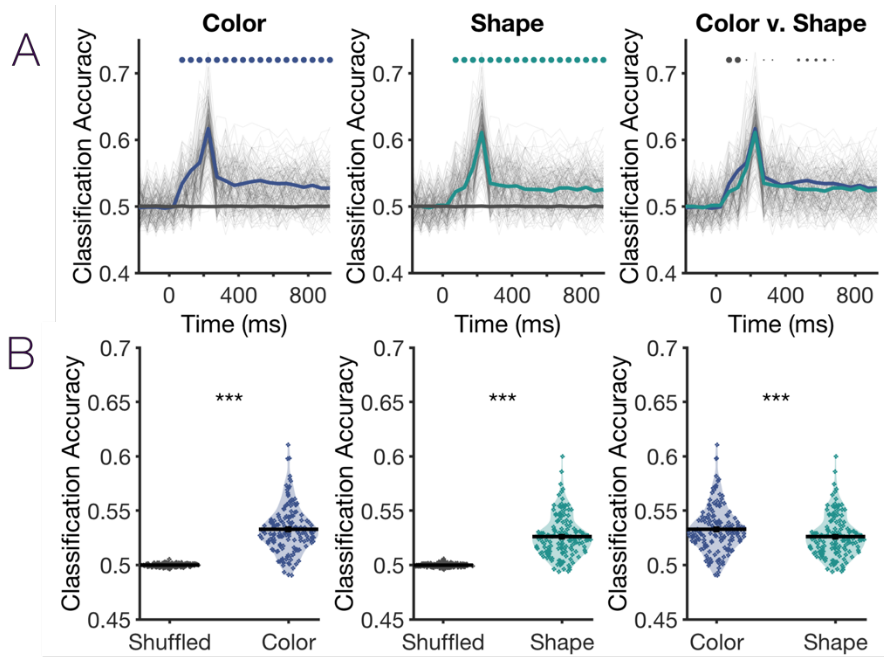

Figure S1. Classification separated by the color and shape conditions. Analyses in Figure 2 (Experiment 1) were performed on data combined across two conditions: remembering colors and remembering shapes. Here, we show that single-trial classification results are qualitatively similar when performed on color or shape data alone. (A) Time course of classification for color and shape conditions. Dots represent Bonferroni corrected significance for 23 time bins (small dots $p <.05$, medium $p < .01$, large $p < .001$). (B) Overall delay period classification. Both shape and color classification were robustly above chance (shuffled baseline), though it was slightly higher for the color condition compared to the shape condition.

