27	.99

50-149 ms window.

See Figure 42a). Consistent with the results of Experiment 2 using the average amplitudes in the 50-149 ms window, we observed a significant effect for the ordinal position of the scene on a trial, $F(92,640) = 2.40$, $p = .01$, BF = 12.97 such that amplitudes of the 1st (M = -0.36, SE = 0.34) and 10th (M = -0.67, SE = 0.37) scenes were significantly more negative than amplitudes in response to the remaining scenes [Second (M = 0.03, SE = 0.34); Third (M = 0.13, SE = 0.34); Fourth (M = -0.19, SE = 0.34); Fifth (M = -0.10, SE = 0.34); Sixth (M = -0.18, SE = 0.34); Seventh (M = -0.10, SE = 0.34); Eighth (M = -0.28 , SE = 0.34); Ninth (M = -0.0005, SE = 0.34)]. Consistent with feed-forward accounts, we did not observe a significant effect for spatiotemporal coherence, $F(1,24) = 0.74$, $p = .40$, BF = 0.004, nor an interaction between spatiotemporal coherence and the ordinal position of the scenes on a trial, $F(92, 643) = 0.59$, $p = .80$, BF = 1.97. Thus, consistent with feed-forward accounts of scene processing, we have no evidence from the analyses of the early component to suggest that the event model facilitates early perceptual analysis. None of the remaining effects were statistically significant. See Table 15 for details.

150-249 ms window.

See Figure 42b). Results were consistent with what we observed in Experiment 2, and thus inconsistent with feed-forward accounts of scene perception. We found a statistically significant main effect on the P200 for region, $F(22, 640) = 92.40$, $p < .001$, BF >1,000; and for the ordinal position of the scene on each trial, $F(92, 640) = 7.40$, $p < .001$, BF >1,000. More importantly, we found a significant interaction between spatiotemporal coherence and the ordinal position of the scenes, $F(92,643) = 2.31$, $p = .01$, BF = 4.82. Amplitudes were significantly more positive to scenes shown in the randomized sequences for the 4th, 5th, and 7th ordinal positions. Comparisons between the coherent and randomized conditions were not significant at any of the remaining ordinal positions after adjusting the p values for multiple comparisons. Thus, we found some evidence in both Experiment 2 and 3 to suggest that predictions made prior to viewing a scene facilitates the P200 observed at the parietal/occipital electrodes (McLean et al., 2021)). None of the remaining interactions were statistically significant. See Table 15 for details.

Exploratory analyses of image predictability and image similarity

As in Experiment 2, we also evaluated how the visual similarity between the target and prime, as quantified from their shared spatial envelope, and image predictability from Experiment 1 correlated with voltage within each window for scenes shown in coherent and randomized sequences. As mentioned previously, it is possible that the activation of similar features in the prime and target resulted in the benefit observed in the coherent sequences. Alternatively, both the predictability of the scenes and their shared spectral information could have produced the benefit in coherent sequences. Our results supported the later hypothesis. As in Experiment 2, we removed behaviorally incorrect trials before running the analyses. Partial correlations are provided in Table 16, and scatterplots are provided in Figures 43 through 46.

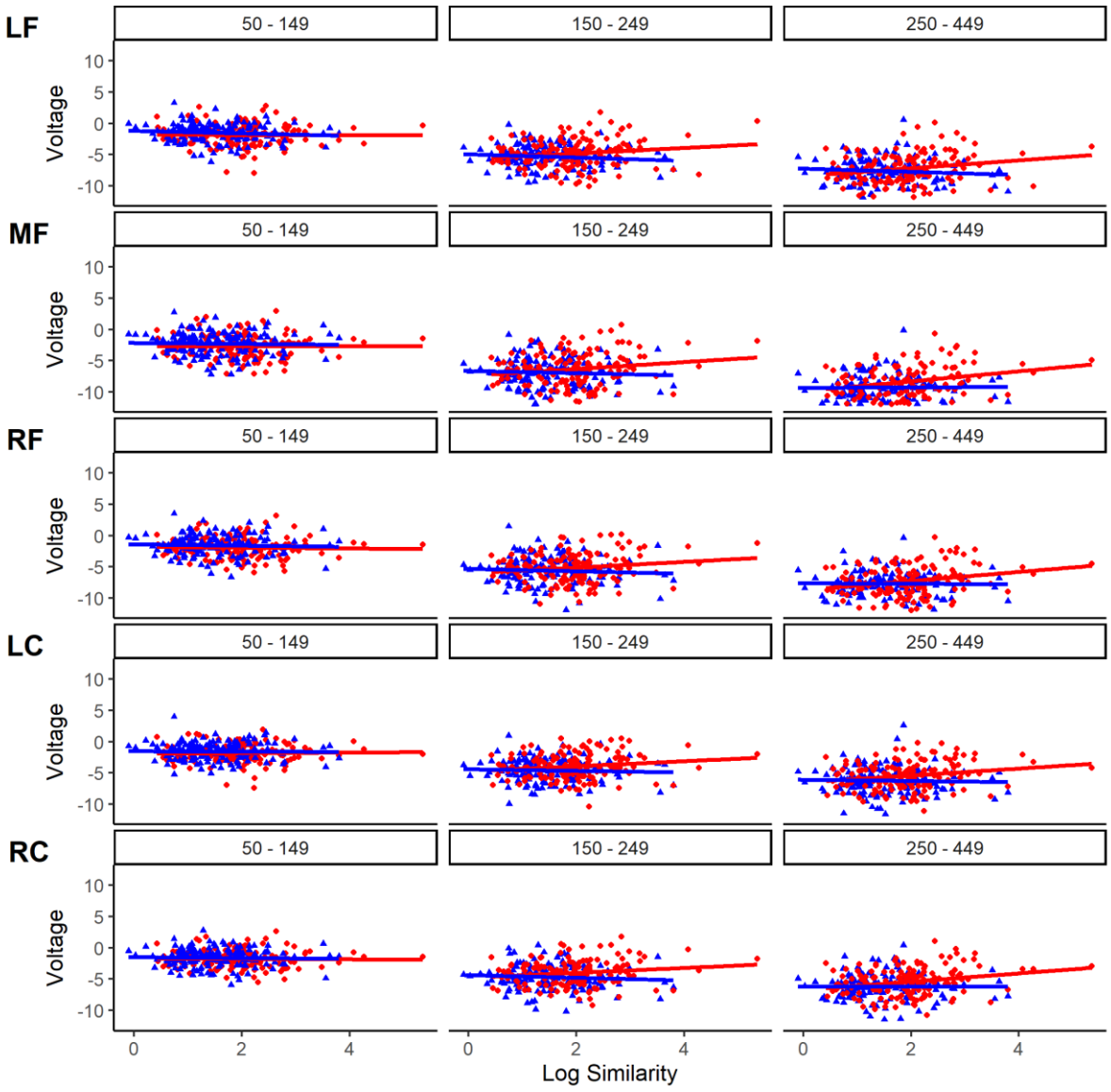


Figure 43. Exp 3: Scatterplots between ***image similarity*** and voltage at LF) left, MF) middle, and RF) right ***frontal*** as well as LC) left and RC) right ***central*** regions.

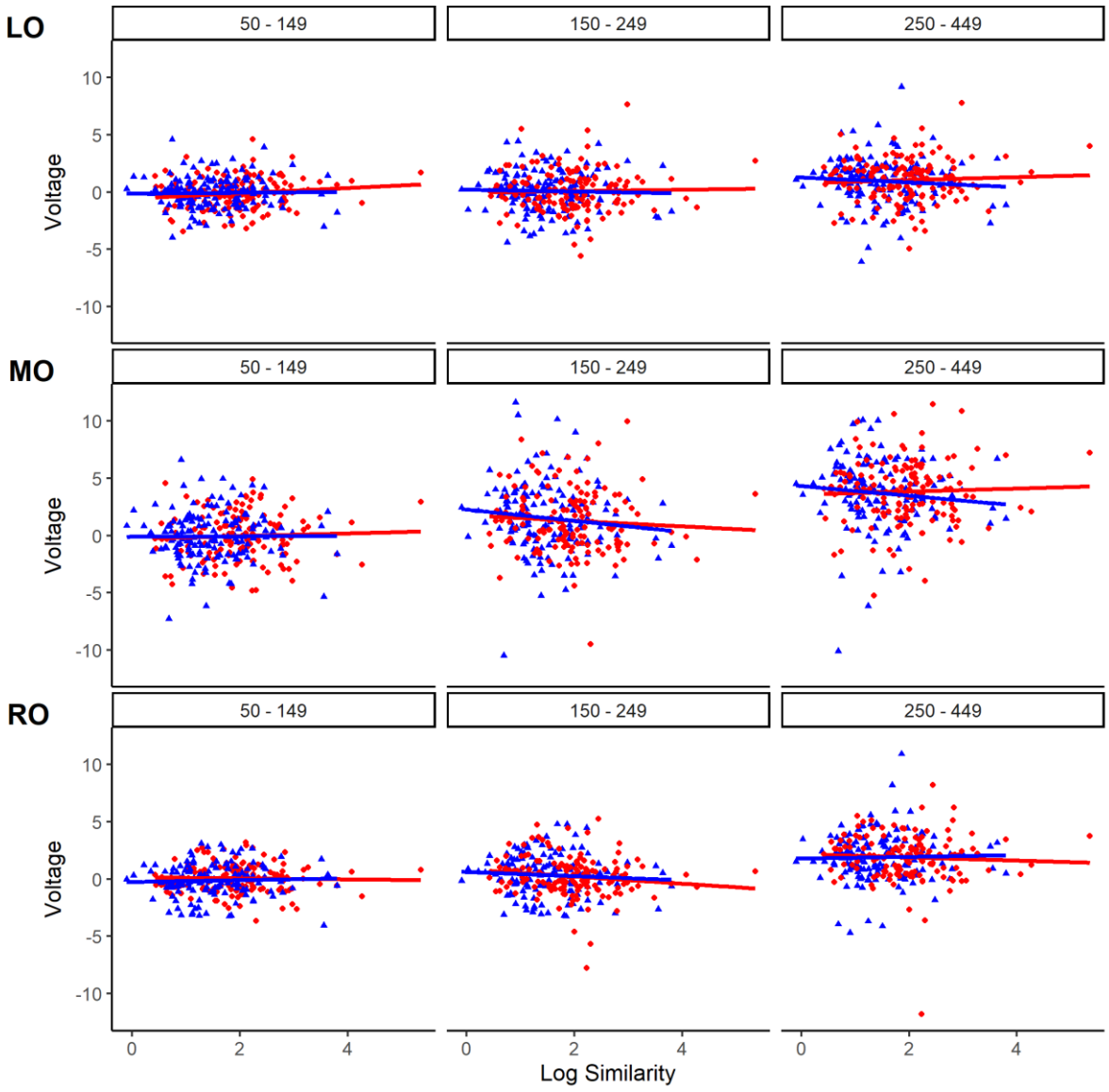


Figure 44. Exp 3: Scatterplots between the log of ***image similarity*** and voltage at LO) left, MO) middle, and RO) right ***parietal/occipital*** regions.

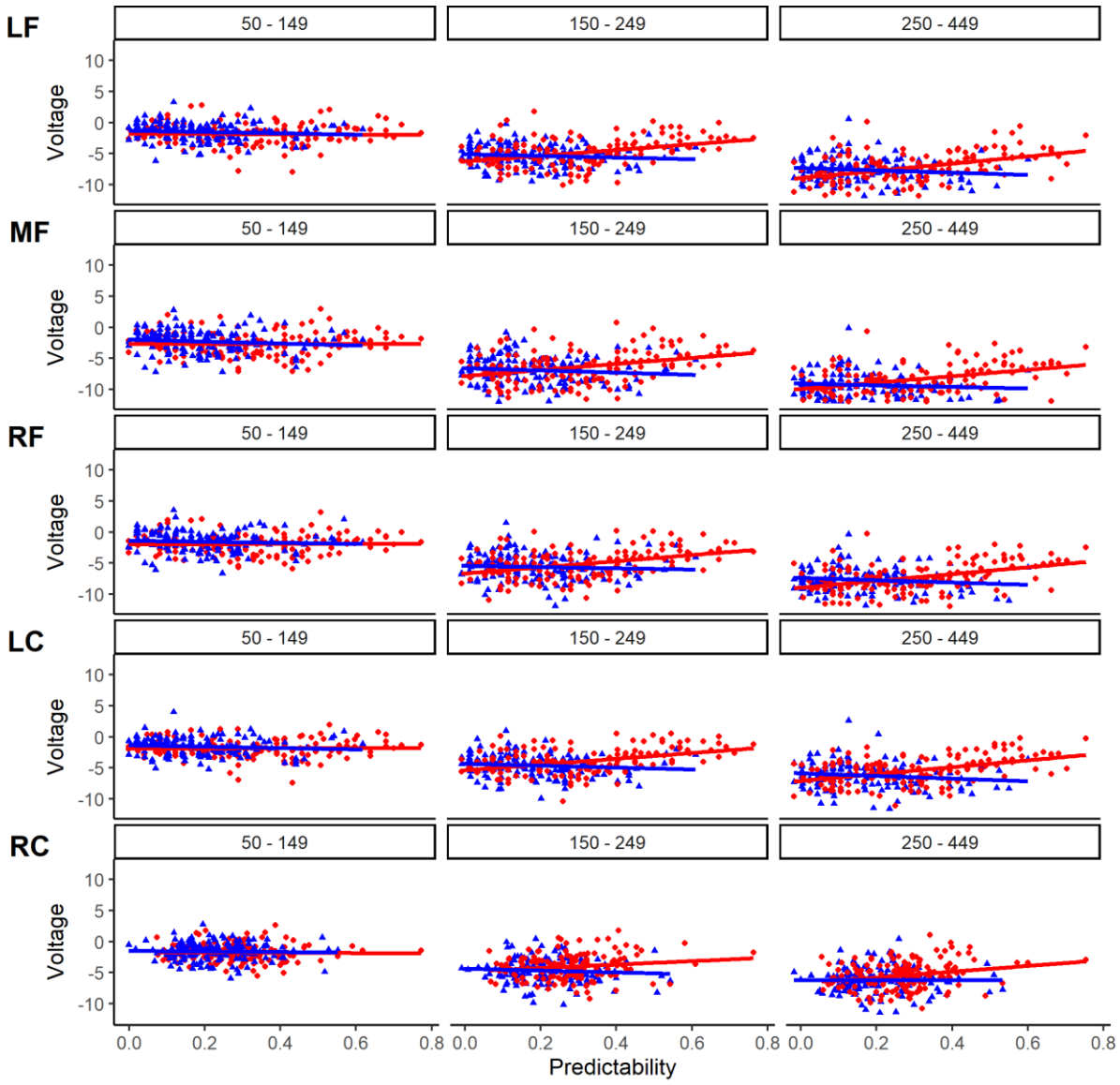


Figure 45. Exp 3: Scatterplots between ***image predictability*** and voltage at LF) left, MF) middle, and RF) right ***frontal*** as well as LC) left and RC) right ***central*** regions.

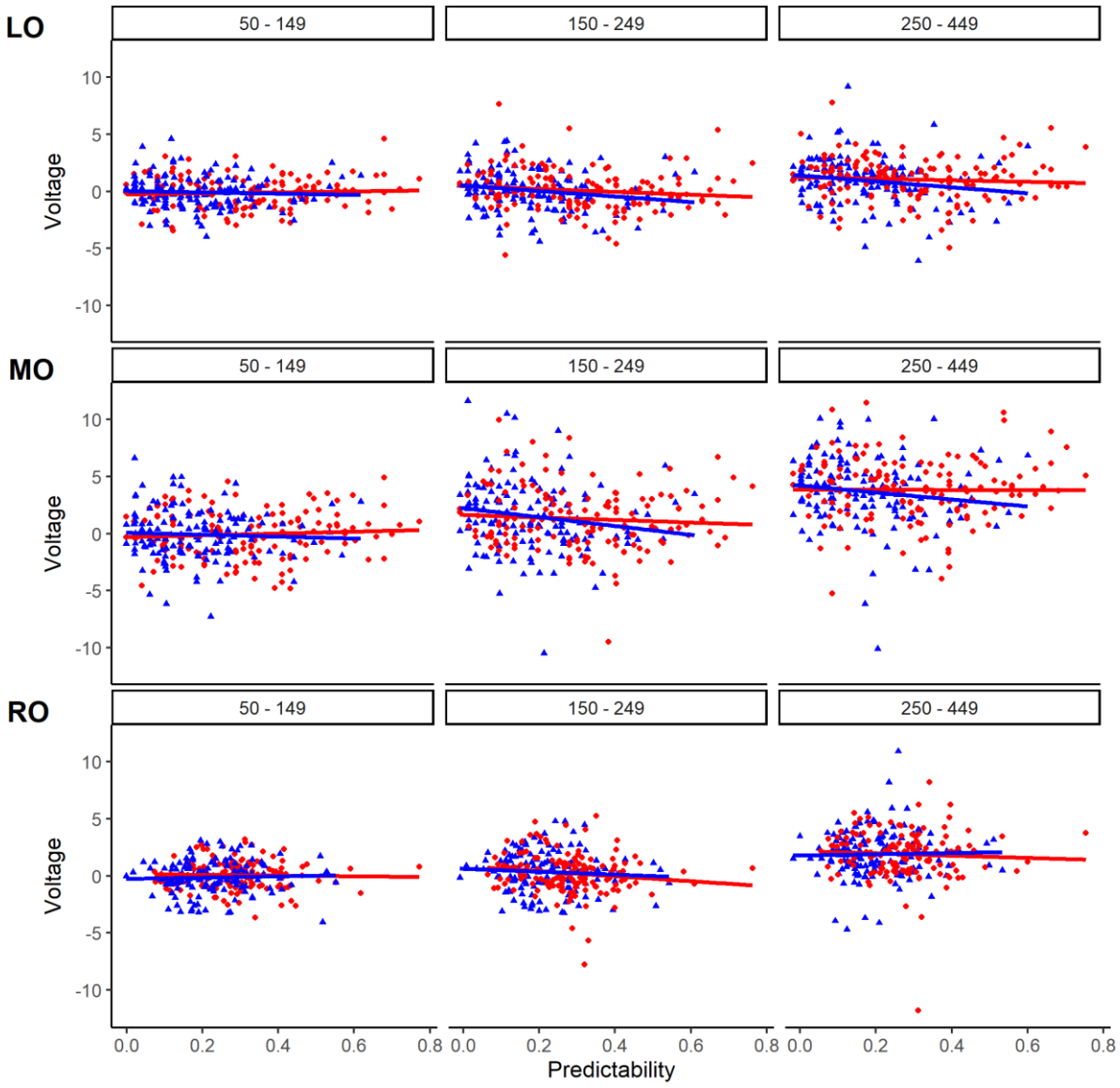


Figure 46. Exp 3: Scatterplots between ***image predictability*** and voltage at LO) left, MO) middle, and RO) right ***parietal/occipital*** regions.

Table 16. Exp 3: *Partial correlation coefficients between the log of image similarity and mean amplitude, controlling for image predictability. Partial correlation coefficients between image predictability and the mean amplitudes within each of the time windows (50-149), (150-249), and (250-449) controlling for the effect of log of image similarity*.

Region	Spatiotemporal Coherence	50-149		150-249		250-449	
		Log Sim	Pred	Log Sim	Pred	Log Sim	Pred
Left Frontal	Randomized	-0.07	-0.07	-0.08	-0.07	-0.03	0.15*
	Coherent	0.00	-0.02	0.03	0.22*	0.05	0.23*
Middle Frontal	Randomized	0.01	-0.11	-0.04	-0.10	0.04	-0.14
	Coherent	0.00	0.01	0.05	0.22*	0.11	0.23*
Right Frontal	Randomized	-0.02	-0.06	-0.05	-0.04	0.09	-0.13
	Coherent	-0.03	0.02	0.03	0.25*	0.09	0.23*
Left Central	Randomized	0.02	-0.10	-0.02	-0.10	0.02	-0.03
	Coherent	0.03	0.00	0.02	0.24*	0.04	0.24*
Right Central	Randomized	-0.01	-0.05	-0.06	-0.05	0.06	-0.08
	Coherent	-0.03	0.04	0.03	0.26*	0.09	0.28*
Left Parietal/ Occipital	Randomized	0.04	-0.06	0.03	-0.15	-0.02	0.10
	Coherent	0.12	0.02	0.03	-0.13	0.03	-0.09

Middle Parietal/ Occipital	Randomized	0.07	-0.07	-0.18*	-0.02	-0.02	-0.12
	Coherent	0.03	0.06	-0.05	-0.07	0.02	-0.04
Right Parietal/ Occipital	Randomized	0.04	0.00	-0.04	-0.10	0.05	-0.09
	Coherent	-0.05	0.07	-0.10	-0.15	-0.02	-0.11

Note: Log Sim = Log Similarity between the target image and its immediately preceding prime. Pred = Image Predictability from Experiment 1. * denotes $p < .05$.

In Experiment 2, we found that image similarity and image predictability accounted for a significant amount of variance in voltage recorded at frontal and central electrodes, and that image predictability had a stronger relationship with voltage in the coherent sequences. Statistically significant partial correlations in Experiment 2 between image predictability and voltage in frontal and central regions ranged from .23 to .44. Interestingly, we also found that image similarity, but not predictability correlated significantly with voltage recorded at parietal/occipital electrodes. Statistically significant partial correlations in Experiment 2 between similarity and voltage, controlling for image predictability, in parietal/occipital regions ranged from -0.10 to -0.23 in the middle parietal/occipital region. Partial correlations were not as strong in Experiment 3.

After controlling for the influence of image predictability, similarity did not correlate with amplitudes recorded at any of the electrodes in the frontal and central regions in Experiment 3 (See Figure 43 and Table 16). The cause of this difference is unknown, but assumedly they may have been due to either the presence of a backward noise mask immediately after the target, or due to the reduction in the duration of each scene on each trial, though the prime images had

the same SOA as in Experiment 2. Because, sensory memory has a very short duration (Sperling, 1967), processing of the low-level features in the prime immediately prior to the target may have decayed before the onset of the target, weakening low-level visual priming of the target due to visual similarity. In addition, image similarity correlated negatively with amplitudes of the P200 in the middle parietal/occipital region, r = -.18, but not in any of the other parietal/occipital regions (See Figure 44). Unlike Experiment 2, the weak interaction effect between spatiotemporal coherence and ordinal position on mean amplitude in the 150-249 ms window (Figure 42) does not seem to be uniquely attributable to visual similarity between prime and target (Table 16). Perhaps a different measure of similarity could explain more variance in the amplitudes.

As evident in Table 16 and Figure 45, predictability of the scenes accounted for unique variance in amplitudes at frontal and central regions, and the influence of predictions start in the 150-249 ms window consistent with what we found in Experiment 2. Partial correlations between image predictability and voltage, controlling for the influence of image similarity ranged from .15 in the left frontal N400 to .28 in the right central N400. Furthermore, consistent with Experiment 2, the correlations between image predictability and voltage were stronger in the coherent than randomized sequences. Thus, predictions made prior to viewing a scene facilitates scene processing above and beyond the influence of similarity between prime and target as measured from the similarity in the prime and target's spectral energy. These results are consistent with the hypotheses of SPECT that the event model facilitates both the rapid categorization scenes, and their integration into the event model during mapping. It is inconsistent, however, with purely feed-forward accounts of rapid scene categorization.

Lastly, we replicated results from Experiment 2 showing that the predictability of the images did not account for a significant amount of variance in the parietal/occipital electrodes. This result suggests that facilitation on the P200 at the 4th, 5th, and 7th positions (Figure 42) also does not seem to be attributed to the predictability of the scenes within the sequences (Table 16).

Neural Decoding of Image Categories

As in Experiment 2, we also explored the temporal dynamics of scene decoding accuracy and how emerging categorical representations contributed to scene categorization. Early facilitation accounts propose that feed-back mechanisms influence the construction of the structural description of a scene. Accordingly, we predicted that neural decoding accuracy and correlations between responses made by neural decoders and humans would be greater for scenes shown in coherent sequences before 150 milliseconds. Alternatively, matching accounts propose that feed-back mechanisms influence the matching process, which should begin after 150 milliseconds (Fabre-Thorpe et al., 2001; Thorpe et al., 1996; VanRullen & Thorpe, 2001c). Results from Experiment 2 suggest that the event model influences matching processes and perhaps early perceptual processing as well, though, the early effects were previously associated with small Bayes factors in favor of the alternative hypothesis. However, scene categorization performance was quite high in Experiment 2, possibly due to the long SOAs (i.e., 800 ms) used in Experiment 2.

Thus, we sought to reduce human categorization performance by decreasing the duration that scenes were shown and by adding a visual mask immediately after the offset of the target. By decreasing rapid scene categorization performance, we should increase off-diagonal entries in the behavioral confusion matrices. So long as confusions made by both humans and neural decoders are systematic, reducing accuracy should increase the correlation between responses

made by decoders and human observers. Thus, we should be able to gain a greater insight into the timing of the facilitation effect observed for coherent sequences by making the target harder to categorize.

However, decoding accuracy in Experiment 3 was very similar to what we observed in Experiment 2. As shown in Figure 47, the average decoding accuracy was at chance level performance before the onset of the images in all 3 regions (Occipital: 12.36%; Central: 12.40%; Frontal: 12.50%). Accuracy of neural decoders rose significantly above chance for both the coherent and randomized sequences in all three regions around 50 milliseconds, and peaked earlier for parietal/occipital regions (140 milliseconds) than central (144 milliseconds) regions than frontal (152 milliseconds) regions. The peak over parietal/occipital regions was slightly later than the peak we observed in Experiment 2 (i.e., 121 ms). As shown in Figure 47, as we found in Experiment 2 (Figure 27) differences between the coherent and randomized sequences appear to arise after the initial peak in decoding accuracy, except for the difference over parietal/occipital regions, which occurs earlier. Thus, results were consistent across experiments despite the addition of the mask.

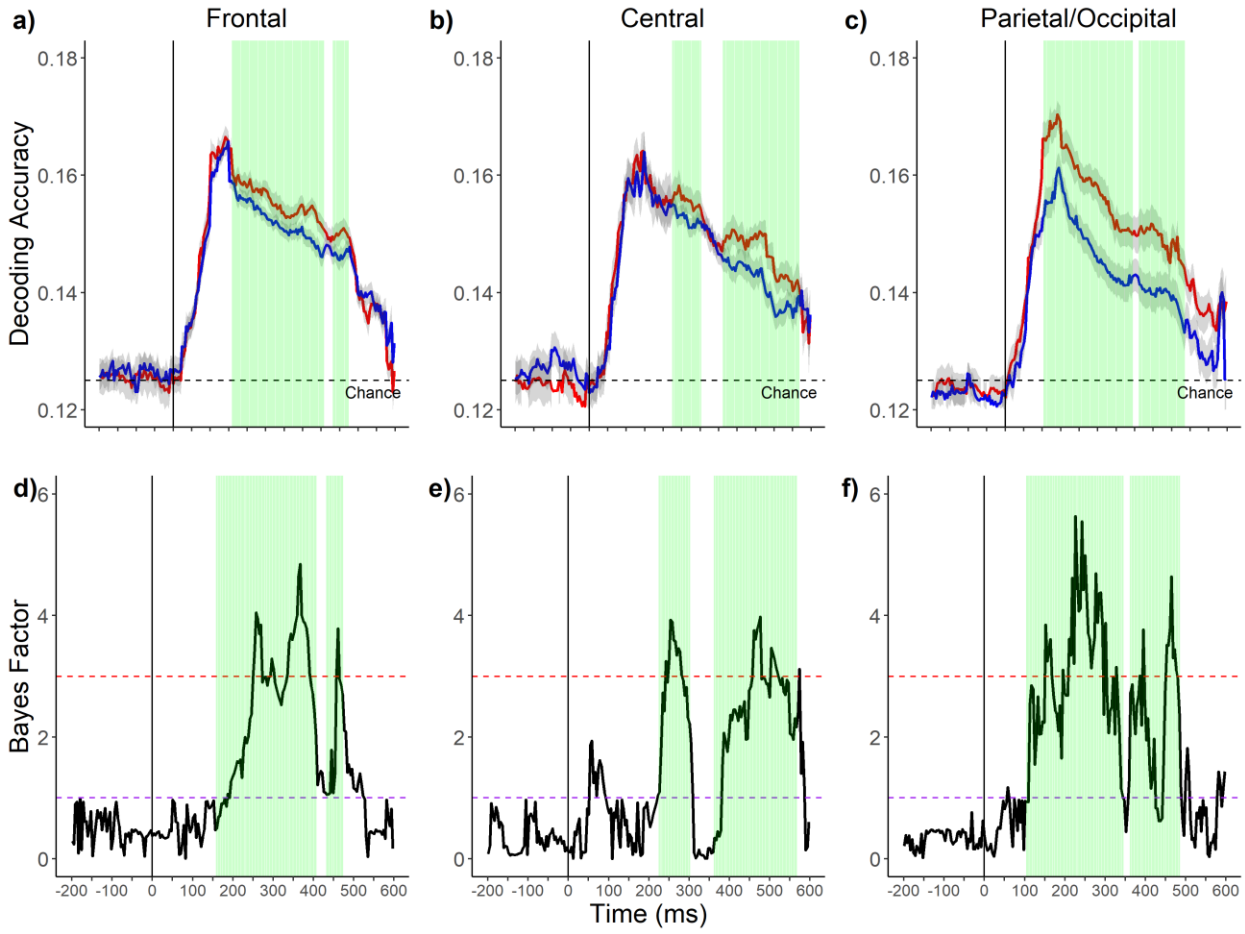


Figure 47. Exp 3: ***Decoding accuracy*** as a function of time in the epoch for the a) Frontal, b) Central, and c) Parietal/Occipital regions. Bayes Factors for each of the paired sample t tests within the epoch for d) Frontal e) Central, and f) Parietal/Occipital electrodes are provided in the bottom row. Green patches represent clusters of statistically significant comparisons. Red dashed lines in the Bayes Factors plots represent a Bayes Factor of 3 and purple lines represent a Bayes Factor of 1 and -1 respectively. Error ribbons correspond to 95% confidence intervals around the means.

To evaluate if decoding accuracy was greater, and thus scene representations were more detailed, when the sequence was coherent, we ran a linear mixed effects model on the average decoding accuracy after the onset of the target scene across participants. Models included the

fixed effects of the channel region (left, middle, and right frontal and parietal/occipital regions, and left and right central regions), the location where the photographs were taken (on-campus vs. off-campus), the effect of spatiotemporal coherence (coherent vs. randomized), and all their interactions. We specified the maximal model in the random effect structure.

Decoding accuracy for each participant and the least square means from the model are shown in Figure 48. We observed a marginally significant effect for region, $F(7,580) = 1.99$, $p = .05$, BF = 2.63. Decoding accuracy was better at parietal/occipital electrodes [Left Occipital ($M = 0.14$, $SE = .002$); Middle Occipital ($M = 0.14$, $SE = .002$); Right Occipital ($M = .15$, $SE = .002$)] than at central and frontal regions [Left Central ($M = 0.14$, $SE = .002$); Right Central ($M = 0.14$, $SE = .002$); Left Frontal ($M = 0.14$, $SE = .002$); Middle Frontal ($M = 0.14$, $SE = .002$); Right Frontal ($M = .14$, $SE = .002$)]. Like rapid scene categorization performance, we also found that decoding accuracy was better for coherent ($M = 0.16$, $SE = 0.002$) than randomized ($M = 0.14$, $SE = 0.002$) sequences, $F(1,20) = 54.50$, $p < .001$, BF = 21.41. This important effect suggests that participants represented scenes in coherent sequences more effectively than scenes in the randomized sequences, though it is important to note that decoding accuracy was still significantly above chance when scenes were shown in randomized sequences. Decoding accuracy in Experiment 3 was very similar to that for coherent sequences in Experiment 2. However, in Experiment 3, decoding accuracy decreased in the randomized sequences relative to the randomized sequences in Experiment 2. These results are consistent with the behavioral results of rapid scene categorization performance between Experiments 2 and 3. Masking the target scene did not reduce accuracy by much. Consistent with Experiment 2, there was also an interaction between spatiotemporal coherence and the location where the images were photographed, $F(1, 580) = 17.44$, $p < .001$, BF = 5.13. Decoding accuracy was significantly

better in the coherent than the randomized sequences for both locations; however, the difference was larger in the off-campus, $\beta = 0.02$, $SE = 0.002$, $t = 8.40$, $p < .001$ than the on-campus sequences, $\beta = 0.01$, $SE = 0.002$, $t = 5.32$, $p = .0001$, again showing that the facilitation of decoding accuracy was greater when the categorization task was harder for the participants. None of the remaining interactions in the analysis were statistically significant.

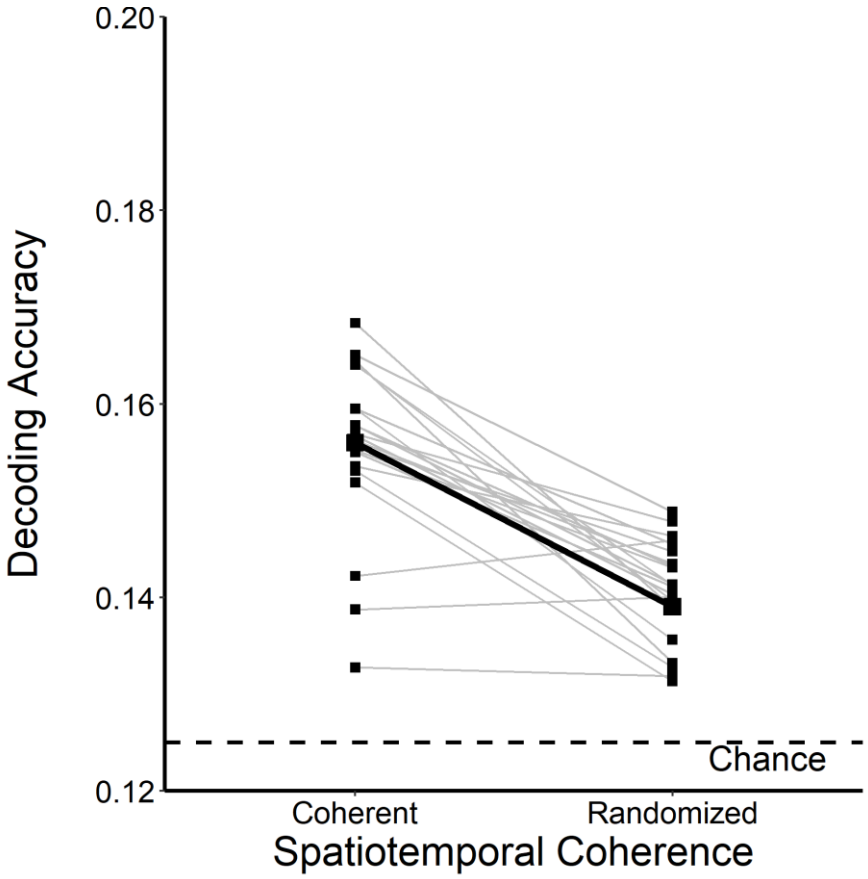


Figure 48. Exp 3: Decoding accuracy after the onset of the images as a function of the spatiotemporal coherence of the sequences. Decoding accuracy for individual participants are represented by the light gray lines, and least square means generated from the estimated regression equation are represented by the thick black line and dots. The dashed line at 12% represents chance level performance.

We also examined when decoding accuracy between coherent and randomized sequences diverged. The time point of this divergence is informative because it suggests when in the time course of scene processing the event model begins to influence rapid scene categorization. Prior to running these analyses, we averaged decoding accuracy at each time point (i.e., each ms) across the image location, and the hemisphere, within each region, since the benefit for the coherent sequences was found in both locations and in each of the 8 regions. We then conducted a paired samples t-test using decoding accuracy as the dependent measure at each time point within the epoch, and evaluated the Bayes factor associated with each statistical test.

Results are shown in Figure 49. Decoding accuracy for coherent sequences was significantly better in frontal electrodes starting from 160 and lasting until 406 milliseconds, consistent with the hypothesis that the event model facilitates matching and semantic integration processes. Decoding accuracy again became significantly better for scenes in the coherent sequences at 433 milliseconds and the effect lasted until 472 milliseconds. Bayes factors were greater than 3 at 250 milliseconds and this lasted until 281 milliseconds. Bayes factors again became greater than 3 at 335 milliseconds, which lasted until 390 milliseconds. Decoding accuracy was significantly better for coherent sequences in central regions starting 226 milliseconds, which lasted until 300 milliseconds. Accuracy again became greater at 363 milliseconds and remained significant at each time point until 566 milliseconds. Bayes factors were greater than 3 from 246 to 281 and from 460 to 488 milliseconds. Importantly, decoding accuracy was significantly better in coherent than randomized sequences starting at 105 milliseconds (35 milliseconds later than what we observed in Experiment 2) in parietal/occipital regions, and this effect lasted until 343 milliseconds though this early difference was associated with a small Bayes factor of 2.81. Decoding accuracy became significantly better for coherent

sequences again at 363 milliseconds and this effect lasted until 484 milliseconds. These later effects were supported by Bayes factors greater than 3 from 152 milliseconds to 164, and then again from 210 to 300 milliseconds. There was also an advantage for coherent sequences much later in the epoch (453-476) milliseconds. Together, these analyses suggest that the event model facilitates the perception of scene category representations as they emerge over time, and that it may start to do so as early as 100 milliseconds after the onset of a scene. This result is consistent with early facilitation accounts, since the difference emerged considerably before 150 milliseconds.

Simultaneity of visual representation and behavioral categorization

As in Experiment 2, we correlated the proportion of confusions made by the support vector machine at each time point in the epoch with confusions made behaviorally by participants. Behavioral confusion matrices averaged across all of the participants for the coherent and randomized on- and off-campus image sequences are represented in Figure 49. As evident from comparing the confusion matrices in Figure 49 a) and c) (coherent) with the matrices in b) and d) (randomized), participants made more errors in the randomized than the coherent sequences. In addition, the errors that participants made were systematic. For example, confusions among basic level categories were chiefly made within their respective superordinate indoor versus outdoor categories, rather than between them (e.g., parks were commonly misidentified as sidewalks and city centers), but rarely did an indoor category get confused with an outdoor category or vice versa (e.g., classrooms were never misidentified as parking lots).

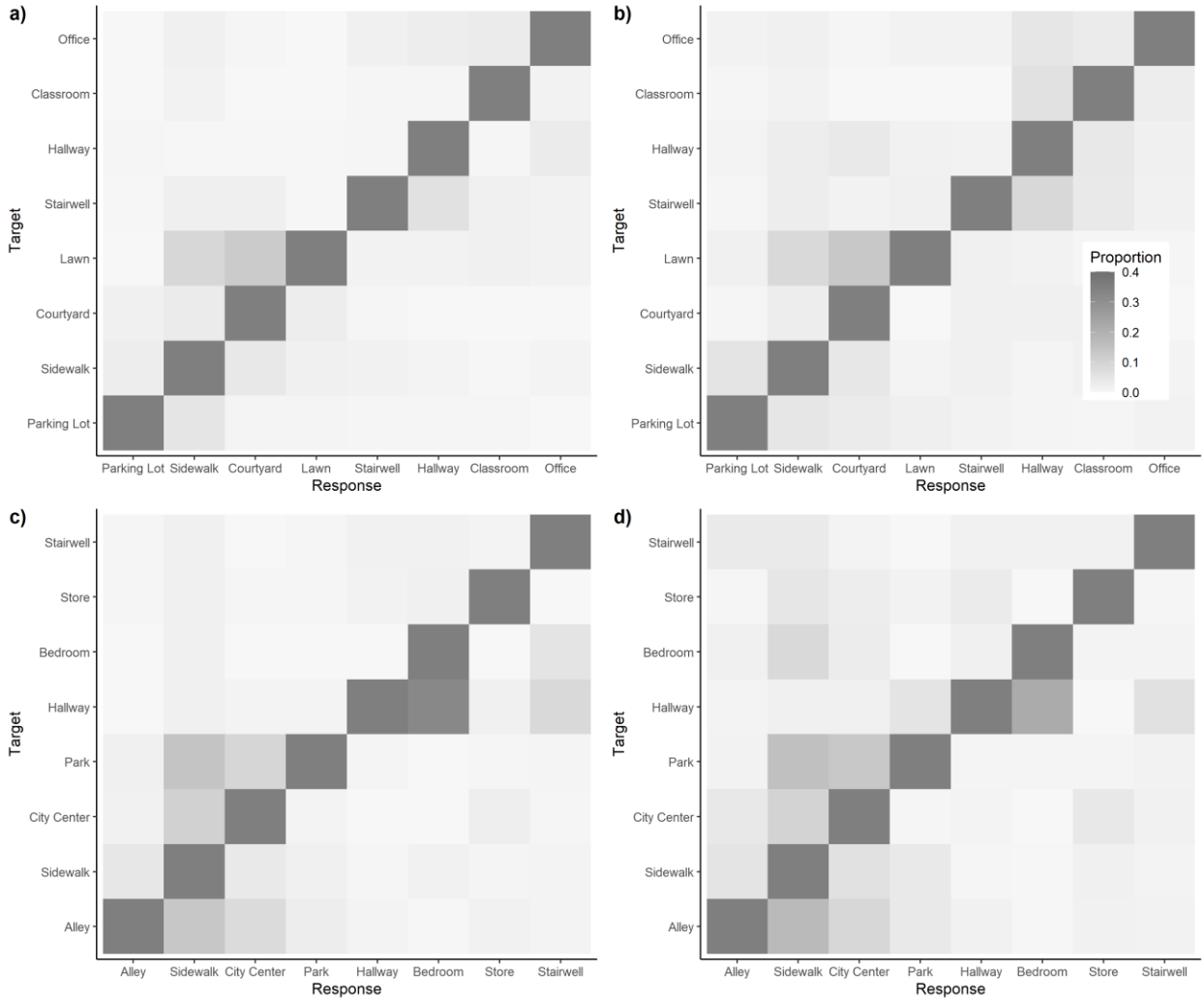


Figure 49. Exp 3: Confusion matrices for coherent and randomized image sequences for on- (top row) and off-campus (bottom row) images. Confusions in coherent sequences across participants are represented in a) and c). Confusions in randomized sequences across participants are represented in b) and d). Rows represent the target image category, and columns represent the average responses made for each response category. Thus, responses on the main diagonal are correct responses. Images belonging to indoor categories were often confused with other indoor categories, and images belonging to outdoor categories were often confused with other outdoor categories.

Figure 50 shows the variance in the neural decoder responses accounted for by behavioral confusion matrices. Correlations followed a similar pattern to what we observed with decoding accuracy. Correlations began to increase 50 milliseconds after the scene onset. They peaked earlier for parietal/occipital regions (164 milliseconds) than central (187 milliseconds) and frontal regions (203 milliseconds). In addition, as we predicted, the peak correlations were also notably higher in Experiment 3 than in Experiment 2, though the correlations were still very small. The strongest R^2 observed in Experiment 2 was .085 at frontal regions, and the strongest R^2 observed in Experiment 3 was .10 at parietal/occipital regions. This is consistent with the hypothesis that correlations should be greater when the target image is masked due to there being greater systematic variability in the errors that participants made.

As with our results of decoding accuracy, our results of the correlations between the responses made by neural decoders and human observers supported both the matching and early facilitation accounts of rapid scene categorization facilitation by spatiotemporal coherence. Correlations in frontal regions were significantly larger in the coherent condition from 187 to 402 milliseconds. Bayes factors were greater than 3 between 230 and 394 milliseconds and 441 to 464 milliseconds. In central regions, correlations were significantly greater in the coherent sequences between 253 to 281 milliseconds and again from 437 to 585 milliseconds. These significant differences were supported by Bayes factors greater than 3 from 277 to 289 milliseconds and again from 500 to 527 milliseconds. In parietal/occipital regions, correlations were significantly greater in coherent sequences between 105 to 343 milliseconds and again from 363 to 562 milliseconds consistent with decoding accuracy. Bayes Factors were greater than 3 in parietal/occipital regions from 167 to 253 milliseconds, from 367 to 429, from 449 to 480, and at 550 milliseconds. Given the time course of these correlations, this could be evidence that the

event model can also facilitate rapid scene categorization as early as 105 milliseconds, which would suggest that the event model influences early perceptual analysis of scenes shown in coherent sequences. However, while the increase in decoding accuracy due to spatiotemporal coherence was large at 100 ms (Fig. 51c), the R^2 is relatively smaller (Fig. 54c), as is the Bayes factor at 100 ms (Fig. 54f). Thus, evidence for facilitation as early as 100 ms in Experiment 3 is rather tenuous.

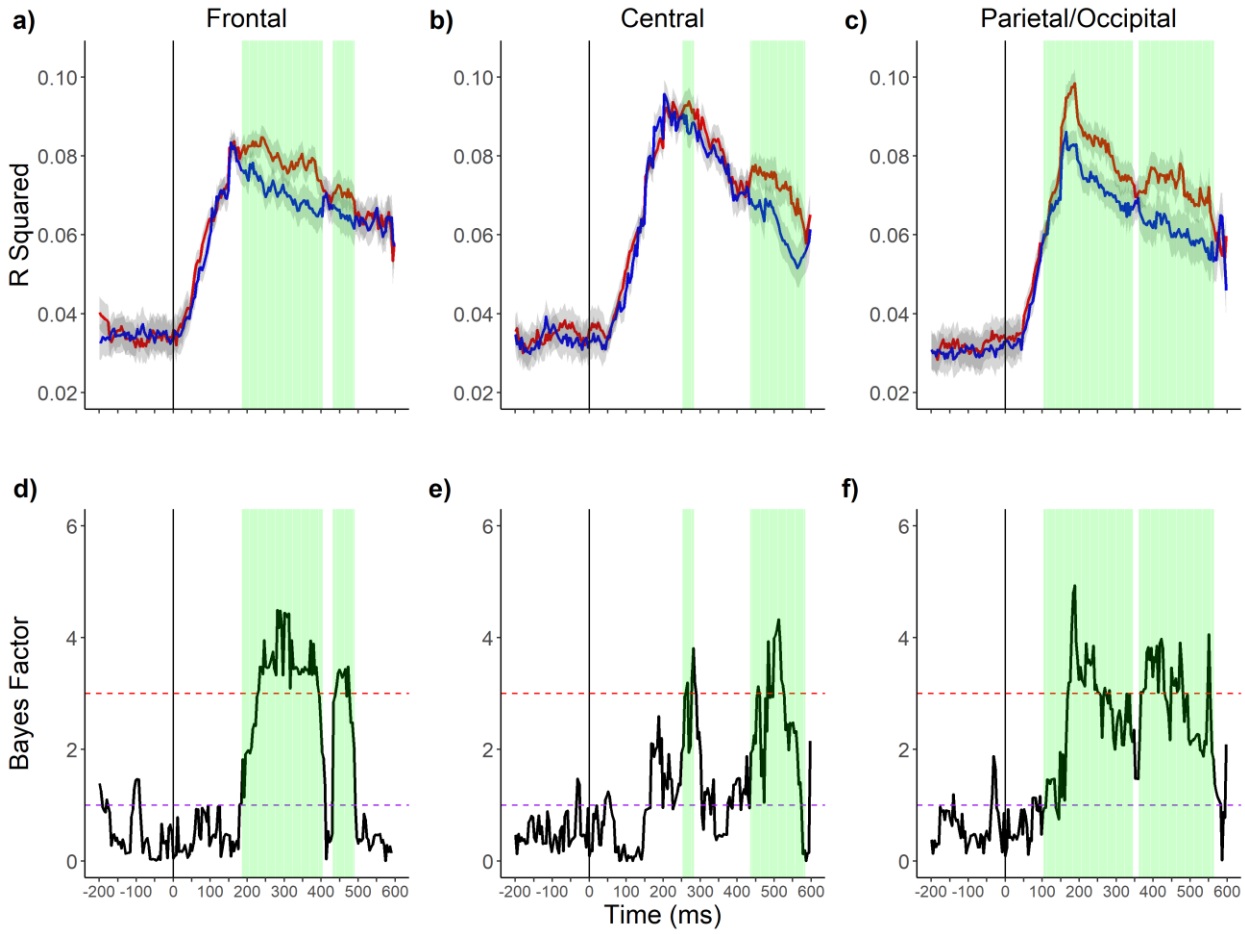


Figure 50. Exp 3: ***Unique variance*** in the behavioral confusion matrices explained by confusions made by the neural decoders over time. Error bars represent between subject 95% confidence intervals at each time point.

Discussion

The purpose of Experiment 3 was two-fold. We sought to replicate the results from Experiment 2, and we sought to obtain stronger correlations between responses made by neural decoders and humans. We were able to successfully replicate many of the findings from Experiment 2. Again, we found that rapid scene categorization performance was better when scenes were shown in spatiotemporally coherent predictable sequences than when the sequences were randomized. We also found that scene categorization performance improved as the ordinal position of the scene on each trial increased. These results are consistent with hypotheses derived from SPECT, which proposes that the extent to which facilitation will be found depends upon the degree of spatiotemporal coherence between contents of the current event model and the new incoming scene information on each eye fixation (Loschky et al., 2020).

Facilitation of vERPs

We also found evidence that the event model feeds back to influence rapid scene categorization in the vERPs. Specifically, we found that amplitudes were more positive in the coherent than randomized conditions in the 150-249 window and in the N400. These results were consistent with what we observed in Experiment 2, and they are in opposition to the notion that scene gist perception is accomplished by purely feed-forward mechanisms (Serre et al., 2007; VanRullen, 2007; VanRullen & Thorpe, 2002). Finding that waveforms differed as a function of spatiotemporal coherence between 150 and 249 milliseconds is important because prior work has found that enough visual information is processed in 150 milliseconds to activate higher-level representations (Johnson & Olshausen, 2002; Thorpe et al., 1996; VanRullen & Thorpe, 2001c). Thus, these results are consistent with matching accounts of perceptual facilitation (Bar, 2004;

Bar & Ullman, 1996; Friedman, 1979; Trapp & Bar, 2015). Matching accounts have typically proposed that top-down influences arise after the construction of a structural description of a scene, but more recent work suggests structural descriptions and matching processes may operate in parallel (Caddigan et al., 2017; Peterson, 1994; Ramkumar et al., 2016). Semantic predictions facilitate rapid scene categorization by limiting the number of alternative representations to compare with the visual input (Bar, 2004; Bar & Ullman, 1996; Friedman, 1979; Trapp & Bar, 2015). This process is much more resource intensive when observers view scenes in randomized sequences, resulting in more negative amplitudes when the scenes are less predictable or less visually similar to the scenes that preceded them.

Again, we failed to find much evidence that predictions made prior to viewing a scene modulated the amplitudes of the P200, and this lack of an effect was again associated with a small bayes factors in favor of the alternative hypothesis. Thus, we replicated the null effect observed in Experiment 2. Our results agree with claims that the P200 is not influenced by top-down factors (Hansen et al., 2018) though it may be influenced by low-level similarity between successive presentations of scenes (See Figures 23 and 43).

Neural mechanisms underlying the effect of predictions on scene processing

We also explored the potential sources for the observed difference in the waveforms between scenes shown in coherent and randomized sequences. We found that differences may have originated from four clusters of components. One of the clusters was localized in the ventral posterior cingulate cortex consistent with what we observed in Experiment 2. The second, which we did not observed in Experiment 2, was localized in Brodmann area 31 in the dorsal posterior cingulate cortex. In addition, we also failed to find differences from a cluster localized in the precuneus as we did in Experiment 2. The reason for the differences in the results between

experiments is unknown; however, Brodmann area 31 is notably situated between the ventral posterior cingulate and the precuneus. Given the lack of spatial resolution in EEG, it is possible that the activity we observed from Brodmann area 31 likely originated from the precuneus. Differences in the topography between participants and minor discrepancies in electrode placement could have also resulted in shifts in where the centroid of the dipoles in the cluster analysis were placed. In addition, the method we used to identify the potential neural source of the differences in the waveforms was a data-driven approach. A better source localization approach may be able to identify if facilitation originated from the dorsal posterior cingulate or the precuneus, or both. Regardless, we have evidence to suggest from the source localization technique that differences in the waveforms may have originated from regions that have previously been associated with the generation and maintenance of the event model (Hasson et al., 2008; Inhoff & Ranganath, 2017; Ranganath & Ritchey, 2012; Stawarczyk et al., 2019). Differences from these regions align with the hypothesis that participants generated an event model when the sequences were coherent, but not when they were randomized, and that differences in the waveforms originated from regions that are associated with the event model.

Consistent with Experiment 2, we found that a source of the difference between the raw waveforms may have originated in Brodmann area 6. Brodmann area 6 is composed of the premotor and supplementary motor area. Previous work has found that it is part of a network that is responsible for higher level control of movement in space, and it is involved in the detection of unexpected stimuli when navigating (Iaria et al., 2008). It is also involved in updating spatial information in working memory (Tanaka et al., 2005). Given that coherent sequences appeared as if an observer were navigating from one location in the environment to another and that coherent sequences were more predictable, it makes sense that differences in the waveforms

would arise from motor related areas that are involved in navigation and in the updating of spatial information.

In addition, we also found evidence that the differences in the waveforms may have originated from the fusiform gyrus in Brodmann area 37. Some work has found evidence that the fusiform gyrus is involved in categorization and semantic processing (Ardila et al., 2015). Specifically, it is involved in converting semantic information to phonological representations (Usui et al., 2003) and is therefore engaged in accessing the names and the meanings of pictures and words (McDermott et al., 2003). According to matching accounts of facilitation, predictions made prior to viewing a scene facilitate scene categorization by reducing the number of categorical representations to match the visual input; therefore, it is possible that predictions may facilitate rapid scene categorization by influencing the ability to find a name for the visual input, which could be subserved by activity in the fusiform gyrus.

Decoding of brain signals

The primary purpose of Experiment 3 was to gain more information to use to correlate responses from neural decoders with human behavioral responses. Rapid scene categorization performance was quite high in Experiment 2, but accuracy of neural decoders was not. As a consequence, the majority of behavioral responses were on the main diagonal of the confusion matrices (See Figure 29). In Experiment 3, we sought to gain more information to use to correlate responses by flashing each scene for a shorter duration and by immediately following the target scene with a perceptual mask to make the target harder to categorize (Bacon-Mace et al., 2005; Enns & Di Lollo, 2000; Loschky, Sethi, et al., 2007). We hypothesized that there would be more off-diagonal observations, and thus more systematic variance to capture, in both the human behavior and in the decoder when the scene was harder to perceive. We found that following the target

scene with a mask reduced accuracy in the randomized sequences, and to a lesser degree in the coherent sequences though the decrease in accuracy was not as much as we hypothesized. Smith and Loschky (2019) presented target and primes for 24 milliseconds followed immediately by the same perceptual masks we used here, and they found that accuracy was 54% and 34% in the coherent and randomized sequences, respectively. Therefore, we assumed before running the experiment that accuracy would decrease substantially by flashing and masking targets in sequences for 48 milliseconds. The reason why masking the target did not reduce performance by very much in Experiment 3 is unknown, though it could be due to differences in the stimuli or in the participants. Only 2/3rds of the scenes used in Experiment 3 were those used by Smith and Loschky (2019), and the new stimuli could have been easier to categorize. Further, participants completed only 48 trials in Smith and Loschky (2019), but they completed 288 trials in Experiment 3, and it could be expected that participants performed the task better over time. Lastly, sixteen of the 24 participants who took part in Experiment 3 were compensated with a monetary payment for completing the experiment; however, participants in both Experiment 2 and Smith and Loschky (2019) were compensated with course credit. Prior work has found that participants perform better on complex cognitive tasks when they are compensated with a monetary incentive (Brase, 2009; Robinson et al., 2019). Thus, participants in Experiment 3 may have performed the scene categorization task almost as well as they did in Experiment 2 because they were more motivated to do so.

This is not to say that decreasing the duration for which images were shown, and that masking the target had no effect on rapid scene categorization. Specifically, event related potentials time locked to the onset of the target in both frontal and occipital regions appeared very different from what they were in Experiment 2 (c.f., Figures 10 vs. 33, and Figures 11 vs.

34). While the mask reduced amplitudes in all of the regions, the largest effects we observed were found from activity recorded over parietal/occipital regions, as we would expect, given prior work that investigated how masking affects vERPs (Bacon-Mace et al., 2005; Robinson et al., 2019). Neural decoding accuracy over time within the epoch also showed evidence of masking the target scene. Accuracy of neural decoders peaked at similar time points, and with similar values, though the peak was later in Experiment 3 than in Experiment 2. Furthermore, in Experiment 2, decoding accuracy decreased only slightly after the initial peak. This could have been due to the long SOAs used for the target scene. Conversely, in Experiment 3, decoding accuracy more rapidly declined after the initial peak. We assume that this constant decrease in decoding accuracy after the initial peak was due to the onset of the perceptual mask 48 milliseconds after the onset of the target, since visual masking prevents the accumulation of visual information (Bacon-Mace et al., 2005; Kovacs et al., 1995; Massaro & Loftus, 1996; Rieger et al., 2005; Rolls et al., 1999).

As a consequence of failing to reduce the accuracy of the human observers and the neural decoders, we found that masking also did not substantially change the correlations we observed over time between decoders and human behavioral responses. Nevertheless, correlations were slightly better in the parietal/occipital regions in Experiment 3 (c.f., Figures 30 vs. 50). Future researchers could decrease the duration of the scenes, use a more efficient mask, or manipulate the stimulus properties so that participants make more confusions between categories in their behavioral responses. If the confusions made by decoders and human participants are systematic (Loschky et al., 2015; Walther et al., 2009; Walther & Shen, 2014), then correlations between decoder and behavior responses should improve.

Regardless, the differences we found between the coherent and randomized sequences in Experiment 3 mirrored what we found in Experiment 2. Consistent with prior work (Cichy et al., 2017; Greene & Hansen, 2020; Ramkumar et al., 2016) scene category representations could maximally be decoded from neural signals between 150 and 250 milliseconds. In addition, the maximum amount of similarity between the representations from the neural decoders and the behavioral responses were also found between 150 and 250 milliseconds, for both the coherent and randomized sequences. Together, these finding support the idea that observers begin to recognize the meaning of a scene approximately 150 milliseconds after scene onset (Johnson & Olshausen, 2002; Johnson & Olshausen, 2005; Thorpe et al., 1996; VanRullen & Thorpe, 2001c).

More importantly, we found a decrease in performance when we compared decoding accuracy between scenes presented in coherent and randomized sequences. As mentioned previously, we observed a similar drop in accuracy in rapid scene categorization performance. These results provide strong evidence to suggest there is a connection between the perceived category membership of a scene and its underlying neural representation. Furthermore, we assume that both decoding accuracy and similarity between responses made by neural decoders and human behavioral responses was better for scenes shown in coherent sequences because scenes in coherent sequences were more predictable and more visually similar than scenes shown in randomized sequences. Interestingly, however, in Experiment 3, image predictability accounted for much more variability in the amplitudes of the vERPs than did visual similarity between primes and targets.

We also found that neural decoding accuracy and similarity between behavior and decoding responses were greater in coherent sequences between 150-249 milliseconds in frontal

and central regions. Remarkably, we also found an early divergence in parietal/occipital regions as early as 105 milliseconds. Assuming that activity before 150 milliseconds corresponds to processing of stimulus features and activity after 150 milliseconds corresponds to higher-level categorical decisions (Johnson & Olshausen, 2002; VanRullen & Thorpe, 2001c),, we assume that the differences we observed before 150 milliseconds indicates facilitation very early in perceptual analysis (Aitken et al., 2020; Biederman et al., 1982; Edwards et al., 2017; Palmer, 1975b) and the differences we observed after 150 milliseconds supports the matching account of facilitation (Bar, 2004; Bar & Ullman, 1996; Mudrik et al., 2010; Smith & Federmeier, 2020; Trapp & Bar, 2015).

One may wonder why decoding accuracy before the onset of targets in coherent sequences was at chance-level performance, even though they were predictable. Afterall, we assume that participants made predictions for upcoming scenes before the target onset. There is some evidence that categorical information can be decoded from object-specific brain regions before the presentation of an object (Peelen et al., 2009), and that preparatory activity aids rapid scene categorization (Puri et al., 2009). If viewers generated an event model for the sequences and made predictions for upcoming scene categories, then one could hypothesize that there would be more category specific information in the neural signals before the onset of the scenes in the coherent sequences. The lack of a difference we observed may be due to the baseline correction we applied to the vERPs prior to submitting voltage to the support vector machine. In the future, researchers should investigate preparatory activity prior to the onset of target scenes shown in coherent and randomized sequences.

Chapter 5 - General Discussion

The roles that feed-forward and feed-back processing play in recognition remain critical to theories of cognition and perception; however, the unique contribution of each remains largely unknown and heavily disputed (Bar & Ullman, 1996; Firestone & Scholl, 2016; Hollingworth & Henderson, 1998, 1999; Kveraga et al., 2007; Summerfield & De Lange, 2014). Our results suggest that cognitive factors such as predictions made prior to viewing a scene change how scenes are perceived. This is in opposition to purely feed-forward models of rapid scene categorization, which argue that predictions made prior to viewing a scene should not feedback and influence initial scene processing (Fabre-Thorpe et al., 2001; Serre et al., 2007; VanRullen, 2007; VanRullen & Thorpe, 2002).

To test our hypotheses, we presented the same photographs of real-world scenes in either spatiotemporally coherent predictable, visually similar, sequences or in randomized sequences that were locally less visually similar, and both locally and globally less predictable. Consistent with our prior psychophysical research using the same or similar stimuli (Smith & Loschky, 2019), we found that both the ability to predict scenes in Experiment 1, and rapid scene categorization performance in Experiments 2 and 3, were better when scenes were viewed in coherent sequences. We also found that rapid scene categorization performance improved as viewers' putative event model was created and updated across successive views of multiple scenes. These results were consistent with hypotheses generated from the Scene Perception & Event Comprehension Theory (SPECT), which proposes that rapid scene categorization should be improved to the extent that scenes cohere with one's current event model (Loschky et al., 2020). Thus, our results challenge models of rapid scene categorization that assume only feed-forward mechanisms are involved, and that top-down influences are either absent, negligible, or

operate only at post-perceptual levels of processing (Ganis & Kutas, 2003; Hollingworth & Henderson, 1998; Potter et al., 2014; VanRullen & Thorpe, 2002). However, we do not argue that our results necessarily reject the primacy of feed-forward processing of features during rapid scene categorization. The standard method of researching human rapid scene categorization over the past 50 years has been to present scenes briefly (e.g., for 10-100 ms), in random order, and all such studies have shown remarkably good performance (Biederman, 1972; Greene & Oliva, 2009a; Loschky et al., 2010; Loschky, Sethi, et al., 2007; Potter, 1976; Rousselet et al., 2005; Schyns & Oliva, 1994). That is very similar to our randomized condition, which has near chance predictability, and relatively low pair-wise visual similarity. Thus, it is clear that viewers can rapidly categorize scenes at high levels of accuracy when they have little if any ability to predict what they will be. As such, feature extraction mechanisms are fully capable of distinguishing information from complex natural scenes. However, our study shows that these mechanisms are susceptible to higher-level influences, which feed back to influence processing during the feed-forward sweep. Furthermore, these feed-back processes had a greater impact when stimulus information was more limited. This was shown by the fact that the Bayes factor for spatiotemporal coherence was considerably larger in Experiment 3 than Experiment 2, which differed only in Experiment 3 having shorter stimulus durations, and visual masking of the targets. Moreover, the processes we have investigated by presenting scenes in spatiotemporally coherent sequences better reflects real-life day-to-day rapid scene categorization, in which we constantly generate (likely unconscious) predictions for the scene categories that we will experience next.

To examine when in the time course of scene processing the event model begins to facilitate rapid scene categorization, we recorded participant's vERP signals while they very

briefly viewed each of the scenes. We can shed light on how predictions facilitate rapid scene understanding by knowing when predictions start to influence scene processing. Studies of scene-specific processing suggest that cortical activity is primarily driven by basic stimulus features during the first 150 milliseconds of scene processing, and it is sensitive to scene categorical and conceptual information after 150 milliseconds (Greene & Hansen, 2020; Johnson & Olshausen, 2002; VanRullen & Thorpe, 2001c).

Early facilitation accounts (Biederman et al., 1982; Palmer, 1975b), predicted that we would find differences in vERPs before 150 milliseconds; however, we did not find evidence to support this hypothesis by analyzing vERPs alone. Instead, we found evidence of facilitation on scene perception in later components (150-249 ms [P200] and in the N400). Differences in these later components align with some studies that have investigated how scene recognition influences object (Lauer et al., 2018; Mudrik et al., 2010; Truman & Mudrik, 2018; Võ & Wolfe, 2013) and scene processing (McLean et al., 2021; Sitnikova et al., 2008) though not all of the previous studies have found differences in each of these time windows (Demiral et al., 2012; Ganis & Kutas, 2003). Differences in these P200 and N400 components have previously been interpreted as strengthening perceptual to memory matching processing (Mudrik et al., 2010; Schendan, 2019; Smith & Federmeier, 2020; Truman & Mudrik, 2018), and in reducing the difficulty of integrating current semantic information with previously viewed information (Hagoort et al., 2009; Van Berkum et al., 1999). Early visual responses are involved in transforming the perceptual input into a structural representation. Procedures, starting around 150 milliseconds, (Thorpe et al., 1996; VanRullen & Thorpe, 2001c), match this structural description to representations stored in semantic memory, resulting in identification. Our results clearly show that waveforms to scenes shown in coherent sequences begin to diverge from those

shown in randomized sequences around 150 milliseconds. This divergence lasts through the N400 components. Thus, presenting scenes in coherent sequences influences matching and the assumedly post-perceptual semantic integration processes.

We did however find some evidence in favor of early perceptual facilitation accounts when we compared decoding accuracy between scenes shown in coherent and randomized sequences and when we compared similarity in the representations between neural decoders and behavioral responses between the conditions. We found evidence of early facilitation at 70 milliseconds in Experiment 2 and later around 100 milliseconds in Experiment 3. We assume that these effects reflect facilitation of constructing the structural description of the scene, either by predictions from the current event model, or alternatively, by the overlap of visual features activated in response to successive spatiotemporally coherent scenes. One caveat is that both of the above-noted results at 70 ms and 100 ms post-stimulus had relatively low Bayes Factors in favor of the alternative hypothesis. Thus, these results would be strengthened by further replication.

As noted above, a key question is *how* presenting scenes in coherent sequences facilitates scene processing. We proposed and found some support for at least two mechanisms. Namely, image similarity, as measured by the shared spectral information between successive scenes, facilitated vERP amplitudes in the 50-149 millisecond window as well as amplitudes in the N400 window. However, those results were much clearer in Experiment 2 than Experiment 3. We propose that image similarity facilitates scene perception processes occurring in the feed-forward sweep, by activating the same or similar feature detectors from one scene to the next, producing visual priming (Bar & Biederman, 1998; Shafer-Skelton & Brady, 2019; Sperber et al., 1979). According to this idea, facilitation is greater in coherent sequences, in part, because successive

scenes in coherent sequences share more visual features than successive scenes in randomized sequences.

We also found that image predictability, as quantified by participants in Experiment 1, correlated with amplitudes, in the same time windows, above and beyond the influence of image similarity. We hypothesize, based on SPECT, that predictions are generated from viewer's event model for upcoming scenes. Predictions could originate from regions in the prefrontal cortex, or from regions associated with processing in the event model. Our source localization data support the later claim. Our results may also be consistent with Predictive Coding accounts of visual processing.

According to Predictive Coding Theory (Clark, 2013; Friston, 2018; Friston & Kiebel, 2009; Rao & Ballard, 1999; Summerfield & De Lange, 2014), perceptual processes are supported by feed-back signals that attempt to match feed-forward signals. Backward connections do the "heavy lifting" by carrying predictions for lower-level inputs. Predictive Coding Theory follows Helmholtz's (1860) notion that perception is a process of unconscious hypothesis-driven inference, whereby recognition is cyclical. Hypotheses for upcoming perceptual input are initially generated by higher-level regions. Higher-level regions send feedback signals (priors in a Bayesian framework) to try and "explain away" the driving sensory signal (likelihoods) from the stage below. No further action is needed when regions at higher levels in the visual hierarchy successfully predict lower-level activity; however, prediction errors arise when predictions do not align with incoming information. Importantly, Predictive Coding Theory proposes that feed-forward connections carry the difference (e.g., the prediction errors) between predictions and the observed inputs from lower levels. Prediction error signals play an important role in determining the sensory input since errors update predictions at higher levels.

As such, the mind does not build the event model from the feed-forward low-level information alone. Instead, each level in the visual hierarchy attempts to predict the activity at the level immediately below it, and sends those predictions via feedback connections. The identity of the observed input is determined from the hypothesis that generates the most accurate predictions.

The question also remains open as to precisely when this kind of predictive information interacts with incoming, moment-to-moment operations of the perceptual system. Currently, Predictive Coding Theory proposes that prediction error monitoring is a concomitant aspect of information processing, accounting for processing that occurs between any pair of levels that communicate with each other using feed-forward and feed-back signals. There is evidence to suggest that predictions can facilitate visual processing as early as V1 and even the lateral geniculate nucleus (LGN) (Aitken et al., 2020; Rao & Ballard, 1999). Predictive Coding theory has been proposed to account for repetition suppression (Summerfield et al., 2008) and binocular rivalry (Hohwy et al., 2008). Importantly, it has also been proposed to explain object-scene consistency effects (Kveraga et al., 2007). Thus, Predictive Coding Theory may account for *how* a prediction made prior to viewing a scene can facilitate rapid scene categorization.

It is also possible that predictions facilitated scene gist processing indirectly by influencing *how* participants attended to the scenes shown in coherent compared to randomized sequences (Peelen & Kastner, 2011; Seidl et al., 2012). Attention serves as a gateway to perception, such that visual processing of information at an attended location is facilitated (Posner, 1980). Further, observers often fail to perceive what they do not attend to, as in the case of inattentional and change blindness (Rensink et al., 1997; Simons & Chabris, 1999). Attention enhances processing of visual features, making objects appear clearer or more detailed (Carrasco et al., 2004). In the context of the current experiment, predictions for an upcoming scene (e.g., a

hallway), when the sequence was coherent, may have influenced observer's attentional set, whereby they searched for diagnostic features in the scenes consistent with their predictions (e.g., indoor scene features and rectilinearity) (Schyns, 1998). This proposal is consistent with modular, information encapsulated, views of perception (Firestone & Scholl, 2016), whereby predictions influence the input to the feed-forward sweep, rather than acting directly on the visual input. Future research could evaluate if predictions influence scene perception directly or indirectly through mechanisms involved in shifting attention.

There is also an open question about the nature of these predictions. Despite many decades of research into the features that underly natural scene perception and distinguish scene categories (Greene & Oliva, 2009b; Loschky, Sethi, et al., 2007; Oliva & Torralba, 2001; Renninger & Malik, 2004; Walther & Shen, 2014), we still know very little about the critical features that we use to make such categorizations. Given that neural decoding accuracy was greater in parietal/occipital regions as early as 70 milliseconds post-stimulus, it is possible that predictions of the upcoming scene category may likely contain information, not just about the scene's identity, but also its conceptual features and associated structural description. At one level, one may expect to see a 'hallway', or, at a broader level, a navigable scene (Greene et al., 2016) that is an enclosed space (Greene & Oliva, 2009). On another level, this may be an expectation to see a scene with a particular spatial layout (Oliva & Torralba, 2006; Sanocki, 2003; Schyns & Oliva, 1994). A deeper unconscious expectation may be that the layout is anchored to the perceptual upright (Gregory & McCloskey, 2010; Haji-Khamneh & Harris, 2010; Harris et al., 2011) via the gravitational frame (Loschky et al., 2015), such that an upside down or sideways scene will be harder to recognize (Kelley et al., 2003; Loschky et al., 2015; Walther et al., 2009). Likewise, expectations for a particular scene may extend to a specific

spatial envelope, composed of particular combinations of spatial frequencies and orientations (Oliva & Torralba, 2001, 2006), and one may expect certain colors (Oliva & Schyns, 2000), textures (Renninger & Malik, 2004), and contour-based information (Choo & Walther, 2016; Walther & Shen, 2014). This is an ongoing area of study, so alternative interpretations and explanations are possible. What seems clear from the current study's data, is that viewers generated predictions for upcoming scene categories, and predictions fed back to influence how the scenes were understood, both before and particularly at 150 milliseconds post-scene onset.

The current set of experiments opens several important lines for future investigation. The methodology we used involved presenting spatiotemporally coherent sequences of scenes, each photographed from a first-person viewpoint, while navigating from a starting location to a destination, attempted to better reflect scene processing outside of the lab. But there are limits to how immersive a series of static images can be. To take this further, scenes that served as primes could be replaced by video clips of journeys in the environment. Taken further, one could incorporate VR technology and have people navigate along predetermined paths, and evaluate how predictions influence scene perception. As mentioned earlier, an important question remains about the spatiotemporal dynamics of prediction effects. Other methodologies may include fMRI, MEG, or combined EEG and fMRI (Philiastides et al., 2021). These methodologies may offer further insights into how predictions made prior to viewing a scene facilitate scene gist perception by illuminating their underlying brain networks, their connections, and their time courses their activation.

Future research could also evaluate how familiarity with the routes where photographs of scenes are taken could have on facilitating scene gist processing. It is possible that our results would not generalize to a sample of participants taken from a different University or town. We

consistently found differences between on- and off-campus sequences, such that off-campus sequences were harder to categorize and harder to decode from neural signals. It is possible that participants were more familiar with the on- than the off-campus scenes, since they were recruited from on-campus. Familiarity with the scenes could have resulted in more accurate categorization performance of the on-campus sequences (see however Fabre-Thorpe et al., 2001). As such, one could compare the role of the event model on facilitating scene gist processing when sequences are familiar and unfamiliar to participants.

Questions surrounding the influence of sequential predictions on scene gist perception are not only of theoretical interest, but also have real world applications. Specifically, this research could be applied to artificial vision systems that can navigate their environment by recognizing scenes. For example, caretaker robots for the elderly need to be able to distinguish hallways from offices or stairwells. Self-driving automobiles must recognize the differences between driveways, parking lots, and highways to appropriately drive through them. Importantly, predictions made prior to viewing a scene affects scene perception in human observers. Artificial vision systems may likewise benefit from them.

References

Acar, Z. A., & Makeig, S. (2013). Effects of forward model errors on EEG source localization. *Brain Topography*, *26*(3), 378-396.

Aguirre, G. K., & D'Esposito, M. (1999). Topographical disorientation: A synthesis and taxonomy. *Brain*, *122*(9), 1613-1628. https://doi.org/10.1093/brain/122.9.1613

Aitken, F., Menelaou, G., Warrington, O., Koolschijn, R. S., Corbin, N., Callaghan, M. F., & Kok, P. (2020). Prior expectations evoke stimulus-specific activity in the deep layers of the primary visual cortex. *PLoS biology*, *18*(12), e3001023.

Aminoff, E. M., Kveraga, K., & Bar, M. (2013). The role of the parahippocampal cortex in cognition. *Trends in Cognitive Sciences*, *17*(8), 379-390.

Angelucci, A., Levitt, J. B., Walton, E. J., Hupe, J.-M., Bullier, J., & Lund, J. S. (2002). Circuits for local and global signal integration in primary visual cortex. *Journal of Neuroscience*, *22*(19), 8633-8646.

Anllo-Vento, L., Luck, S. J., & Hillyard, S. A. (1998). Spatio-temporal dynamics of attention to color: Evidence from human electrophysiology. *Human Brain Mapping*, *6*(4), 216-238.

Ardila, A., Bernal, B., & Rosselli, M. (2015). Language and visual perception associations: meta-analytic connectivity modeling of Brodmann area 37. *Behavioural neurology*.

Bach, M. (2006). The Freiburg Visual Acuity Test-Variability unchanged by post-hoc re-analysis [journal article]. *Graefe's Archive for Clinical and Experimental Ophthalmology*, *245*(7), 965-971. https://doi.org/10.1007/s00417-006-0474-4

Bacon-Mace, N., Mace, M. J., Fabre-Thorpe, M., & Thorpe, S. J. (2005). The time course of visual processing: Backward masking and natural scene categorisation. *Vision Research*, *45*, 1459-1469.

Bankson, B. B., Hebart, M. N., Groen, I. I., & Baker, C. I. (2018). The temporal evolution of conceptual object representations revealed through models of behavior, semantics and deep neural networks. *Neuroimage*, *178*, 172-182.

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, *5*(8), 617.

Bar, M., & Biederman, I. (1998). Subliminal Visual Priming. *Psychological Science*, *9*(6), 464-468. https://doi.org/10.1111/1467-9280.00086

Bar, M., & Ullman, S. (1996). Spatial context in recognition. *Perception*, *25*(3), 343-352.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255-278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Bello, N. M., Kramer, M., Tempelman, R. J., Stroup, W. W., St-Pierre, N. R., Craig, B. A., Young, L. J., & Gbur, E. E. (2016). On recognizing the proper experimental unit in animal studies in the dairy sciences. *Journal of dairy science*, *99*(11), 8871-8879.

Bello, N. M., & Renter, D. G. (2018). Invited review: Reproducible research from noisy data: Revisiting key statistical principles for the animal sciences. *Journal of dairy science*, *101*(7), 5679-5701.

Biederman, I. (1972). Perceiving real-world scenes. *Science*, *177*(4043), 77-80. <Go to ISI>://A1972M852500026

Biederman, I., Mezzanotte, R., & Rabinowitz, J. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, *14*, 143-177.

Biederman, I., Rabinowitz, J., Glass, A., & Stacy, E. (1974). On the information extracted from a glance at a scene. *Journal of Experimental Psychology, 103*, 597-600.
Bilalić, M., Lindig, T., & Turella, L. (2019). Parsing rooms: the role of the PPA and RSC in perceiving object relations and spatial layout. *Brain Structure and Function, 224*(7), 2505-2524.
Block, N. (2008). The Perception/Cognition Border. Proceedings of the Aristotelian Society,
Blonder, L. X., Smith, C. D., Davis, C. E., Kesler, M. L., Garrity, T. F., Avison, M. J., & Andersen, A. H. (2004). Regional brain response to faces of humans and dogs. *Cognitive Brain Research, 20*(3), 384-394.
Boehler, C., Schoenfeld, M., Heinze, H.-J., & Hopf, J.-M. (2008). Rapid recurrent processing gates awareness in primary visual cortex. *Proceedings of the National Academy of Sciences, 105*(25), 8742-8747.
Boyce, S., & Pollatsek, A. (1992). An exploration of the effects of scene context on object identification. In K. Rayner (Ed.), *Eye movements and visual cognition* (pp. 227-242). Springer-Verlag.
Brase, G. L. (2009). How different types of participant payments alter task performance. *Judgment and Decision Making, 4*(5), 419.
Brewer, W. F., & Treyens, J. C. (1981). Role of schemata in memory for places. *Cognitive Psychology, 13*(2), 207-230.
Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: anatomy, function, and relevance to disease.
Bullier, J. (2001). Integrated model of visual processing. *Brain Research Reviews, 36*(2-3), 96-107.
Burge, T. (2014). Reply to Block: Adaptation and the upper border of perception. *Philosophy and phenomenological research, 89*(3), 573-583.
Caddigan, E., Choo, H., Fei-Fei, L., & Beck, D. M. (2017). Categorization influences detection: A perceptual advantage for representative exemplars of natural scene categories. *Journal of Vision, 17*(1), 21-21.
Camprodon, J. A., Zohary, E., Brodbeck, V., & Pascual-Leone, A. (2010). Two phases of V1 activity for visual recognition of natural images. *Journal of Cognitive Neuroscience, 22*(6), 1262-1269.
Carrasco, M., Ling, S., & Read, S. (2004). Attention alters appearance [10.1038/nn1194]. *Nature Neuroscience, 7*(3), 308-313. http://dx.doi.org/10.1038/nn1194
Castelhano, M. S., & Pollatsek, A. (2010). Extrapolating spatial layout in scene representations. *Memory & Cognition, 38*(8), 1018-1025.
Cavanna, A. E., & Trimble, M. R. (2006). The precuneus: a review of its functional anatomy and behavioural correlates. *Brain, 129*(3), 564-583.
Cermeño-Aínsa, S. (2021). Is Perception Stimulus-Dependent? *Review of Philosophy and Psychology*, 1-20.
Cho, J., & Sharp, P. E. (2001). Head direction, place, and movement correlates for cells in the rat retrosplenial cortex. *Behavioral neuroscience, 115*(1), 3.
Choo, H., & Walther, D. B. (2016). Contour junctions underlie neural representations of scene categories in high-level human visual cortex. *Neuroimage, 135*, 32-44.
Cichy, R. M., Khosla, A., Pantazis, D., & Oliva, A. (2017). Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *Neuroimage, 153*, 346-358.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence [Article]. *Scientific Reports*, *6*, 27755. https://doi.org/10.1038/srep27755

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181-204.

Clark, V. P., Fan, S., & Hillyard, S. A. (1994). Identification of early visual evoked potential generators by retinotopic and topographic analyses. *Human Brain Mapping*, *2*(3), 170-187.

Coderre, E. L., O'Donnell, E., O'Rourke, E., & Cohn, N. (2020). Predictability modulates neurocognitive semantic processing of non-verbal narratives. *Scientific Reports*, *10*(1), 1-11.

Cohn, N., & Foulsham, T. (2020). Zooming in on the cognitive neuroscience of visual narrative. *Brain And Cognition*, *146*, 105634.

Cohn, N., & Kutas, M. (2015). Getting a cue before getting a clue: Event-related potentials to inference in visual narrative comprehension. *Neuropsychologia*, *77*, 267-278.

Cohn, N., Paczynski, M., Jackendoff, R., Holcomb, P. J., & Kuperberg, G. R. (2012). (Pea)nuts and bolts of visual narrative: Structure and meaning in sequential image comprehension. *Cognitive Psychology*, *65*(1), 1-38. https://doi.org/http://dx.doi.org/10.1016/j.cogpsych.2012.01.003

Cooper, B. G., & Mizumori, S. J. (1999). Retrosplenial cortex inactivation selectively impairs navigation in darkness. *Neuroreport*, *10*(3), 625-630.

Daselaar, S. M., Rice, H. J., Greenberg, D. L., Cabeza, R., LaBar, K. S., & Rubin, D. C. (2008). The spatiotemporal dynamics of autobiographical memory: neural correlates of recall, emotional intensity, and reliving. *Cerebral Cortex*, *18*(1), 217-229.

Davenport, J. L. (2007). Consistency effects between objects in scenes. *Memory & Cognition*, *35*(3), 393-401.

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science*, *15*(8), 559-564. https://doi.org/https://doi.org/10.1111/j.0956-7976.2004.00719.x

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117-1121.

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9-21.

Demiral, Ş. B., Malcolm, G. L., & Henderson, J. M. (2012). ERP correlates of spatially incongruent object identification during scene viewing: Contextual expectancy versus simultaneous processing. *Neuropsychologia*, *50*(7), 1271-1285.

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, *72*, 78-89.

Dilks, D. D., Julian, J. B., Paunov, A. M., & Kanwisher, N. (2013). The Occipital Place Area Is Causally and Selectively Involved in Scene Perception. *The Journal of Neuroscience*, *33*(4), 1331-1336. https://doi.org/10.1523/jneurosci.4081-12.2013

Draschkow, D., Heikel, E., Võ, M. L.-H., Fiebach, C. J., & Sassenhagen, J. (2018). No evidence from MVPA for different processes underlying the N300 and N400 incongruity effects in object-scene processing. *Neuropsychologia*, *120*, 9-17.
Edwards, G., Vetter, P., McGruer, F., Petro, L. S., & Muckli, L. (2017). Predictive feedback to V1 dynamically updates with sensory input. *Scientific Reports*, *7*(1), 1-12.
Enns, J. T., & Di Lollo, V. (2000). What's new in visual masking? *Trends in Cognitive Sciences*, *4*(9), 345-352.
Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598-601.
Epstein, R. A. (2005). The cortical basis of visual scene processing. *Visual Cognition*, *12*(6), 954.
Epstein, R. A., & Baker, C. I. (2019). Scene perception in the human brain.
Evans, K. K., Horowitz, T. S., & Wolfe, J. M. (2011). When Categories Collide: Accumulation of Information About Multiple Categories in Rapid Scene Perception. *Psychological Science*, *22*(6), 739-746. https://doi.org/10.1177/0956797611407930
Fabre-Thorpe, M., Delorme, A., Marlot, C., & Thorpe, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *Journal of Cognitive Neuroscience*, *13*(2), 171-180.
Federmeier, K. D., & Kutas, M. (2001). Meaning and modality: Influences of context, semantic memory organization, and perceptual predictability on picture processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(1), 202.
Ferstl, E. C., Neumann, J., Bogler, C., & Von Cramon, D. Y. (2008). The extended language network: a meta-analysis of neuroimaging studies on text comprehension. *Human Brain Mapping*, *29*(5), 581-593.
Filik, R., & Leuthold, H. (2008). Processing local pragmatic anomalies in fictional contexts: Evidence from the N400. *Psychophysiology*, *45*(4), 554-558.
Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and Brain Sciences*, *39*.
Foxe, J. J., & Simpson, G. V. (2002). Flow of activation from V1 to frontal cortex in humans. *Experimental Brain Research*, *142*(1), 139-150.
Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, *291*(5502), 312-316.
Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *Journal of Neuroscience*, *23*(12), 5235-5246.
Friedman, A. (1979). Framing pictures: the role of knowledge in automatized encoding and memory for gist. *Journal of experimental psychology. General*, *108*(3), 316-355.
Friston, K. (2018). Does predictive coding have a future? *Nature Neuroscience*, *21*(8), 1019-1021.
Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1211-1221.
Ganis, G., & Kutas, M. (2003). An electrophysiological study of scene effects on object identification. *Cognitive Brain Research*, *16*(2), 123-144.
Ganis, G., Kutas, M., & Sereno, M. I. (1996). The search for "common sense": An electrophysiological study of the comprehension of words and pictures in reading. *Journal of Cognitive Neuroscience*, *8*(2), 89-106.

Germeys, F., & d'Ydewalle, G. (2001). Revisiting scene primes for object locations. *The Quarterly Journal of Experimental Psychology: Section A*, *54*(3), 683-693.
Gernsbacher, M. A. (1990). *Language comprehension as structure building* (Vol. xi). Lawrence Erlbaum Associates, Inc.
Goffaux, V., Jacques, C., Mouraux, A., Oliva, A., Schyns, P. G., & Rossion, B. (2005). Diagnostic colours contribute to the early stages of scene categorization: Behavioural and neurophysiological evidence. *Visual Cognition*, *12*(6), 878-892. <Go to ISI>://000232033700003
Green, P., & MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493-498.
Greene, M. R., Baldassano, C., Esteva, A., Beck, D. M., & Fei-Fei, L. (2016). Visual scenes are categorized by function. *Journal of Experimental Psychology: General*, *145*(1), 82.
Greene, M. R., Botros, A. P., Beck, D. M., & Fei-Fei, L. (2015). What you see is what you expect: rapid scene understanding benefits from prior experience [journal article]. *Attention, Perception & Psychophysics*, *77*(4), 1239-1251. https://doi.org/10.3758/s13414-015-0859-8
Greene, M. R., & Hansen, B. C. (2018). Shared spatiotemporal category representations in biological and artificial deep neural networks. *PLOS Computational Biology*, *14*(7), e1006327. https://doi.org/10.1371/journal.pcbi.1006327
Greene, M. R., & Hansen, B. C. (2020). Disentangling the Independent Contributions of Visual and Conceptual Features to the Spatiotemporal Dynamics of Scene Categorization. *Journal of Neuroscience*, *40*(27), 5283-5299.
Greene, M. R., & Oliva, A. (2009a). The briefest of glances: The time course of natural scene understanding. *Psychological Science*, *20*(4), 464-472.
Greene, M. R., & Oliva, A. (2009b). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, *58*(2), 137-176. https://doi.org/10.1016/j.cogpsych.2008.06.001
Gregory, E., & McCloskey, M. (2010). Mirror-image confusions: Implications for representation and processing of object orientation. *Cognition*, *116*(1), 110-129. https://doi.org/http://dx.doi.org/10.1016/j.cognition.2010.04.005
Gregory, R. L. (1990). *Eye and brain: The psychology of seeing* (4th ed.). Princeton University Press.
Hagoort, P. (2007). The fractionation of spoken language understanding by measuring electrical and magnetic brain signals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 1055-1069.
Hagoort, P., Baggio, G., & Willems, R. M. (2009). Semantic unification. In *The cognitive neurosciences, 4th ed.* (pp. 819-836). MIT press.
Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, *304*(5669), 438-441.
Haji-Khamneh, B., & Harris, L. R. (2010). How different types of scenes affect the Subjective Visual Vertical (SVV) and the Perceptual Upright (PU). *Vision Research*, *50*(17), 1720-1727. https://doi.org/10.1016/j.visres.2010.05.027
Halgren, E., Mendola, J., Chong, C. D., & Dale, A. M. (2003). Cortical activation to illusory shapes as measured with magnetoencephalography. *Neuroimage*, *18*(4), 1001-1009.

Hamm, J. P., Johnson, B. W., & Kirk, I. J. (2002). Comparison of the N300 and N400 ERPs to picture stimuli in congruent and incongruent contexts. *Clinical Neurophysiology*, *113*(8), 1339-1350.

Han, S., Jiang, Y., Mao, L., Humphreys, G. W., & Qin, J. (2005). Attentional modulation of perceptual grouping in human visual cortex: ERP studies. *Human Brain Mapping*, *26*(3), 199-209.

Hansen, B. C., & Loschky, L. C. (2013). The contribution of amplitude and phase spectra defined scene statistics to the masking of rapid scene categorization. *Journal of Vision*, *13*(13), 1–21. https://doi.org/10.1167/13.13.21

Hansen, N. E., Noesen, B. T., Nador, J. D., & Harel, A. (2018). The influence of behavioral relevance on the processing of global scene properties: An ERP study. *Neuropsychologia*, *114*, 168-180.

Harel, A., Groen, I. I., Kravitz, D. J., Deouell, L. Y., & Baker, C. I. (2016). The temporal dynamics of scene processing: A multifaceted EEG investigation. *Eneuro*, *3*(5).

Harel, A., Kravitz, D. J., & Baker, C. I. (2013). Deconstructing visual scenes in cortex: gradients of object and spatial layout information. *Cerebral Cortex*, *23*(4), 947-957.

Harel, A., Mzozoyana, M. W., Al Zoubi, H., Nador, J. D., Noesen, B. T., Lowe, M. X., & Cant, J. S. (2020). Artificially-generated scenes demonstrate the importance of global scene properties for scene perception. *Neuropsychologia*, *141*, 107434.

Harris, L. R., Jenkin, M., Dyde, R. T., & Jenkin, H. (2011). Enhancing visual cues to orientation: Suggestions for space travelers and the elderly. In A. M. Green, C. E. Chapman, J. F. Kalaska, & F. Lepore (Eds.), *Progress in Brain Research* (Vol. 191, pp. 133-142). Elsevier.

Hashimoto, R., Tanaka, Y., & Nakano, I. (2010). Heading disorientation: a new test and a possible underlying mechanism. *European neurology*, *63*(2), 87-93.

Hassabis, D., Kumaran, D., & Maguire, E. A. (2007). Using imagination to understand the neural basis of episodic memory. *Journal of Neuroscience*, *27*(52), 14365-14374.

Hassabis, D., & Maguire, E. A. (2009). The construction system of the brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1263-1271.

Hasson, U., Chen, J., & Honey, C. J. (2015). Hierarchical process memory: memory as an integral component of information processing. *Trends in Cognitive Science*, *19*(6), 304-313.

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience*, *28*(10), 2539-2550.

Hebart, M. N., & Baker, C. I. (2018). Deconstructing multivariate decoding for the study of brain function. *Neuroimage*, *180*, 4-18.

Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour*, *1*(10), 743-747.

Hillyard, S. A., Vogel, E. K., & Luck, S. J. (1998). Sensory gain control (amplification) as a mechanism of selective attention: electrophysiological and neuroimaging evidence. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *353*(1373), 1257-1270.

Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, *108*(3), 687-701.

Holcomb, P. J., & Mcpherson, W. B. (1994). Event-related brain potentials reflect semantic priming in an object decision task. *Brain And Cognition*, *24*(2), 259-276.

Hollingworth, A., & Henderson, J. M. (1998). Does consistent scene context facilitate object perception? *Journal of Experimental Psychology: General*, *127*(4), 398-415. https://doi.org/http://dx.doi.org/10.1037/0096-3445.127.4.398

Hollingworth, A., & Henderson, J. M. (1999). Object identification is isolated from scene semantic constraint: Evidence from object type and token discrimination. *Acta Psychologica*, *102*(2-3), 319-343.

Humphreys, G. W., Price, C. J., & Riddoch, M. J. (1999). From objects to names: A cognitive neuroscience approach. *Psychological Research*, *62*(2-3), 118-130.

Iaria, G., Fox, C. J., Chen, J. K., Petrides, M., & Barton, J. J. (2008). Detection of unexpected events during spatial navigation in humans: Bottom-up attentional system and neural mechanisms. *European Journal of Neuroscience*, *27*(4), 1017-1025.

Inhoff, M. C., & Ranganath, C. (2017). Dynamic Cortico-hippocampal Networks Underlying Memory and Cognition: The PMAT Framework. In D. E. Hannula & M. C. Duff (Eds.), *The Hippocampus from Cells to Systems: Structure, Connectivity, and Functional Contributions to Memory and Flexible Cognition* (pp. 559-589). Springer International Publishing. https://doi.org/10.1007/978-3-319-50406-3_18

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434-446. https://doi.org/http://dx.doi.org/10.1016/j.jml.2007.11.007

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving*, *7*(1), 2.

Johnson, J. S., & Olshausen, B. A. (2002). Early target related processing in the discrimination of natural objects. Vision Sciences Society 2002, Sarasota, FL.

Johnson, J. S., & Olshausen, B. A. (2005). The earliest EEG signatures of object recognition in a cued-target task are postsensory. *Journal of Vision*, *5*(4), 2-2.

Kahn, I., Andrews-Hanna, J. R., Vincent, J. L., Snyder, A. Z., & Buckner, R. L. (2008). Distinct cortical anatomy linked to subregions of the medial temporal lobe revealed by intrinsic functional connectivity. *Journal of Neurophysiology*, *100*(1), 129-139.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, *17*(11), 4302-4311.

Kar, K., & DiCarlo, J. J. (2021). Fast recurrent processing via ventrolateral prefrontal cortex is needed by the primate ventral stream for robust core visual object recognition. *Neuron*, *109*(1), 164-176. e165.

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, *22*(6), 974-983.

Kass, R. E., & Rafferty, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *89*(430), 773-795.

Kelley, T. A., Chun, M. M., & Chua, K.-P. (2003). Effects of scene inversion on change detection of targets matched for visual salience. *Journal of Vision*, *3*(1), 1-5. https://doi.org/10.1167/3.1.1

Kenemans, J., Kok, A., & Smulders, F. (1993). Event-related potentials to conjunctions of spatial frequency and orientation as a function of stimulus parameters and response requirements. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, *88*(1), 51-63.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 983-997.
Key, A. P. F., Dove, G. O., & Maguire, M. J. (2005). Linking brainwaves to the brain: an ERP primer. *Developmental neuropsychology*, *27*(2), 183-215.
Kirchner, H., & Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research*, *46*(11), 1762-1776. https://doi.org/10.1016/j.visres.2005.10.002
Kosslyn, S. M., Alpert, N. M., Thompson, W. L., Chabris, C. F., Rauch, S. L., & Anderson, A. K. (1994). Identifying objects seen from different viewpoints A PET investigation. *Brain*, *117*(5), 1055-1071.
Kovacs, G., Vogels, R., & Orban, G. A. (1995). Cortical correlate of pattern backward-masking. *Proceedings of the National Academy of Sciences of the United States of America*, *92*(12), 5587-5591.
Kravitz, D. J., Saleem, K. S., Baker, C. I., & Mishkin, M. (2011). A new neural framework for visuospatial processing [10.1038/nrn3008]. *Nature Reviews Neuroscience*, *12*(4), 217-230. https://doi.org/10.1038/nrn3008
Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26). Springer.
Kumar, M., Federmeier, K. D., & Beck, D. M. (2021). The N300: An Index For Predictive Coding Of Complex Visual Objects and Scenes. *Cerebral Cortex Communications*, *2*(2), tgab030.
Kumle, L., Vo, M. L.-H., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods*, 1-16.
Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, *4*(12), 463-470.
Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, *62*, 621-647.
Kutas, M., & Hillyard, S. (1983). Event-related brain potentials to grammatical errors and semantic anomalies. *Memory & Cognition*, *11*(5), 539-550. https://doi.org/10.3758/bf03196991
Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203-205.
Kveraga, K., Ghuman, A. S., & Bar, M. (2007). Top-down predictions in the cognitive brain. *Brain And Cognition*, *65*(2), 145-168.
Larson, A. M., Freeman, T. E., Ringer, R. V., & Loschky, L. C. (2014). The spatiotemporal dynamics of scene gist recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(2), 471-487. https://doi.org/10.1037/a0034986
Lauer, T., Cornelissen, T. H., Draschkow, D., Willenbockel, V., & Võ, M. L.-H. (2018). The role of scene summary statistics in object recognition. *Scientific Reports*, *8*(1), 1-12.
Leech, R., & Sharp, D. J. (2014). The role of the posterior cingulate cortex in cognition and disease. *Brain*, *137*(1), 12-32.
Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2018). Emmeans: Estimated marginal means, aka least-squares means. *R package version*, *1*(1), 3.
Leroy, A., Faure, S., & Spotorno, S. (2020). Reciprocal semantic predictions drive categorization of scene contexts and objects even when they are separate. *Scientific Reports*, *10*(1), 1-12.

Libby, L. A., Ekstrom, A. D., Ragland, J. D., & Ranganath, C. (2012). Differential connectivity of perirhinal and parahippocampal cortices within human hippocampal subregions revealed by high-resolution functional imaging. *Journal of Neuroscience, 32*(19), 6550-6560.

Loftus, G. R., & Mclean, J. E. (1999). A front end to a theory of picture recognition. *Psychonomic Bulletin and Review, 6*(3), 394-411.

Loschky, L. C., Hansen, B. C., Sethi, A., & Pydimari, T. (2010). The role of higher-order image statistics in masking scene gist recognition. *Attention, Perception & Psychophysics, 72*(2), 427-444.

Loschky, L. C., Hutson, J. P., Smith, M. E., Smith, T. J., & Magliano, J. P. (2018). Viewing Static Visual Narratives Through the Lens of the Scene Perception and Event Comprehension Theory (SPECT). In J. Laubrock, J. Wildfeuer, & A. Dunst (Eds.), *Empirical Comics Research: Digital, Multimodal, and Cognitive Methods* (pp. 217-238). Routledge.

Loschky, L. C., Larson, A., Smith, T. J., & Magliano, J. P. (2020). The scene perception & event comprehension theory (SPECT) applied to visual narratives. *Topics in Cognitive Science*, 1-41.

Loschky, L. C., Larson, A. M., Smerchek, S., & Finan, S. (2008). The superordinate natural/man-made distinction is perceived before basic level distinctions in scene gist recognition. *Journal of Vision, 8*(6), 738-738. http://journalofvision.org/8/6/738/

Loschky, L. C., Ringer, R. V., Ellis, K., & Hansen, B. C. (2015). Comparing rapid scene categorization of aerial and terrestrial views: A new perspective on scene gist. *Journal of Vision, 15*(6:11), 1–29. https://doi.org/doi:10.1167/15.6.11

Loschky, L. C., Sethi, A., Simons, D. J., Pydimari, T., Ochs, D., & Corbeille, J. (2007). The importance of information localization in scene gist recognition. *Journal of Experimental Psychology: Human Perception & Performance, 33*(6), 1431-1450.

Loschky, L. C., Simons, D. J., Smerchek, S., Matz, E., Bilyeu, B., & Artman, L. (2007). Is unlocalized amplitude information of any use for scene Gist recognition? *Journal of Vision, 7*(9), 1051-1051. https://doi.org/10.1167/7.9.1051

Lowe, M. X., Rajsic, J., Ferber, S., & Walther, D. B. (2018). Discriminating scene categories from brain activity within 100 milliseconds. *Cortex, 106*, 275-287.

Lu, Z., Williamson, S., & Kaufman, L. (1992). Behavioral lifetime of human auditory sensory memory predicted by physiological measures. *Science, 258*(5088), 1668-1670.

Luck, S. J. (2014). *An introduction to the event-related potential technique*. MIT press.

Macknik, S. L., & Martinez-Conde, S. (2007). The role of feedback in visual masking and visual processing. *Advances in Cognitive Psychology, 3*(1-2), 125.

Malcolm, G. L., Groen, I. I. A., & Baker, C. I. (2016). Making sense of real-world scenes. *Trends in Cognitive Sciences, 20*(11), 843-856. https://doi.org/10.1016/j.tics.2016.09.003

Malouin, F., Richards, C. L., Jackson, P. L., Dumas, F., & Doyon, J. (2003). Brain activations during motor imagery of locomotor-related tasks: A PET study. *Human Brain Mapping, 19*(1), 47-62.

Mandelbaum, E. (2018). Seeing and conceptualizing: Modularity and the shallow contents of perception.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* W. H. Freeman.

Maruyama, Y., Ogata, Y., Martínez-Tejada, L. A., Koike, Y., & Yoshimura, N. (2020). Independent Components of EEG Activity Correlating with Emotional State. *Brain Sciences*, *10*(10), 669.
Massaro, D. W., & Loftus, G. R. (1996). Sensory and perceptual storage: Data and theory. In *Memory* (pp. 67-99). Elsevier.
Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305-315.
McDermott, K. B., Petersen, S. E., Watson, J. M., & Ojemann, J. G. (2003). A procedure for identifying regions preferentially activated by attention to semantic and phonological relations using functional magnetic resonance imaging. *Neuropsychologia*, *41*(3), 293-303.
McLean, D., Renoult, L., & Malcolm, G. L. (2021). Expectation-Based Gist Facilitation: Rapid Scene Understanding and the Role of Top-Down Information. *BioRxiv*.
McPherson, W. B., & Holcomb, P. J. (1999). An electrophysiological investigation of semantic priming with pictures of real objects. *Psychophysiology*, *36*(1), 53-65.
Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.-C., Lin, C.-C., & Meyer, M. D. (2019). Package 'e1071'. *The R Journal*.
Meyer, D., & Wien, F. T. (2015). Support vector machines. *The Interface to libsvm in package e1071*, *28*.
Molenberghs, P., Cunnington, R., & Mattingley, J. B. (2009). Is the mirror neuron system involved in imitation? A short review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, *33*(7), 975-980.
Morey, R. D., Homer, S., & Proulx, T. (2018). Beyond Statistics: Accepting the Null Hypothesis in Mature Sciences. *Advances in Methods and Practices in Psychological Science*, *1*(2), 245-258. https://doi.org/10.1177/2515245918776023
Moss, J., Schunn, C. D., Schneider, W., McNamara, D. S., & VanLehn, K. (2011). The neural correlates of strategic reading comprehension: Cognitive control and discourse comprehension. *Neuroimage*, *58*(2), 675-686.
Muckli, L., De Martino, F., Vizioli, L., Petro, Lucy S., Smith, Fraser W., Ugurbil, K., Goebel, R., & Yacoub, E. (2015). Contextual Feedback to Superficial Layers of V1. *Current Biology*, *25*(20), 2690-2695. https://doi.org/10.1016/j.cub.2015.08.057
Muckli, L., & Petro, L. S. (2013). Network interactions: non-geniculate input to V1. *Current opinion in neurobiology*, *23*(2), 195-201. https://doi.org/http://dx.doi.org/10.1016/j.conb.2013.01.020
Mudrik, L., Lamy, D., & Deouell, L. Y. (2010). ERP evidence for context congruity effects during simultaneous object–scene processing. *Neuropsychologia*, *48*(2), 507-517.
Munneke, J., Brentari, V., & Peelen, M. (2013). The influence of scene context on object recognition is independent of attentional focus [Original Research]. *Frontiers in Psychology*, *4*(552). https://doi.org/10.3389/fpsyg.2013.00552
O'Craven, K. M., & Kanwisher, N. (2000). Mental imagery of faces and places activates corresponding stimulus-specific brain regions. *Journal of Cognitive Neuroscience*, *12*(6), 1013-1023.
Oliva, A., & Schyns, P. G. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology*, *41*(2), 176-210.
Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145-175.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research, Special Issue on Visual Perception*, *155*, 23-36.
Onton, J., & Makeig, S. (2006). Information-based modeling of event-related brain dynamics. *Progress in Brain Research*, *159*, 99-120.
Oostenveld, R., & Oostendorp, T. F. (2002). Validating the boundary element method for forward and inverse EEG computations in the presence of a hole in the skull. *Human Brain Mapping*, *17*(3), 179-192.
Palmer, S. E. (1975a). The effects of contextual scenes on the identification of objects. *Memory and Cognition*, *3*, 519-526.
Palmer, S. E. (1975b). Visual perception and world knowledge: Notes on a model of sensory-cognitive interaction. In D. A. Norman, D. E. Rumelhart, & t. L. R. Group (Eds.), *Explorations in cognition* (pp. 279-307). Freeman.
Park, S., Brady, T. F., Greene, M. R., & Oliva, A. (2011). Disentangling scene content from spatial boundary: complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes. *The Journal of Neuroscience*, *31*(4), 1333-1340. https://doi.org/10.1523/JNEUROSCI.3885-10.2011
Park, S., Intraub, H., Yi, D.-J., Widders, D., & Chun, M. M. (2007). Beyond the edges of a view: boundary extension in human scene-selective visual cortex. *Neuron*, *54*(2), 335-342.
Peelen, M. V., Fei-Fei, L., & Kastner, S. (2009). Neural mechanisms of rapid natural scene categorization in human visual cortex. *Nature*, *460*(7251), 94-97.
Peelen, M. V., & Kastner, S. (2011). A neural basis for real-world visual search in human occipitotemporal cortex. *Proceedings of the National Academy of Sciences*, *108*(29), 12125-12130.
Peirce, J., & MacAskill, M. (2018). *Building experiments in PsychoPy*. Sage.
Persichetti, A., Weiller, S., Zorn, A., Williams, K., & Dilks, D. (2016). Distinct neural and cognitive systems selectively involved in navigation and categorization of scenes. *Journal of Vision*, *16*(12), 525-525.
Peterson, M. A. (1994). Object recognition processes can and do operate before figure–ground organization. *Current Directions in Psychological Science*, *3*(4), 105-111.
Petro, L., Vizioli, L., & Muckli, L. (2014). Contributions of cortical feedback to sensory processing in primary visual cortex [Review]. *Frontiers in Psychology*, *5*(1223). https://doi.org/10.3389/fpsyg.2014.01223
Pezdek, K., Whetstone, T., Reynolds, K., Askari, N., & Dougherty, T. (1989). Memory for real-world scenes: The role of consistency with schema expectation. *Journal Of Experimental Psychology-Learning Memory And Cognition*, *15*(4), 587-595. <Go to ISI>://A1989AC84100005
Philiastides, M. G., Tu, T., & Sajda, P. (2021). Inferring Macroscale Brain Dynamics via Fusion of Simultaneous EEG-fMRI. *Annual Review of Neuroscience*, *44*.
Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *Neuroimage*, *198*, 181-197.
Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*(1), 3-25.
Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning & Memory*, *2*(5), 509-522.

Potter, M. C., Wyble, B., Hagmann, C. E., & McCourt, E. S. (2014). Detecting meaning in RSVP at 13 ms per picture. *Attention, Perception, & Psychophysics*, *76*(2), 270-279.
Puri, A. M., Wojciulik, E., & Ranganath, C. (2009). Category expectation modulates baseline and stimulus-evoked activity in human inferotemporal cortex. *Brain Research*, *1301*, 89-99.
Radvansky, G. A., & Zacks, J. M. (2014). *Event Cognition*. Oxford University Press.
Raffin, E., Mattout, J., Reilly, K. T., & Giraux, P. (2012). Disentangling motor execution from motor imagery with the phantom limb. *Brain*, *135*(2), 582-595.
Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., & Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences*, *98*(2), 676-682.
Ramkumar, P., Hansen, B. C., Pannasch, S., & Loschky, L. C. (2016). Visual information representation and rapid-scene categorization are simultaneous across cortex: An MEG study. *Neuroimage*, *134*, 295-304. https://doi.org/http://dx.doi.org/10.1016/j.neuroimage.2016.03.027
Ranganath, C., & Ritchey, M. (2012). Two cortical systems for memory-guided behaviour. *Nature Reviews Neuroscience*, *13*(10), 713-726.
Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79 - 87.
Reinitz, M. T., Wright, E., & Loftus, G. R. (1989). Effects of semantic priming on visual encoding of pictures. *Journal of Experimental Psychology: General*, *118*(3), 280-297.
Renninger, L. W., & Malik, J. (2004). When is scene identification just texture recognition? *Vision Research*, *44*, 2301–2311.
Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, *8*(5), 368-373. https://doi.org/10.1111/j.1467-9280.1997.tb00427.x
Rieger, J. W., Braun, C., Bulthoff, H. H., & Gegenfurtner, K. R. (2005). The dynamics of visual pattern masking in natural scene processing: A magnetoencephalography study. *Journal of Vision*, *5*(3), 275-286. <Go to ISI>://000228128200010
Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019-1025.
Ringach, D. L. (2002). Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *Journal of Neurophysiology*, *88*(1), 455-463.
Ritchey, M., Yonelinas, A. P., & Ranganath, C. (2014). Functional connectivity relationships predict similarities in task activation and pattern information during associative memory encoding. *Journal of Cognitive Neuroscience*, *26*(5), 1085-1099.
Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics*, *12*(1), 1-8.
Robinson, A. K., Grootswagers, T., & Carlson, T. A. (2019). The influence of image masking on object representations during rapid serial visual presentation. *Neuroimage*, *197*, 224-231.
Rolls, E., Tovée, M. J., & Panzeri, S. (1999). The neurophysiology of backward visual masking: Information analysis. *Journal of Cognitive Neuroscience*, *11*(3), 300-311.
Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356-374.

Rousselet, G. A., Fabre-Thorpe, M., & Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nature Neuroscience*, *5*(7), 629.
Rousselet, G. A., Joubert, O. R., & Fabre-Thorpe, M. (2005). How long to get to the "gist" of real-world natural scenes? *Visual Cognition*, *12*(6), 852-877. <Go to ISI>://000232033700002
Salin, P. A., & Bullier, J. (1995). Corticocortical connections in the visual system: structure and function. *Physiological reviews*, *75*(1), 107-154.
Sanocki, T. (2003). Representation and perception of spatial layout. *Cognitive Psychology*, *47*, 43-86.
Sanocki, T., & Epstein, W. (1997). Priming spatial layout of scenes. *Psychological Science*, *8*(5), 374-378.
Schendan, H. E. (2019). Memory influences visual cognition across multiple functional states of interactive cortical dynamics. In *Psychology of Learning and Motivation* (Vol. 71, pp. 303-386). Elsevier.
Schendan, H. E., & Kutas, M. (2002). Neurophysiological evidence for two processing times for visual object identification. *Neuropsychologia*, *40*(7), 931-945.
Schendan, H. E., & Kutas, M. (2007). Neurophysiological evidence for the time course of activation of global shape, part, and local contour representations during visual object categorization and memory. *Journal of Cognitive Neuroscience*, *19*(5), 734-749.
Schendan, H. E., & Lucia, L. C. (2010). Object-sensitive activity reflects earlier perceptual and later cognitive processing of visual objects between 95 and 500 ms. *Brain Research*, *1329*, 124-141.
Schendan, H. E., & Maher, S. M. (2009). Object knowledge during entry-level categorization is activated and modified by implicit memory after 200 ms. *Neuroimage*, *44*(4), 1423-1438.
Schyns, P. G. (1998). Diagnostic recognition: Task constraints, object information, and their interactions. *Cognition*, *67*(1-2), 147-179. <Go to ISI>://000075369700006
Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science*, *5*, 195-200.
Sehatpour, P., Molholm, S., Javitt, D. C., & Foxe, J. J. (2006). Spatiotemporal dynamics of human object recognition processing: an integrated high-density electrical mapping and functional imaging study of “closure” processes. *Neuroimage*, *29*(2), 605-618.
Seidl, K. N., Peelen, M. V., & Kastner, S. (2012). Neural evidence for distracter suppression during visual search in real-world scenes. *Journal of Neuroscience*, *32*(34), 11812-11819.
Sereno, S. C., Rayner, K., & Posner, M. I. (1998). Establishing a time-line of word recognition: evidence from eye movements and event-related potentials. *Neuroreport*, *9*(10), 2195-2200.
Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, *104*(15), 6424-6429. https://doi.org/10.1073/pnas.0700622104
Shafer-Skelton, A., & Brady, T. F. (2019). Scene layout priming relies primarily on low-level features rather than scene layout. *Journal of Vision*, *19*(1), 14-14.
Shallice, T., Fletcher, P., Frith, C., Grasby, P., Frackowiak, R., & Dolan, R. J. (1994). Brain regions associated with acquisition and retrieval of verbal episodic memory. *Nature*, *368*(6472), 633-635.

Shmuel, A., Korman, M., Sterkin, A., Harel, M., Ullman, S., Malach, R., & Grinvald, A. (2005). Retinotopic axis specificity and selective clustering of feedback projections from V2 to V1 in the owl monkey. *Journal of Neuroscience*, *25*(8), 2117-2131.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359-1366.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: sustained inattentional blindness for dynamic events. *Perception*, *28*(9), 1059-1074.

Singmann, H., & Kellen, D. (2019). An introduction to mixed models for experimental psychology. *New methods in cognitive psychology*, *28*, 4-31.

Sitnikova, T., Holcomb, P. J., Kiyonaga, K. A., & Kuperberg, G. R. (2008). Two neurocognitive mechanisms of semantic integration during the comprehension of visual real-world events. *Journal of Cognitive Neuroscience*, *20*(11), 2037-2057.

Smith, C. M., & Federmeier, K. D. (2020). Neural Signatures of Learning Novel Object–Scene Associations. *Journal of Cognitive Neuroscience*, 1-20.

Smith, M. E., & Loschky, L. C. (2019). The influence of sequential predictions on scene-gist recognition. *Journal of Vision*, *19*(12), 1-24.

Sperber, R. D., McCauley, C., Ragain, R. D., & Weil, C. M. (1979). Semantic priming effects on picture and word processing. *Memory & Cognition*, *7*(5), 339-345.

Sperling, G. (1967). Successive approximations to a model for short term memory. *Acta Psychologica*, *27*, 285-292.

Stawarczyk, D., Bezdek, M. A., & Zacks, J. M. (2019). Event representations and predictive processing: The role of the midline default network core. *Topics in Cognitive Science*.

Strotzer, M. (2009). One century of brain mapping using Brodmann areas. *Clinical Neuroradiology*, *19*(3), 179-186.

Stroup, W. W. (2012). *Generalized linear mixed models: modern concepts, methods and applications*. CRC press.

Summerfield, C., & De Lange, F. P. (2014). Expectation in perceptual decision making: neural and computational mechanisms. *Nature Reviews Neuroscience*, *15*(11), 745-756.

Summerfield, C., Trittschuh, E. H., Monti, J. M., Mesulam, M.-M., & Egner, T. (2008). Neural repetition suppression reflects fulfilled perceptual expectations. *Nature Neuroscience*, *11*(9), 1004.

Tanaka, S., Honda, M., & Sadato, N. (2005). Modality-specific cognitive function of medial and lateral human Brodmann area 6. *Journal of Neuroscience*, *25*(2), 496-501.

Thorpe, S. J. (2002). Ultra-rapid scene categorization with a wave of spikes. In H. H. Bulthoff (Ed.), *Biologically Motivated Computer Vision* (pp. 335-351). Springer.

Thorpe, S. J., & Fabre-Thorpe, M. (2001). Seeking categories in the brain. *Science*, *291*(5502), 260-263.

Thorpe, S. J., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*(6582), 520-522.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*(4), 766-786. <Go to ISI>://000241047400003

Torralbo, A., Walther, D. B., Chai, B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2013). Good exemplars of natural scene categories elicit clearer patterns than bad exemplars but not greater BOLD activity. *PLoS ONE*, *8*(3), e58594.

Trapp, S., & Bar, M. (2015). Prediction, context, and competition in visual recognition. *Annals of the new York Academy of Sciences*, *1339*(1), 190-198.
Truman, A., & Mudrik, L. (2018). Are incongruent objects harder to identify? The functional significance of the N300 component. *Neuropsychologia*, *117*, 222-232.
Tversky, B., & Hemenway, K. (1983). Categories of environmental scenes. *Cognitive Psychology*, *15*(1), 121-149.
Usui, K., Ikeda, A., Takayama, M., Matsuhashi, M., Yamamoto, J. I., Satoh, T., Begum, T., Mikuni, N., Takahashi, J. B., & Miyamoto, S. (2003). Conversion of semantic information into phonological representation: a function in left posterior basal temporal area. *Brain*, *126*(3), 632-641.
Utevsky, A. V., Smith, D. V., & Huettel, S. A. (2014). Precuneus is a functional core of the default-mode network. *Journal of Neuroscience*, *34*(3), 932-940.
Van Berkum, J. J., Hagoort, P., & Brown, C. M. (1999). Semantic integration in sentences and discourse: Evidence from the N400. *Journal of Cognitive Neuroscience*, *11*(6), 657-671.
Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal Of Psychophysiology*, *83*(2), 176-190.
Vandenberghe, R., Nobre, A. C., & Price, C. (2002). The response of left temporal cortex to sentences. *Journal of Cognitive Neuroscience*, *14*(4), 550-560.
Vann, S. D., Aggleton, J. P., & Maguire, E. A. (2009). What does the retrosplenial cortex do? *Nature Reviews Neuroscience*, *10*(11), 792-802.
VanRullen, R. (2007). The power of the feed-forward sweep. *Advances in Cognitive Psychology*, *3*(1-2), 167.
VanRullen, R., & Thorpe, S. J. (2001a). Is it a bird? Is it a plane? Ultra-rapid visual categorisation of natural and artifactual objects. *Perception*, *30*(6), 655-668. https://doi.org/10.1068/p3029
VanRullen, R., & Thorpe, S. J. (2001b). Rate coding versus temporal order coding: what the retinal ganglion cells tell the visual cortex. *Neural Computation*, *13*(6), 1255-1283.
VanRullen, R., & Thorpe, S. J. (2001c). The time course of visual processing: From early perception to decision-making. *Journal of Cognitive Neuroscience*, *13*(4), 454-461.
VanRullen, R., & Thorpe, S. J. (2002). Surfing a spike wave down the ventral stream. *Vision Research*, *42*(23), 2593-2615.
Võ, M. L.-H., & Wolfe, J. M. (2013). Differential ERP signatures elicited by semantic and syntactic processing in scenes. *Psychological Science*, *24*(9), 1816.
Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192-196.
Wagner, K., Frings, L., Quiske, A., Unterrainer, J., Schwarzwald, R., Spreer, J., Halsband, U., & Schulze-Bonhage, A. (2005). The reliability of fMRI activations in the medial temporal lobes in a verbal episodic memory task. *Neuroimage*, *28*(1), 122-131.
Walther, D. B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2009). Natural scene categories revealed in distributed patterns of activity in the human brain. *The Journal of Neuroscience*, *29*(34), 10573-10581. https://doi.org/10.1523/jneurosci.0559-09.2009
Walther, D. B., Chai, B., Caddigan, E., Beck, D. M., & Fei-Fei, L. (2011). Simple line drawings suffice for functional MRI decoding of natural scene categories. *Proceedings of the National Academy of Sciences*, *108*(23), 9661-9666. https://doi.org/10.1073/pnas.1015666108

Walther, D. B., & Shen, D. (2014). Nonaccidental properties underlie human categorization of complex natural scenes. *Psychological Science*, *25*(4), 851-860. https://doi.org/10.1177/0956797613512662

Watson, D. M., Andrews, T. J., & Hartley, T. (2017). A data driven approach to understanding the organization of high-level visual cortex. *Scientific Reports*, *7*(1), 1-14.

Watson, D. M., Hartley, T., & Andrews, T. J. (2014). Patterns of response to visual scenes are linked to the low-level properties of the image. *Neuroimage*, *99*, 402-410.

West, W. C., & Holcomb, P. J. (2002). Event-related potentials during discourse-level semantic integration of complex pictures. *Cognitive Brain Research*, *13*(3), 363-375. https://doi.org/10.1016/s0926-6410(01)00129-x

Willems, R. M., Özyürek, A., & Hagoort, P. (2008). Seeing and hearing meaning: ERP and fMRI evidence of word versus picture integration into a sentence context. *Journal of Cognitive Neuroscience*, *20*(7), 1235-1249.

Woodman, G. F. (2010). A brief introduction to the use of event-related potentials in studies of perception and attention. *Attention, Perception, & Psychophysics*, *72*(8), 2031-2046.

Zacks, J. M., Speer, N., Swallow, K., Braver, T., & Reynolds, J. (2007). Event perception: A mind-brain perspective. *Psychological Bulletin*, *133*(2), 273-293. https://doi.org/2007-02367-005 [pii]

10.1037/0033-2909.133.2.273

Zeman, P. M., Till, B. C., Livingston, N. J., Tanaka, J. W., & Driessen, P. F. (2007). Independent component analysis and clustering improve signal-to-noise ratio for statistical analysis of event-related potentials. *Clinical Neurophysiology*, *118*(12), 2591-2604.

Appendix A - Analysis of the N300/N400

Violations of perceptual predictions influence the N400 as well as an earlier component known as the N300 (Hamm et al., 2002; Holcomb & Mcpherson, 1994; Kumar et al., 2021; McPherson & Holcomb, 1999; Võ & Wolfe, 2013). The N300 is a negative deflection in the vERP waveform between 250 – 349 ms. It is thought to be more tightly yoked to image specific perceptual processing than the N400 with generators likely in the occipitotemporal cortex (Hamm et al., 2002; Schendan & Maher, 2009; Sehatpour et al., 2006). It appears before the N400, and it is more frontally distributed than the N400 (Hamm et al., 2002; Holcomb & Mcpherson, 1994). In some cases, differential N300 waveforms can be observed without the presence of the N400 (Cohn & Foulsham, 2020). Consequently, many researchers have separated their analyses according to these components (Demiral et al., 2012; Federmeier & Kutas, 2001; Ganis & Kutas, 2003; Lauer et al., 2018; Mudrik et al., 2010; Sitnikova et al., 2008; Võ & Wolfe, 2013; Willems et al., 2008) so we did here as well to be consistent with prior literature. However, questions remain about whether the N300 and N400 reflect distinct or the same underlying process(Draschkow et al., 2018; Truman & Mudrik, 2018).

Amplitudes of the N300 and N400 were taken as the mean of all data points between 250-349 and 350-449 ms over frontal and central cites, consistent with the prior literature (see Table 17).

Table 17. *Electrodes within each region of interest for each of the 3 vERP components of interest. Authors who demonstrated differential activity between expected and unexpected stimuli at each electrode location are provided in the far-right column.*

Component	Regions of Interest		Electrodes	Authors
P200 (150-249 ms)				
	Hemisphere	Area		
	Left	Parieto-Occipital	P3, O1, P7, 21, 25, 26, 27, 32	(Hansen et al., 2018; Harel et al., 2016; Harel et al., 2020; McLean et al., 2021)
	Middle	Parieto-Occipital	Pz, Oz, 31, 33, 36, 38, 40	(Hansen et al., 2018; Harel et al., 2016; Harel et al., 2020; McLean et al., 2021)
	Right	Parieto-Occipital	P4, O2, P8, 41, 43, 45, 46, 48	(Hansen et al., 2018; Harel et al., 2016; Harel et al., 2020; McLean et al., 2021)
N300 (250-349 ms)				
	Left	Frontal	Fp1, F7, F3, 11, 13 14, 17, 64	(Demiral et al., 2012; Draschkow et al., 2018; Kumar et al., 2021; Lauer et al., 2018; Mudrik et al., 2010; Smith &

				Federmeier, 2020; Truman & Mudrik, 2018)
	Middle	Frontal	Fz, 3, 4, 7, 8, 9, 54	(Demiral et al., 2012; Draschkow et al., 2018; Kumar et al., 2021; Lauer et al., 2018; Mudrik et al., 2010; Truman & Mudrik, 2018)
	Right	Frontal	Fp2,F4, F8, 1, 2, 57, 59, 61	(Demiral et al., 2012; Draschkow et al., 2018; Lauer et al., 2018; Mudrik et al., 2010; Smith & Federmeier, 2020; Truman & Mudrik, 2018)
	Left	Central	T7, C1, 15, 16, 19, 22, 23	(Ganis & Kutas, 2003; Hamm et al., 2002; Kumar et al., 2021; Mudrik et al., 2010; Truman & Mudrik, 2018; Võ & Wolfe, 2013)
	Right	Central	C4, T8, 49, 51, 53, 55, 56	(Demiral et al., 2012; Ganis & Kutas, 2003; Hamm et al., 2002; Lauer et al., 2018; Mudrik et al., 2010; Truman & Mudrik, 2018)
N400 (350-449 ms)				

	Left	Frontal	Fp1, F7, F3, 11, 13 14, 17, 64	(Coderre et al., 2020; Cohn & Foulsham, 2020; Demiral et al., 2012; Lauer et al., 2018; Mudrik et al., 2010; Truman & Mudrik, 2018)
	Middle	Frontal	Fz, 3, 4, 7, 8, 9, 54	(Coderre et al., 2020; Cohn & Foulsham, 2020; Demiral et al., 2012; Lauer et al., 2018; Mudrik et al., 2010; Truman & Mudrik, 2018)
	Right	Frontal	Fp2,F4, F8, 1, 2, 57, 59, 61	(Coderre et al., 2020; Cohn & Foulsham, 2020; Demiral et al., 2012; Lauer et al., 2018; Mudrik et al., 2010; Truman & Mudrik, 2018)
	Left	Central	T7, C1, 15, 16, 19, 22, 23	(Coderre et al., 2020; Cohn & Foulsham, 2020; Demiral et al., 2012; Lauer et al., 2018; McLean et al., 2021; Mudrik et al., 2010; Truman & Mudrik, 2018)
	Right	Central	C4, T8, 49, 51, 53, 55, 56	(Coderre et al., 2020; Cohn & Foulsham, 2020; Demiral et al., 2012; Lauer et al., 2018; McLean et al., 2021; Mudrik et al., 2010; Truman & Mudrik, 2018)

Experiment 2: Analyses including the N300

vERPs to the Target Image

We began by analyzing vERPs in response to the target. Least square means from each of the models are shown in Figure 51 and model output is provided in Table 18. Models contained the fixed effects of region, spatiotemporal coherence, location, and all of their interactions. Models contained the by participant intercepts and within subject slope effects and random effects.

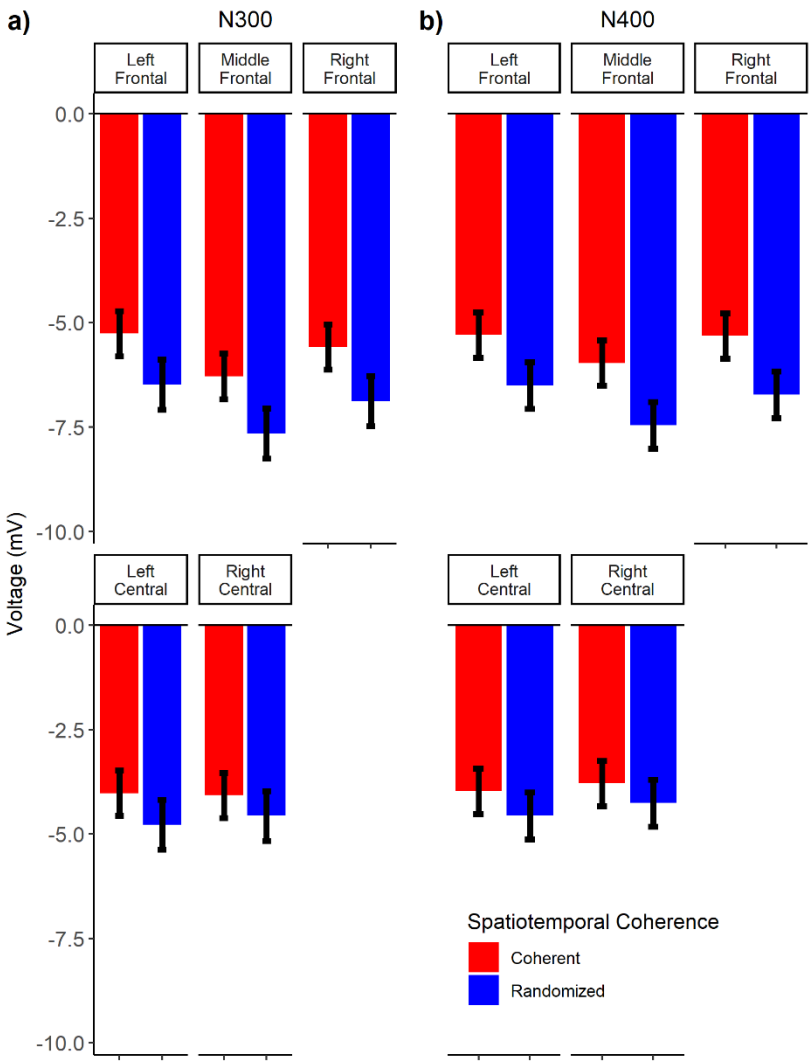


Figure 51. Exp 2: Least square means of amplitudes in response to the target at the frontal and central regions. Amplitudes are reported for the a) N300 and d) N400 components. Error bars correspond to 1 standard error around the estimated means.

Table 18. Exp 2: *Summary of the results for the frontal and central regions. Amplitudes were time locked to the onset of the target scene.*

Component	Factor	*df*	*F*	*β*	*t*	*p*
N300	Region	4,384	105.48			<.001*
	SC	1,24	25.23	1.03	5.02	<.001*
	Location	1,24	4.96	-0.32	-2.22	.04*

	Region*SC	4,384	2.84			.02*
	Paired t-tests (for Region*SC)					
	Left Frontal			1.22	4.23	.003*
	Middle Frontal			1.37	4.75	<.0001*
	Right Frontal			1.29	4.5	.0001*
	Left Central			0.75	2.62	.05
	Right Central			0.49	1.71	.45
	Region*Location	4,384	0.26			.90
	SC*Location	1,24	7.19			.01*
	Paired t-tests (for SC*Location)					
	Off Campus			2.91		.02*
	On Campus			5.03		.0001*
	Region*SC*Location	4,384	1.65			.16
N400	Region	4,384	104.78			<.001*
	SC	1,24	20.47	1.04	4.53	.0001*
	Location	1,24	0.28	-0.09	-0.53	.60
	Region*SC	4,384	4.23			.002*
	Paired t-tests (for Region*SC)					
	Left Frontal			1.52	3.94	.0009*
	Middle Frontal			1.49	4.85	<.0001*
	Right Frontal			1.41	4.59	.0001*
	Left Central			0.6	1.93	.28
	Right Central			0.47	1.53	.65
	Region*Location	4,384	0.33			.86
	SC*Location	1,24	2.46			.13
	Region*SC*Location	4,384	1.65			.16

Note: SC = Spatiotemporal coherence of scene sequences. * denotes p < .05.

N300.

The amplitudes of vERPs to the target image were averaged within the window to capture the N300 for each participant and the resulting amplitudes were submitted to a linear mixed effects model. See Figure 51a). We found a significant main effect for region, $F(4,384) = 105.48$, $p < .001$, BF >1,000 and for spatiotemporal coherence, $F(1,24) = 25.23$, $p < .001$, BF = 11.66. Consistent with the hypothesis that scenes that are easier to categorize elicit a reduced N300, we found that amplitudes were more positive to target scenes in coherent ($M = -5.04$, $SE = 0.52$) than in randomized sequences ($M = -6.07$, $SE = 0.58$). Consistent with prior explorations of the

N300 and its common distribution across the scalp (Hamm et al., 2002; McPherson & Holcomb, 1999), we also found that the difference in response to scenes presented in coherent and randomized sequences was larger in the frontal regions [Right Frontal, β = 1.29, *SE* = 0.29, *t* =4.50, *p* = .0001; Middle Frontal, β = 1.37, *SE* = 0.29, *t* = 4.75, p < .001; Left Frontal, β = 1.22, *SE* = 0.29, *t* =4.26, *p* = .0003] than at central sites [Right Central, β = 0.49, *SE* = 0.29, *t* = 1.71, *p* = .45; Left Central, β = 0.75, *SE* = 0.29, *t* =2.62, p = .05]. This was evident after we probed a significant interaction between spatiotemporal coherence and region, *F*(4,384) = 2.84, *p* = .02, BF > 1,000. Amplitudes were also significantly more positive for on- (*M* = -5.40, *SE* = 0.55) than off-campus sequences (*M* = -5.72, *SE* = 0.55), *F*(1,24) = 4.96, *p* = .04, BF = 1.01, Further, the difference between responses to targets in coherent and randomized sequences was larger for on-, β = 1.42, *SE* = 0.28, *t* =5.03, *p* = .0001, than for off-campus targets, β = 0.63, *SE* = 0.22, *t* =2.91, *p* = .02 as evident from a significant interaction between spatiotemporal coherence and location, *F*(1,24) = 7.19, *p* = .01, BF = 10.01. We did not observe a significant interaction between the region and location, *F*(4,384) = 0.26, *p* = .90, BF <.001 nor a significant three way interaction between the regions, spatiotemporal coherence, and location, *F*(4,384) = 1.65, *p* = .16, BF < .001. See Table 18 for details.

N400.

The amplitudes of vERPs to the target image were averaged within the predetermined time window of 350-449 ms to capture the N400 component. Results were consistent with what we observed in the N300. Importantly, we observed an effect for spatiotemporal coherence, such that amplitudes were more positive in coherent (*M* = -4.87, *SE* = 0.52) than randomized (*M* = -5.90, *SE* = 0.54) sequences, *F*(1,24) = 20.47, *p* < .001, BF = 34.76. Like the N300, the difference between the targets shown in the coherent and randomized sequences was larger at the frontal

[Right Frontal, β = 1.41, *SE* = 0.31, *t* =4.59, *p* = .0001; Middle Frontal, β = 1.49, *SE* = 0.31, *t* = 4.85, *p* < .001; Left Frontal, β = 1.21, *SE* = 0.31, *t* =3.94, *p* = .0009] than at central sites [Right Central, β = 0.47, *SE* = 0.31, *t* = 1.53, *p* = .65; Left Central, β = 0.60, *SE* = 0.31, *t* =1.93, *p* = .28]. This was evident after we probed a significant interaction between region and spatiotemporal coherence, $F(1,384) = 4.23$, $p = .002$, BF = 7.81. Unlike the results from the N300, we did not observe a significant effect for location, $F(1,24) = 0.28$, $p = .60$, BF = 0.03; nor a significant interaction between the spatiotemporal coherence and location, $F(1,384) = 0.33$, $p = .86$, BF = 0.42. We also did not observe an interaction between channel sites and location, $F(4,384) = 1.65$, $p = .16$, BF <.001; or a three-way interaction, $F(4,384) = 1.65$, $p = .16$, BF <.001. See Table 19 for details.

vERPs to all of the images

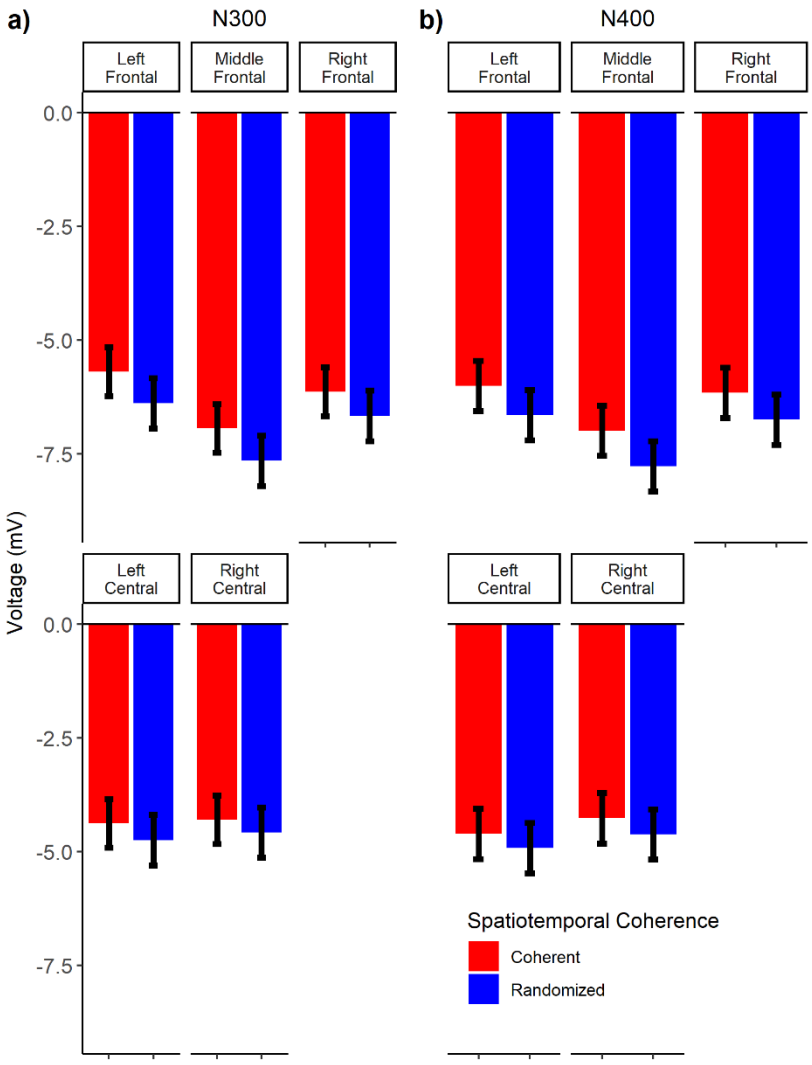


Figure 52. Exp 2: Least square means of amplitudes in response to the all of the scenes excluding behaviorally incorrect trials at the frontal and central regions. Amplitudes are reported for the a) N300 and d) N400 components. Error bars correspond to 1 standard error around the estimated means.

Table 19. Exp 2: *Summary of the results for the frontal/central electrodes. Amplitudes were time locked to the onset of the scenes in the experiment. Observations from the first image within each sequence were removed from the analyses as well as behaviorally incorrect responses to the target.*

Component	Factor	*df*	*F*	*β*	*t*	*p*
N300	Region	4,384	105.48			<.001*
	SC	1,24	25.23	1.03	5.02	<.001*
	Location	1,24	4.96	-0.32	-2.22	.04*
	Region*SC	4,384	2.84			.02*
	Paired t-tests (for Region*SC)					
	Left Frontal			1.22	4.23	.003*
	Middle Frontal			1.37	4.75	<.0001*
	Right Frontal			1.29	4.5	.0001*
	Left Central			0.75	2.62	.05
	Right Central			0.49	1.71	.45
	Region*Location	4,384	0.26			.90
	SC*Location	1,24	7.19			.01*
	Paired t-tests (for SC*Location)					
	Off Campus			2.91		.02*
	On Campus			5.03		.0001*
	Region*SC*Location	4,384	1.65			.16
N400	Region	4,384	104.78			<.001*
	SC	1,24	20.47	1.04	4.53	.0001*
	Location	1,24	0.28	-0.09	-0.53	.60
	Region*SC	4,384	4.23			.002*
	Paired t-tests (for Region*SC)					
	Left Frontal			1.52	3.94	.0009*
	Middle Frontal			1.49	4.85	<.0001*
	Right Frontal			1.41	4.59	.0001*
	Left Central			0.60	1.93	.28
	Right Central			0.47	1.53	.65
	Region*Location	4,384	0.33			.86
	SC*Location	1,24	2.46			.13
	Region*SC*Location	4,384	1.65			.16

N300.

See Figure 52a) . Consistent with responses time locked to the target scene, the N300 was significantly more positive in the central channels [Right Frontal (M = -6.40, SE = 0.53); Middle Frontal (M = -7.30, SE = 0. 53); Left Frontal (M = -6.04, SE = 0. 53); Right Central (M = -4.44, SE = 0. 53); Left Central (M = -4.56, SE = 0. 53)], $F(4,432) = 181.31$, $p < .001$, BF > 1,000. Importantly, the N300 was also significantly more positive for images shown in the coherent (M = -5.49, SE = 0.54) than the randomized (M = -6.00, SE = 0.52) sequences, $F(1,24.04) = 12.70$, p = 002, BF = 3.10; though evidence in favor of this effect is notably small as evident from a small Bayes factor in favor of the alternative hypothesis. We also observed a significant effect of location such that amplitudes were significantly more positive for on- (M = -5.96, SE = 0.55) than off-campus scenes (M = -5.96, SE = 0.55), $F(1,77.44) = 24.26$, $p < .001$, BF = 1.36. None of the interactions were statistically significant. See Table 19.

N400.

See Figure 52b). Results were consistent with what we observed in the N300. Again, we found a main effect of region, $F(4,432) = 173.90$, $p < .001$, BF > 1,000. Importantly, the N400 was significantly more positive for images shown in the coherent sequences (M = -5.61, SE = 0.54) than randomized sequences (M = -6.14, SE = 0.54), $F(1,24.04) = 12.22$, p = 002, BF = 1.45 as we hypothesized. Consistent with the N300, we also observed a significant effect of location such that amplitudes were significantly more positive for on- (M = -5.71, SE = 0.53) than off-campus sequences (M = -6.04, SE = 0.54) , $F(1,252.06) = 15.80$, $p < .001$, BF = 2.96. Again, none of the interactions were statistically significant. See Table 19.

Changes in vERPs within a trial

We also evaluated how the N300 and N400 changed as a function of the ordinal position of the scenes on each trial. Results are reported in Table 20 and least square means from the models are provided in Figure 53.

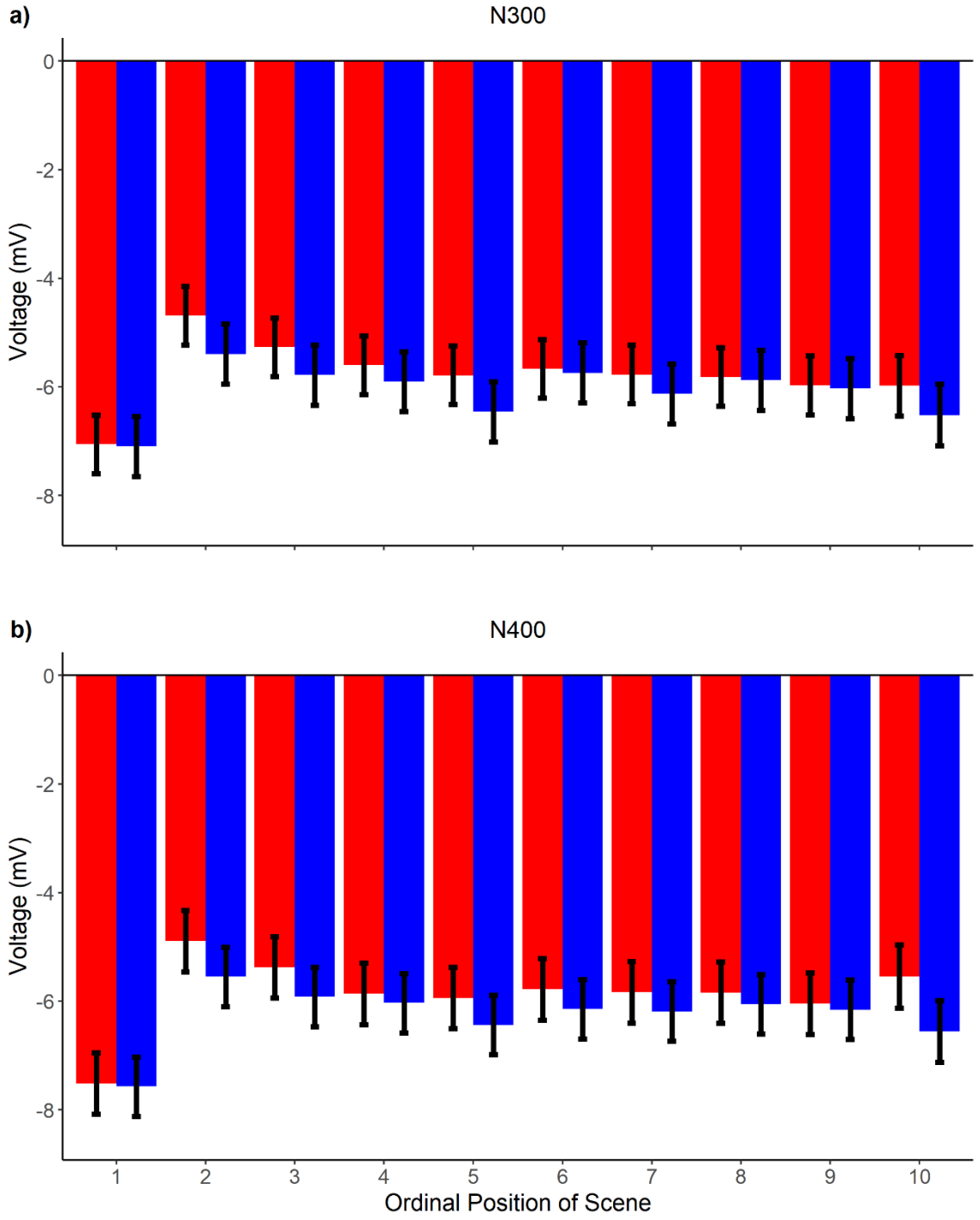


Figure 53. Exp 2: Amplitudes at each ordinal position (1-10), excluding behaviorally incorrect trials. Responses to images in coherent sequences are in red, and responses to images in randomized sequences are in blue.

Table 20. Exp 2: *Summary of the results for the frontal/central electrodes. Amplitudes were time locked to the onset of the scenes in the experiment in the 1st through the 10th position.*

Component	Factor	*df*	*F*	*p*
N300	Region	4,408	530.08	<.001*
	SC	1,24	3.09	.09
	Location	1,25	30.49	<.001*
	Ordinal Position	94,408	53.15	<.001*
	Region*SC	44,408	1.34	.25
	Region*Location	44,408	0.63	.64
	SC*Location	14,408	0.01	.92
	Region* Ordinal Position	364,408	0.56	.98
	SC* Ordinal Position	94,410	3.28	<.001*
	Location* Ordinal Position	94,394	6.48	<.001*
	Region*SC*Location	44,408	0.19	.95
	Region*SC* Ordinal Position	364,408	0.16	.99
	Region*Location* Ordinal Position	364,408	0.33	.99
	SC*Location* Ordinal Position	94,408	2.35	.01*
	Channels*SC*Location* Ordinal Position	364,408	0.31	.99
N400	Region	4,408	465.07	<.001*
	SC	1,24	3.88	.06
	Location	1,25	22.21	<.001*
	Ordinal Position	94,408	58.66	<.001*
	Region*SC	44,408	1.59	.17
	Region*Location	44,408	0.35	.84
	SC*Location	14,408	3.04	.08
	Region* Ordinal Position	364,408	0.81	.79
	SC* Ordinal Position	94,410	2.34	.01*
	Location* Ordinal Position	94,394	5.05	<.001*
	Region*SC*Location	44,408	0.20	.94
	Region*SC* Ordinal Position	364,408	0.26	.99
	Region*Location* Ordinal Position	364,408	0.29	.99
	SC*Location* Ordinal Position	94,408	3.17	<.001*
	Channels*SC*Location* Ordinal Position	364,408	0.27	.99

N300.

We found a main effect for region, F(4,408) = 530.08, p <.001. BF > 1, 000 as well as a marginally significant effect of spatiotemporal coherence, $F(1,24) = 3.09$, $p = .09$, BF = 1.02.

Consistent with the analysis of the 150-249 ms window reported in the paper, the significant three-way interaction between spatiotemporal coherence, the ordinal position of the image, and the location the image was taken (on-campus vs. off-campus), $F(9,4408) = 2.35$, $p = .01$, BF = 3.41, revealed that amplitudes did not significantly differ in response to images in coherent and randomized sequences for the first image on a trial for both on-, $\beta = 0.21$, $SE = 0.27$, $t = 0.77$, $p = .44$ and off-campus sequences, $\beta = -0.14$, SE = 0.27, $t = -0.50$, $p = .62$, but they did at the remaining ordinal positions, as we hypothesized. For on-campus images, amplitudes were significantly more positive in response to images in the coherent sequences in the 2nd, 3rd, 4th, 5th, and 7th positions, and they were numerically greater in the remaining positions. Further, amplitudes in coherent off-campus sequences were significantly more positive than in the randomized sequences in the 2nd, 4th, 5th, and 7th positions (all Bonferroni corrected p values < .05). Amplitudes were numerically greater in the remaining positions for the images taken off-campus. Thus, the event model begins to facilitate scene perception after the first image on a trial. The remaining significant interactions were the same as those observed in the analysis of 50-149 and the 150-249 ms windows. See Table 20.

N400.

Results in the N400 were analogous to what we observed in the N300. We again observed a significant three-way interaction between spatiotemporal coherence, location the image was taken, and the location of the image on the trial, $F(9, 4,408) = 3.17$, $p < .001$, BF > 1,000. Amplitudes in response to images in the first position did not significantly differ for on-campus, $\beta = -0.001$, $SE = 0.29$, $t = -0.005$, $p = .99$ or off-campus scenes, $\beta = 0.11$, $SE = 0.29$, $t = 0.38$, $p = .70$ as we hypothesized; however, they did after the first scene. Coherent on-campus sequences were significantly more positive in the 2nd, 3rd, 6th, and 10th positions. Coherent off-

campus sequences were significantly more positive than randomized off-campus sequences in the 2nd, 5th, 7th, and 9th positions. Thus, we observed evidence of facilitation after the first image in both on- and off-campus sequences. The remaining significant interactions were consistent with the previous 3 analyses. See Table 20 for details.

Experiment 3: Analyses including the N300

vERPs to the Target Image

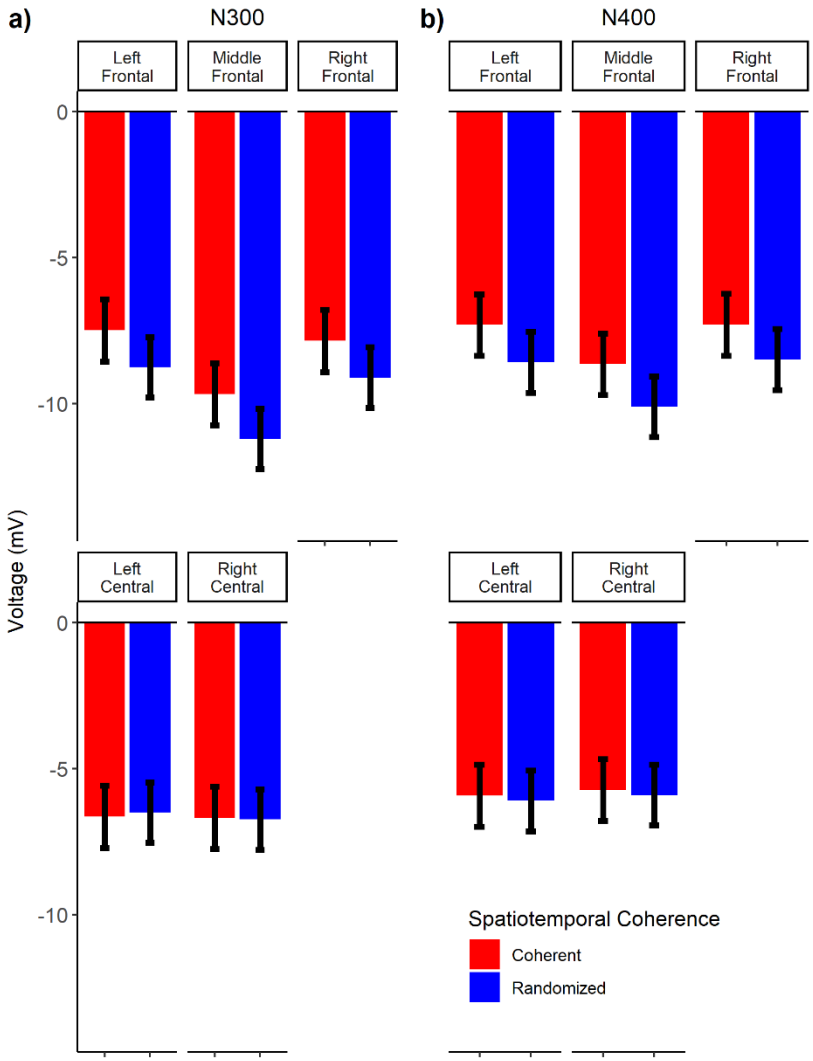


Figure 54. Exp 3: Least square means of amplitudes in response to the target scene at the frontal and central regions. Amplitudes are reported for the a) N300 and d) N400 components. Error bars correspond to 1 standard error around the estimated means.

Table 21. Exp 3: *Summary of the results for the frontal and central regions. Amplitudes were time locked to the onset of the target scene.*

Component	Factor	*df*	*F*	*β*	*t*	*p*
N300	Region	4,384	68.74			<.0001*
	SC	1,24	18.37	-0.05	-5.09	<.001*
	Location	1,24	0.04	-0.24	-0.39	0.84
	Region*SC	4,384	4.18			0.002*
	Paired t-tests (for Region*SC)					
	Left Frontal			1.26	3.26	0.01*
	Middle Frontal			1.53	3.97	<.001*
	Right Frontal			1.25	3.25	0.01*
	Left Central			-0.15	-0.38	0.99
	Right Central			0.06	0.16	0.99
	Region*Location	4,384	0.04			
	SC*Location	1,24	2.42			0.12
	Region*SC*Location	4,384	1.37			0.24
N400	Region	4,384	54.2			<.0001*
	SC	1,24	15.12	-0.4	-4.69	<.001*
	Location	1,24	0.06	-0.3	-0.46	0.8
	Region*SC	4,384	2.4			0.04*
	Paired t-tests (for Region*SC)					
	Left Frontal			1.28	3.02	0.01*
	Middle Frontal			1.46	3.45	<.001
	Right Frontal			1.19	2.8	0.03*
	Left Central			0.18	0.42	0.99
	Right Central			0.18	0.41	0.99
	Region*Location	4,384	0.02			0.99
	SC*Location	1,24	1.14			0.29
	Region*SC*Location	4,384	1.11			0.35

N300.

The effects we observed in the N300 were consistent with what we found in Experiment 2. As shown in Figure 54a), we observed a significant main effect for the region, $F(4,384) = 68.74$, $p < .001$, BF > 1,000, and spatiotemporal coherence, $F(1,24) = 18.37$, $p < .001$, BF = 6.94; such that amplitudes in coherent ($M = -7.67$, $SE = 1.04$) sequences were more positive than

amplitudes in the randomized (M = -8.46, SE = 1.01) sequences. As is typical of the N300, the difference in the amplitudes between the conditions was larger in frontal [Right Frontal, β = 1.25, SE = 0.39, t =3.25, p = .006; Middle Frontal, β = 1.53, SE = 0.39, t = 3.97, p < .001; Left Frontal, β = 1.26, SE = 0.39, t =3.26, p < .01] electrodes than at central regions [Right Central, β = 0.06, SE = 0.39, t = 0.16, p = .99; Left Central, β = -0.15, SE = 0.39, t =-0.38, p = .99]. We found this after we probed a significant interaction between spatiotemporal coherence and region, F(4,384) = 4.18, p = .002, BF = 12.85. None of the remaining effects were statistically significant. See Table 21 for details.

N400.

As shown in Figure 54b), we observed analogous effects to what we observed in the N300 in the N400. We observed a significant effect of region, F(4,384) = 54.2, p <.001, BF > 1,000; spatiotemporal coherence, F(1,24) = 15.12, p <.001, BF = 6.92; and an interaction between region and spatiotemporal coherence, F(4,384) = 2.40, p = .04, BF = 5.17. Again, the difference in the average amplitudes of coherent (M = -6.98, SE = 1.03) and randomized (M = -7.83, SE = 1.01) sequences was larger in the frontal electrodes [Right Frontal, β = 1.19, SE = 0.42, t =2.80, p = .03; Middle Frontal, β = 1.46, SE = 0.42, t = 3.45, p = .004; Left Frontal, β = 1.28, SE = 0.42, t =3.02, p = .01] than at central electrodes [Right Central, β = 0.18, SE = 0.42, t = 0.41, p = .99; Left Central, β = 0.18, SE = 0.42, t = 0.42, p = .99]. None of the remaining effects were statistically significant. See Table 21 for details.

vERPs to all of the images

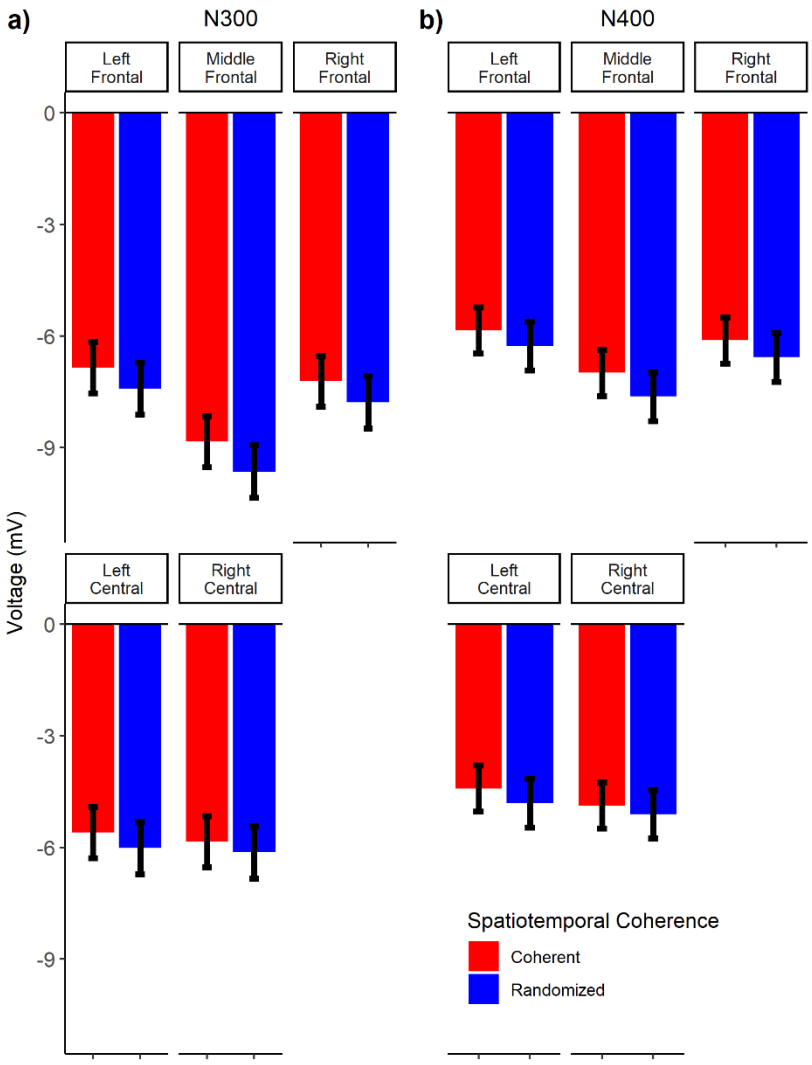


Figure 55. Exp 3: Least square means of amplitudes in response to all of the scenes, excluding behaviorally incorrect trials, at the frontal and central regions.

Table 22. Exp 3: *Summary of the results for the frontal and central electrodes. Amplitudes were time locked to the onset of all of the scenes within a trial excluding observations from the first image within each trial and behaviorally incorrect responses to the target.*

Component	Factor	*df*	*F*	*β*	*t*	*p*

N300	Region	4,432	96.73			<.001*
	SC	1,24	17.24	-0.51	-3.27	<.001*
	Location	1,77	1.11	-0.10	-0.26	.29
	Region*SC	4,432	0.46			.76
	Region*Location	4,432	0.13			.97
	SC*Location	1,147	0.27			.60
	Region*SC*Location	4,432	0.02			.99
N400	Region	4,432	70.18			<.001*
	SC	1.24	9.16	-0.45	-2.19	.01*
	Location	1,252	0.46	-0.11	-0.30	.50
	Region*SC	4,432	0.31			.87
	Region*Location	4,432	0.13			.97
	SC*Location	1,96	0.52			.47
	Region*SC*Location	4,432	0.08			.99

Note: SC = Spatiotemporal coherence of scene sequences. * denotes p < .05.

N300.

See Figure 55a). Again, the N300 was more positive at central [Right Central (*M* = -5.99, *SE* = 0.91); Left Central (*M* = -5.81, *SE* = 0.91)] than frontal regions [Right Frontal (*M* = -7.50, *SE* = 0.91); Middle Frontal (M = -9.24, SE = 0.91); Left Frontal (*M* = -7.13, *SE* = 0.91)], *F*(4, 432) = 81.89, *p* <.001 , BF>1,000. The N300 was also more positive for scenes shown in coherent (*M* = -6.87, *SE* = 0.90) than randomized (*M* = -7.40, *SE* = 0.92) sequences, *F*(1, 24) = 25.26, *p* <.001, BF = 6.88. None of the remaining effects were statistically significant. See Table 22 for details.

N400.

See Figure 55b). Again, we observed a significant main effect of region, *F*(4, 432) = 70.18, *p* < .001, BF > 1,000. Importantly, the N400 was significantly more positive in the coherent (*M* = -5.65, *SE* = 0.81) than in the randomized (*M* = -6.08, *SE* = 0.86) sequences, *F*(1, 24) = 9.16, *p* = .01, BF = 3.63. Thus, we replicated the important effects we observed in Experiment 2 when we time locked the waveforms to all of the scenes. None of the remaining effects were significant. See Table 22.

Changes in vERPs within a trial

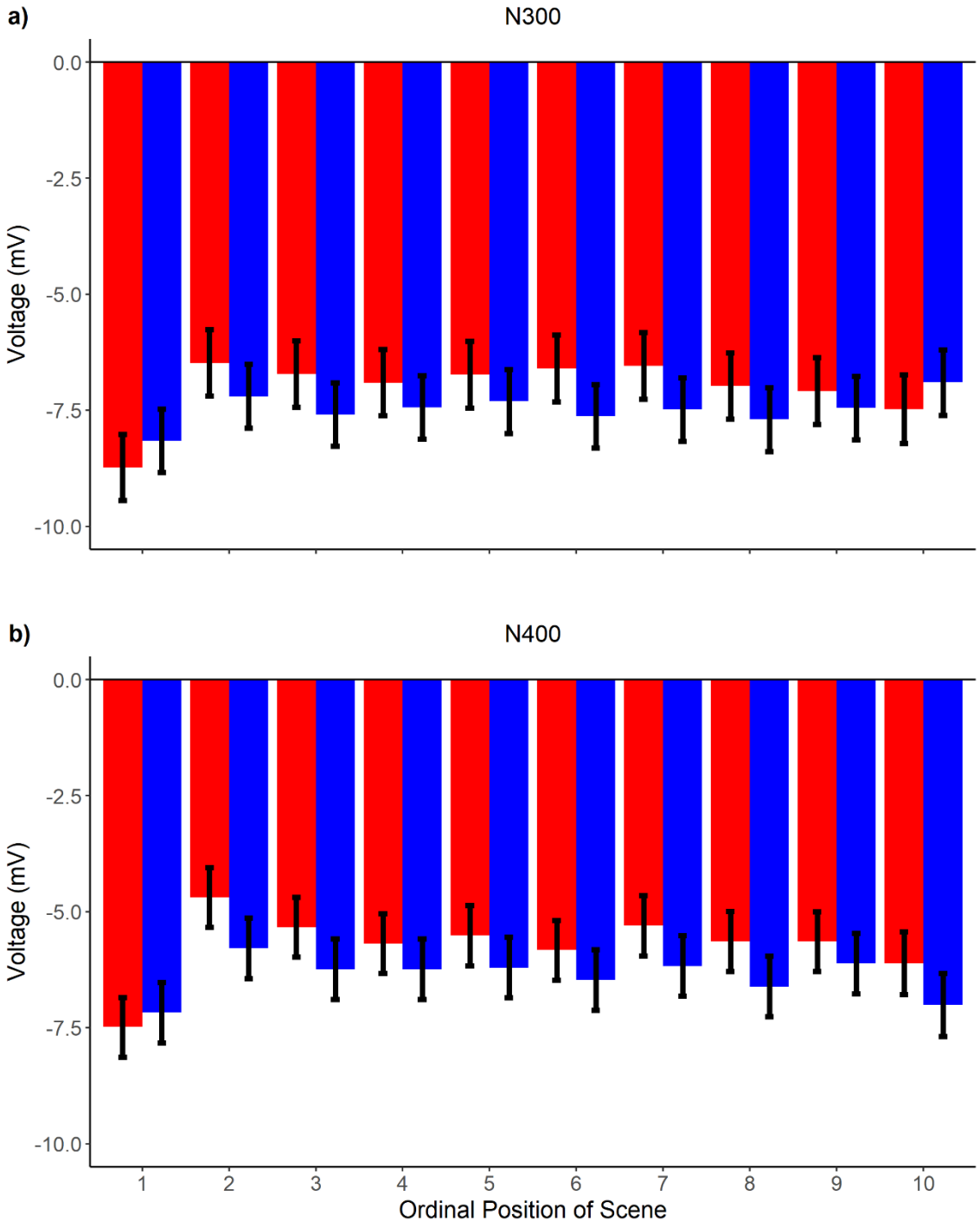


Figure 56. Exp 3: Amplitudes at each ordinal position (1-10), excluding behaviorally incorrect trials. Responses to images in coherent sequences are in red, and responses to images in randomized sequences are in blue.

Table 23. Exp 3: *Summary of the results for the frontal/central electrodes. Amplitudes were time locked to the onset of the scenes in the experiment in the 1st through the 10th position.*

Component	Factor	*df*	*F*	*p*
N300	Region	4.4408	301.68	<.001*
	SC	1,24	13.57	<.01*
	Location	1,25	1.71	.20
	Ordinal Position	94,408	14.93	<.01*
	Region*SC	44,408	0.63	.64
	Region*Location	44,408	4.99	<.001*
	SC*Location	14,408	21.59	<.001*
	Region* Ordinal Position	364,408	1.39	.06
	SC* Ordinal Position	94,410	3.02	<.001*
	Location* Ordinal Position	94,394	11.23	<.001*
	Region*SC*Location	44,408	0.13	.97
	Region*SC* Ordinal Position	364,408	0.23	.99
	Region*Location* Ordinal Position	364,408	1.60	.01*
	SC*Location* Ordinal Position	94,408	2.10	.03*
	Channels*SC*Location* Ordinal Position	364,408	0.15	.99
N400	Region	4.4408	175.57	<.001*
	SC	1,24	18.52	<.001*
	Location	1,25	3.21	.09
	Ordinal Position	94,408	17.80	<.001*
	Region*SC	44,408	0.85	.50
	Region*Location	44,408	5.67	<.001*
	SC*Location	14,408	10.95	<.001*
	Region* Ordinal Position	364,408	1.80	.002*
	SC* Ordinal Position	94,410	1.86	.05
	Location* Ordinal Position	94,394	11.43	<.001*
	Region*SC*Location	44,408	0.06	.99
	Region*SC* Ordinal Position	364,408	0.23	.99
	Region*Location* Ordinal Position	364,408	2.05	<.001*
	SC*Location* Ordinal Position	94,408	2.12	<.001*
	Channels*SC*Location* Ordinal Position	364,408	0.16	0.99

Note: SC = Spatiotemporal coherence of scene sequences. * denotes $p < .05$.

N300.

See Figure 56a). Effects were consistent with what we observed when we limited the analysis to the 150-249 millisecond window reported in the paper except we also observed a statistically significant three-way interaction between spatiotemporal coherence, the ordinal position of the image, and the location where the images were taken (on-campus vs. off-campus), $F(94,408) = 2.10$, $p = .03$, BF = 3.39. The N300 did not significantly differ between coherent and randomized sequences for the first scene in both the on-, $\beta = -0.08$, $SE = 0.35$, $t = -0.24$, $p = .81$ and off-campus sequences, $\beta = -0.56$, $SE = 0.35$, $t = -1.63$, $p = .10$, but it did in the remaining positions. For on-campus sequences, the N300 was significantly more in the coherent sequences in all of the positions after the first scene. Further, amplitudes in coherent off-campus sequences were significantly more positive than in the randomized sequences in the 2nd, 3rd, and the 6th positions (all Bonferroni corrected p values < .05). Amplitudes were numerically greater in the remaining positions for the photographs taken off-campus. The remaining significant interactions were the same as those observed in the analysis of 50-149 and the 150-249 ms windows reported in the paper. See Table 23.

N400.

See Figure 56b). Effects were all consistent with what we observed when we analyzed the N300 including a statistically significant three-way interaction between spatiotemporal coherence, the ordinal position of the image, and the location the image was taken (on-campus vs. off-campus), $F(94,408) = 2.12$, $p < .001$, , BF > 1,000. Like the N300, the N400 did not significantly differ between coherent and randomized sequences for the first image in a sequence in both the on-, $\beta = -0.007$, $SE = 0.37$, $t = 0.02$, $p = .98$ and off-campus sequences, $\beta = -0.13$, $SE = 0.37$, $t = -0.35$, $p = .72$. The N400 was more positive in response to scenes shown in the

coherent sequences at all of the remaining ordinal positions in the on-campus sequences and was more positive in the 2nd 3rd 8th ordinal positions in the off-campus sequences. Amplitudes were numerically more positive in the coherent sequences in the positions that did not show a statistically significant difference. The remaining statistically significant interactions were the same as those observed in the analysis of the N300. See Table 23.

Appendix B - Analysis with N300 removed

vERPs to all of the images

Frontal/Central Electrodes.

We included the same fixed and random effects in the linear mixed effects models as we used when we analyzed vERPs after the onset of the target scene, which were the same as those used in Experiment 2. Figure 57 shows the least square means of amplitude at each window, and Table 24 reports the results of each model.

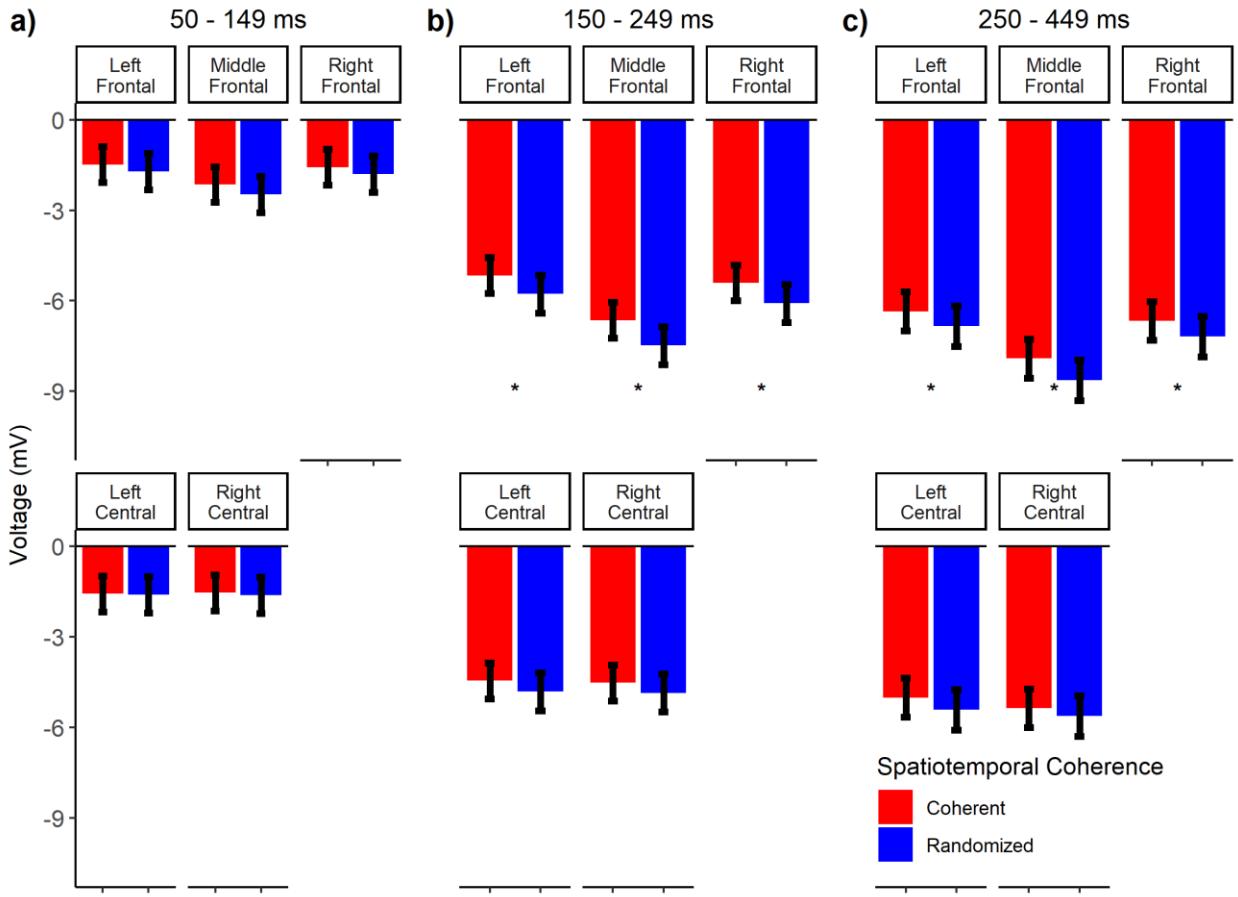


Figure 57. Exp 3: Least square means of amplitudes in response to all of the scenes, excluding behaviorally incorrect trials, at the frontal and central regions. Amplitudes are reported for the a) 50-149, b) 150-249, and c) 250-449 windows.

Table 24. Exp 3: *Summary of the results for the frontal and central electrodes. Amplitudes were time locked to the onset of the scenes. Observations from the first image within each sequence were removed from the analyses as well as behaviorally incorrect responses to the target.*

Window	Factor	*df*	*F*	*β*	*t*	*p*
50-149						
ms	Region	4,384	29.01			<.0001
	SC	1,24	2.68	-0.09	-0.45	.12
	Location	1,24	6.23	-0.19	-1.09	.02*
	Region*SC	4,384	1.00			.41
	Region*Location	4,384	0.01			.99
	SC*Location	1,24	0.45			.51
	Region*SC*Location	4,384	0.04			.99
150-249						
ms	Region	4,432	81.89	-0.45	-3.40	<.001
	SC	1,24	25.26	-0.03	-3.08	<.001
	Location	1,44	1.72			.19
	Region*SC	4,432	0.98			.42
	Region*Location	4,432	0.08			.99
	SC*Location	1,189	0.54			.47
	Region*SC*Location	4,432	0.04			.99
250-449						
ms	Region	4,432	86.33			<.001
	SC	1,24	14.92	-0.48	3.86	<.001
	Location	1,77	0.81	-0.11	0.90	.37
	Region*SC	4,432	0.40			.81
	Region*Location	4,432	0.13			.97
	SC*Location	1,147	0.40			.53
	Region*SC*Location	4,432	0.05			.99

50-149 ms window.

See Figure 57a). Results were consistent with what we observed in Experiment 2, and thus inconsistent with early facilitation accounts of facilitation. There was a significant main effect of region, $F(4,384) = 29.01$, $p < .001$, BF > 1,000. Also consistent with the previous results in the same window, we found that amplitudes in coherent sequences (M = -1.66 SE = 0.28) did not significantly differ from images shown in randomized sequences (M = -1.84, SE = 0.28), $F(1,24) = 2.68$, p =0.12, BF = .06. Thus, we have no evidence from this analysis of the vERPs to suggest that scenes are processed any differently when they are shown in coherent or randomized sequences between 50-149 milliseconds consistent with feed-forward models of scene perception. None of the interactions with spatiotemporal coherence were significant. See Table 24 for details.

150-249 ms window.

See Figure 57b). Results were consistent with Experiment 2. We observed a significant main effect for region, $F(4, 432) = 81.89$, p <.001, BF > 1,000, but more importantly, we found that the event model facilitated processes that occur when matching the structural description of the scene to representations stored in semantic memory, as evident from a significant main effect of spatiotemporal coherence, $F(1, 24) = 25.26$, $p < .001$, BF = 10.09. Again, amplitudes were more negative in response to scenes presented in coherent (M = -5.23, SE = 0.77) than randomized (M = -5.80, SE = 0.83) sequences. None of the remaining effects were significant. See Table 24 for details.

250-449 ms window.

See Figure 57c). Again, we found that scenes shown in coherent sequences were easier to map onto the event model than scenes shown in randomized sequences. We found a significant main effect of region, F(4, 432) = 86.33, $p < .001$, BF > 1,000. More importantly, the N400 was

significantly more positive in the coherent (M = -6.26, SE = 0.84) than in the randomized (M = -6.74, SE = 0.88) sequences, $F(1, 24) = 21.82$, p <.001, BF = 3.11. Thus, we replicated the most important effects we observed in Experiment 2 when we time locked the waveforms to all of the scenes. None of the remaining effects were significant. See Table 24.

Parietal/Occipital Electrodes

Linear mixed effects models for the analysis of the early component (50-149 ms) and the P200 (150-249 ms) contained the same fixed and random effects as the models run on amplitudes time locked to the target, and those used in Experiment 2. Results from the individual analyses at each time window are shown in Figure 58 and Table 25, respectively. We found no evidence to suggest that the event model influences processing in the early component or in the P200 in Experiment 2. Results were consistent in Experiment 3.

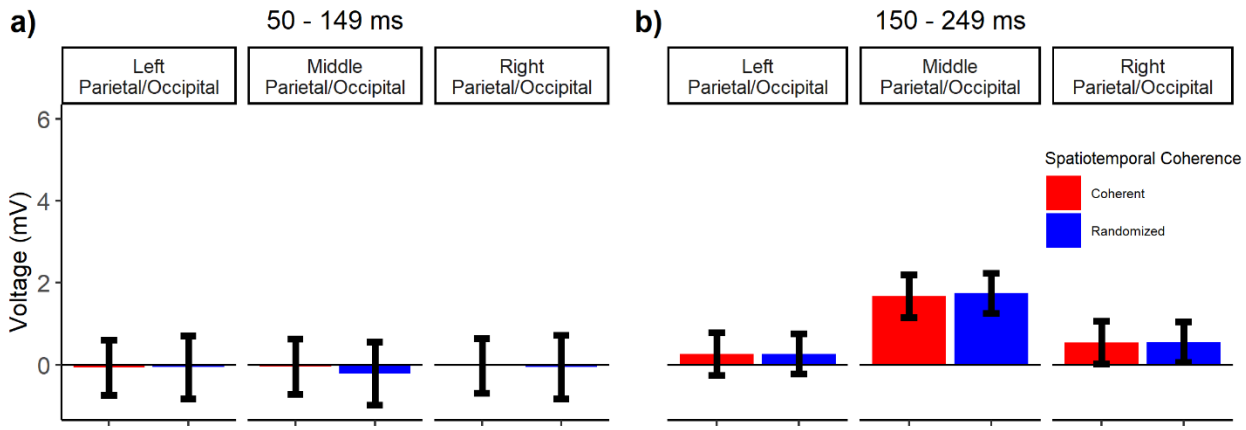


Figure 58. Exp 3: Least square means of amplitudes in response to all of the scenes, excluding behaviorally incorrect trials, at the parietal/occipital regions. Amplitudes are reported for the a) 50-149, b) 150-249, windows.

Table 25. Exp 3: *Summary of the results for the parietal/occipital electrodes. Amplitudes were time locked to the onset of the scenes. Observations from the first scene within each sequence were removed from the analyses as well as behaviorally incorrect responses to the target.*

Window	Factor	*df*	*F*	*β*	*t*	*p*
50-149 ms	Region	2,240	0.18			.84
	SC	1,25.13	0.22	0.0003	0.001	.64
	Location	1,97.16	0.00	-0.06	-0.20	.95
	Region*SC	2,240	0.18			.83
	Region*Location	2,240	0.08			.93
	SC*Location	1,191.47	0.13			.72
	Region*SC*Location	2,240	0.05			.95
150-249 ms	Region	2,264	19.39			<.0001
	SC	1,179.72	0.02	0.03	0.06	.89
	Location	1,104.66	0.75	-0.14	-0.28	.39
	Region*SC	2,264	0.01			.99
	Region*Location	2,264	0.05			.95
	SC*Location	1,209	0.02			.88
	Region*SC*Location	2,264	0.01			.99

50-149 ms window.

See Figure 58a). Unlike in Experiment 2, we did not find a significant effect of region, $F(2, 240) = 0.18$, $p = .84$, $BF = 0.05$. This may have been due to the change in the duration of all of the images within each trial and the onset of the mask in response to target scenes, which were included in this analysis. Consistent with the results of Experiment 2 and inconsistent with early accounts of facilitation, we failed to find a significant effect for spatiotemporal coherence, $F(1, 25) = 0.22$, $p = 64$, $BF = 0.02$. Thus, we have no evidence from vERPs alone to suggest that

predictions made prior to viewing a scene influences early perceptual analysis. None of the remaining effects were significant. See Table 25 for details.

150-249 ms window.

See Figure 58b). The P200 differed between the regions, $F(2,264) = 19.39$, $p < .0001$, BF > 1,000. The P200 was larger in the middle ($M = 1.71$, $SE = 0.48$) parietal/occipital electrodes than it was for the left ($M = 0.26$,$SE = 0.48$), $\beta = -1.45$, $SE = 0.25$, $t = -5.88$, $p < .0001$ and right ($M = 0.55$,$SE = 0.48$) parietal/occipital regions, $\beta = 1.16$, $SE = 0.25$, $t = 4.71$, $p < .0001$ electrodes. In addition, the P200 did not significantly differ between left and right parietal/occipital electrodes, $\beta = -0.29$, $SE = 0.25$, $t = -1.17$, $p = .73$. Consistent with feed-forward models, we failed to find a significant difference in the amplitudes between the coherent ($M = 0.83$, $SE = 0.48$) and randomized ($M = 0.86$, $SE = 0.45$) sequences, $F(1, 179) = 0.02$, $p = .89$, BF = 0.04. Thus, we replicated our initial null result at parietal/occipital electrodes, and this null result was again supported by a small Bayes factor in favor of the alternative and a large Bayes factor in favor of the null (BF = 25).

Analysis of vERP divergence

To examine when waveforms diverged, we conducted a point-by-point t-test at each moment (i.e., each ms) within the epoch. ERPs from all of the scenes in the experiment were included in this analysis, excluding responses to the first image within the trial and all of the incorrect behavioral responses. As in Experiment 2, we averaged the voltage across the electrode regions (Left, Middle, and Right) within the frontal, central, and parietal/occipital sites for the analysis because we did not observe any reliable interactions with region in the component-based analysis in response to all of the images. In addition, we conducted the same Monte Carlo simulation as in Experiment 2. The simulation revealed that a run length of 15 or greater

occurred in 5% of the simulations in the frontal sites, 14 or greater occurred in 5% of the simulations in the central sites, and 15 or greater occurred in 5% of the simulations in the parietal/occipital sites. Thus, this was used as a criterion for considering a given pairwise comparison statistically meaningful. For the frontal region, a difference had to be followed by at least 15 statistically significant time points. For the central region, it had to be followed by at least 14 statistically significant time points, and it had to be followed by at least 15 statistically significant time points in the parietal/occipital regions. We also calculated Bayes factors for each of the statistical comparisons. Results are shown in Figure 59.

As evident Figure 59a) and 59b), results of the analysis of amplitudes for the frontal sites converged with the component-based approach. We replicated the differences we observed in Experiment 2 though waveforms diverged a little later than they did in Experiment 2 (c.f., Figure 17 in the manuscript). Waveforms diverged significantly at 152 milliseconds post scene onset over the frontal region, and the effect remained significant until 375 milliseconds. Waveforms were also significantly different from 425 to 597 milliseconds in the frontal region. Thus, the effect contained the 150 millisecond time point, previously associated with the moment when the visual system begins to discriminate target from non-target scenes in RSVP (Thorpe et al., 1996; VanRullen & Thorpe, 2001a, 2001c), as well as the N400 associated with the ease of semantic integration (Hagoort et al., 2009; Hagoort et al., 2004; Kutas & Federmeier, 2000). This effect was also supported by Bayes factors for each of the individual t-tests. Bayes factors were greater than 3, indicating substantial evidence in favor of the alternative hypothesis, from 164 to 324 milliseconds, and then again at 488 milliseconds. Analogous results were obtained at the central electrode region. Waveforms differed significantly at 195 milliseconds and the differences lasted until 343 milliseconds. Bayes factors were greater than 3 from 207 to 277 milliseconds, and

again from 289 to 328 milliseconds. Consistent with conclusions drawn from Experiment 2, our results suggest that predictions for upcoming scene categories begin to facilitate scene perception approximately 150 milliseconds after scene onset in the frontal and central regions. Interesting differences from Experiment 2 were also shown. First, in Experiment 3, the peak Bayes Factors for the coherence factor were larger for the Frontal and Central regions (max of ~ 9) than in Experiment 2 (max of ~ 5.5), but only from roughly 200-250 ms post-stimulus. Conversely, for the Frontal region, which showed larger Bayes Factors in both Experiments 2 and 3, from roughly 300-550 ms, the Bayes Factors for the coherence factor were smaller in Experiment 3 (BFs < 3) than in Experiment 2 (BFs > 3). If we attribute this to the shorter stimulus durations and masking of the target in Experiment 3, then it suggests they created a larger effect for spatiotemporal coherence in the 200-250 ms window (i.e., during matching), but a smaller effect for spatiotemporal coherence in the 300-550 ms window (i.e., event model mapping). This is consistent with masking limiting the time window for sensory and perceptual information accumulation.

Results in the parietal/occipital regions also converged with the component-based approach, as well as with the results from Experiment 2. Waveforms in the coherent and randomized sequences differed significantly at -75 milliseconds; however, this difference was not followed by at least 15 consecutive significant differences, and none of the Bayes factors were greater than 1.

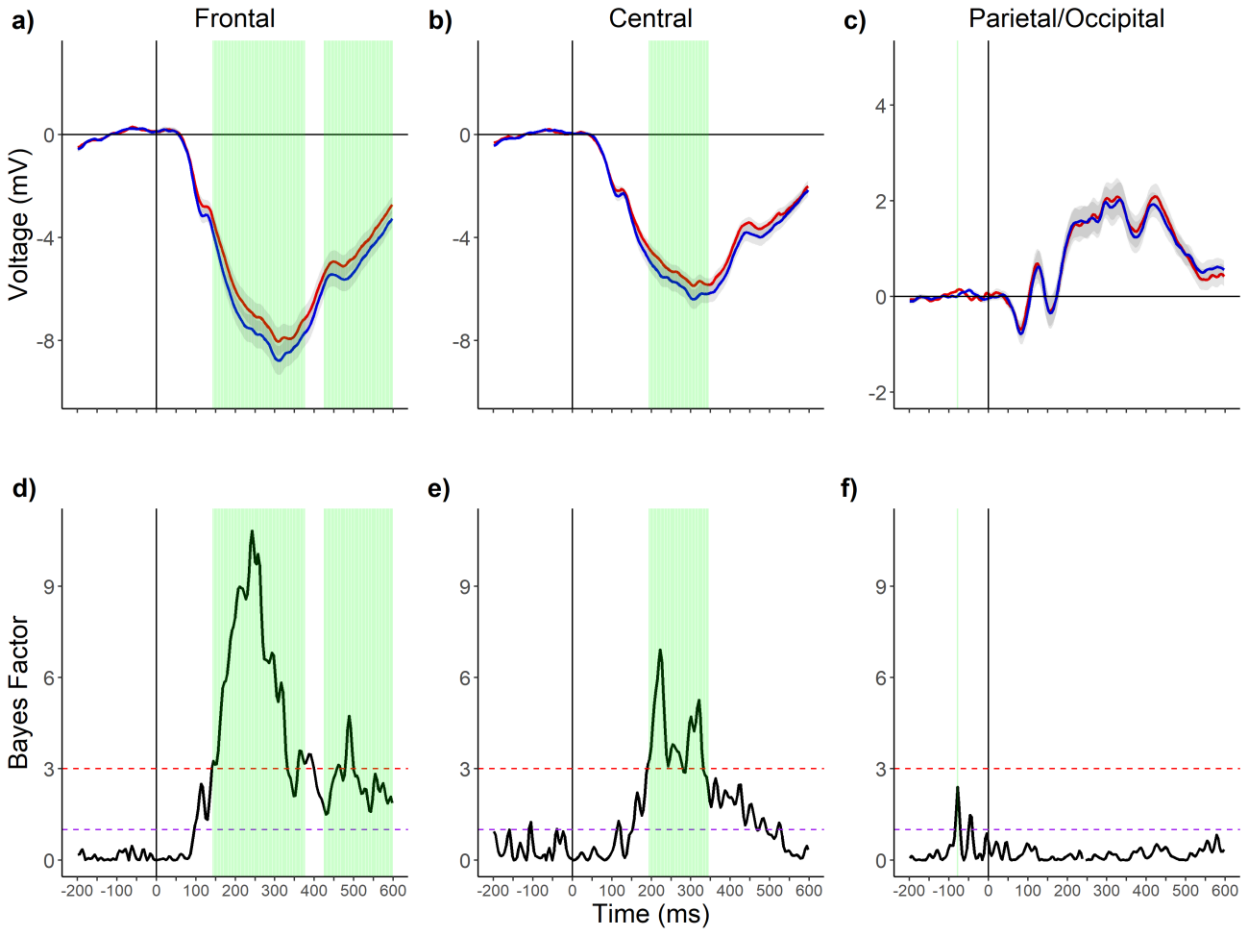


Figure 59. Exp 3: Grand average vERPs time locked to the onset of the scenes at time 0. Scenes were presented in either coherent or randomized sequences. Average waveforms at a) Frontal b) Central, and c) Parietal/Occipital sites are on the top row. Bayes factors for each of the paired sample t-tests within the epoch for d) Frontal e) Central, and f) Parietal/Occipital electrodes are provided in the bottom row. Green patches represent clusters of statistically significant comparisons. Red dashed lines in the Bayes factors plots d) through f) represent a Bayes Factor of 3 and purple lines represent a Bayes factor of 1 and -1, respectively.