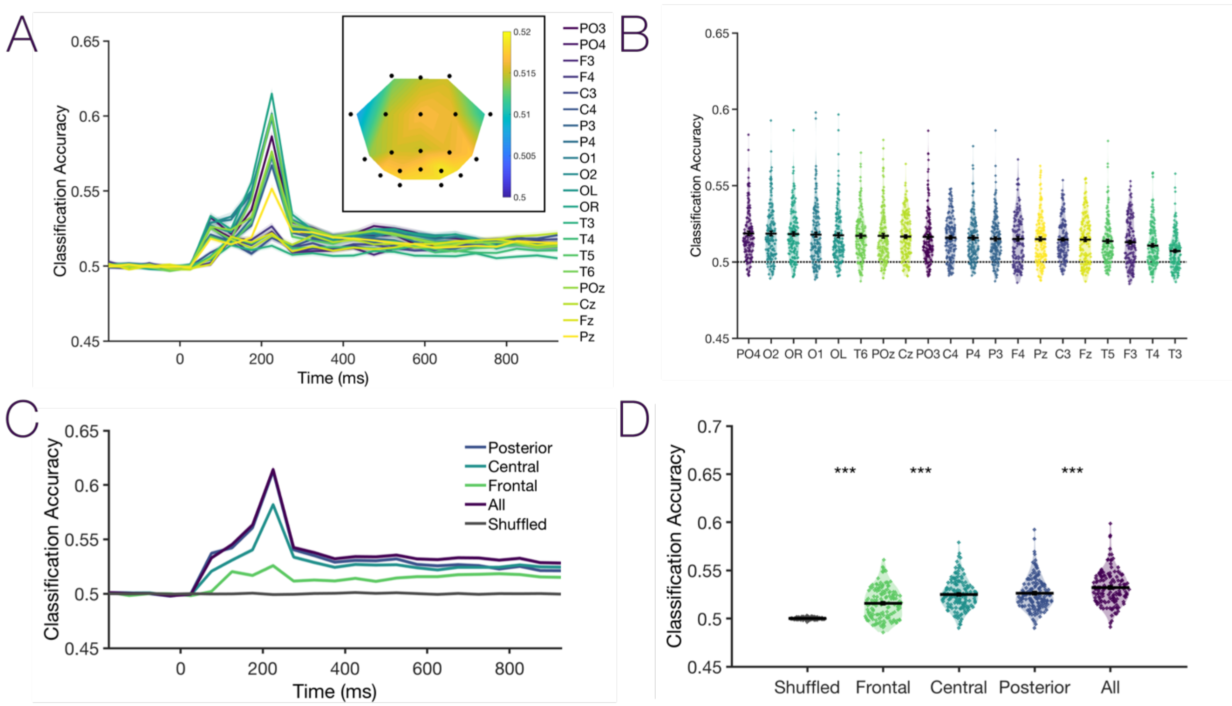

Figure S2. Classification for single electrodes and electrode groups. (A) We ran the single-trial classification analysis separately for each electrode (i.e., using only 1 predictor). The main figure shows classification accuracy over time for each electrode. The inset shows a topographic plot of average decoding accuracy during the delay period. (B) Average decoding accuracy during the delay period. All electrodes significantly outperformed a shuffled baseline ($p < .001$ Bonferroni-corrected for 20 channels). Electrodes are plotted from left to right in relative descending order of classification accuracy. (C). We likewise ran the single-trial classification analysis separately for broad groups of electrodes, including posterior (O1,O2,OL,OR,PO3,PO4,T5,T6), central (C3,Cz,C4,P3,Pz,P4,T3,T4), and frontal (F3,F4,Fz). (D) Average classification accuracy for groups of trials electrodes during the delay period. All groups significantly predicted set size, though frontal electrodes were had lower classification accuracy than central or posterior electrodes. However, no single group of electrodes achieved the same classification accuracy as the full analysis (all 20 electrode predictors)

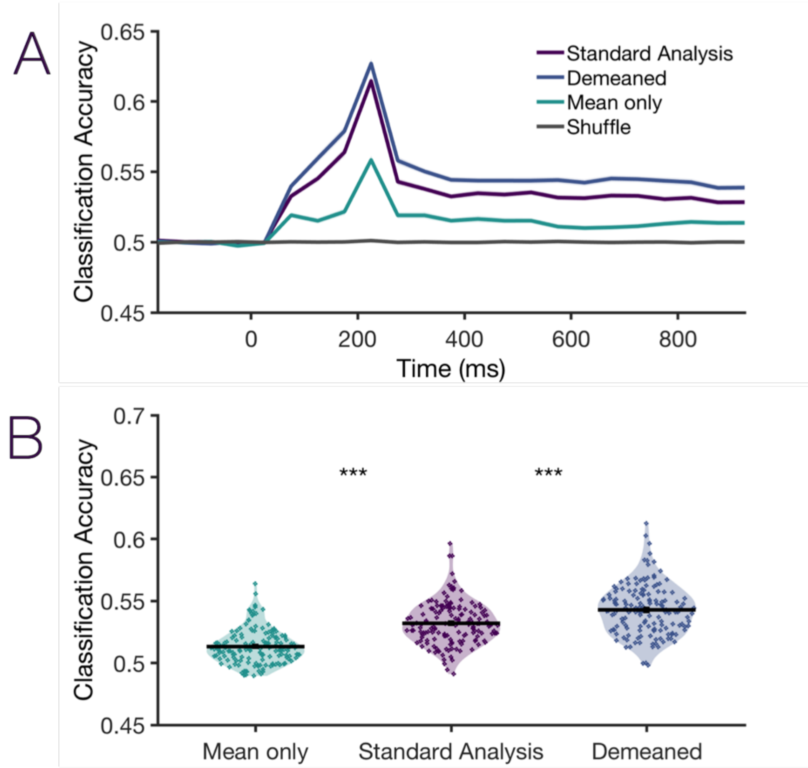

Figure S3. Classification using the global signal or the de-meaned pattern across electrodes. Given that even single electrodes (e.g., 1 predictor as in Figure 2) predicted working memory load, we wanted to test whether the multivariate aspect of this approach is truly beneficial or if, instead, the data could be equally well accounted for by a simple univariate signal change. To test this, we compared our standard analysis (raw voltage values from 20 electrode predictors) with demeaned values (i.e., take subtract the trial-wise mean from the 20 electrode predictors) and with the mean univariate signal alone (i.e., use the global univariate signal as our only predictor, quantified as 1 average voltage value for all electrodes). Overall, we found that the multivariate aspect of this signal (i.e., the pattern of voltage across electrodes) had more predictive power than a univariate signal alone. Although the univariate signal predicted working memory load significantly above chance ($p < .001$), it was significantly less predictive than the standard analysis ($p < .001$) which was in turn less effective than the trial-wise demeaned analysis ($p < .001$). (A) Time course of classification accuracy. (B) Violin plots of mean classification accuracy during the delay period.

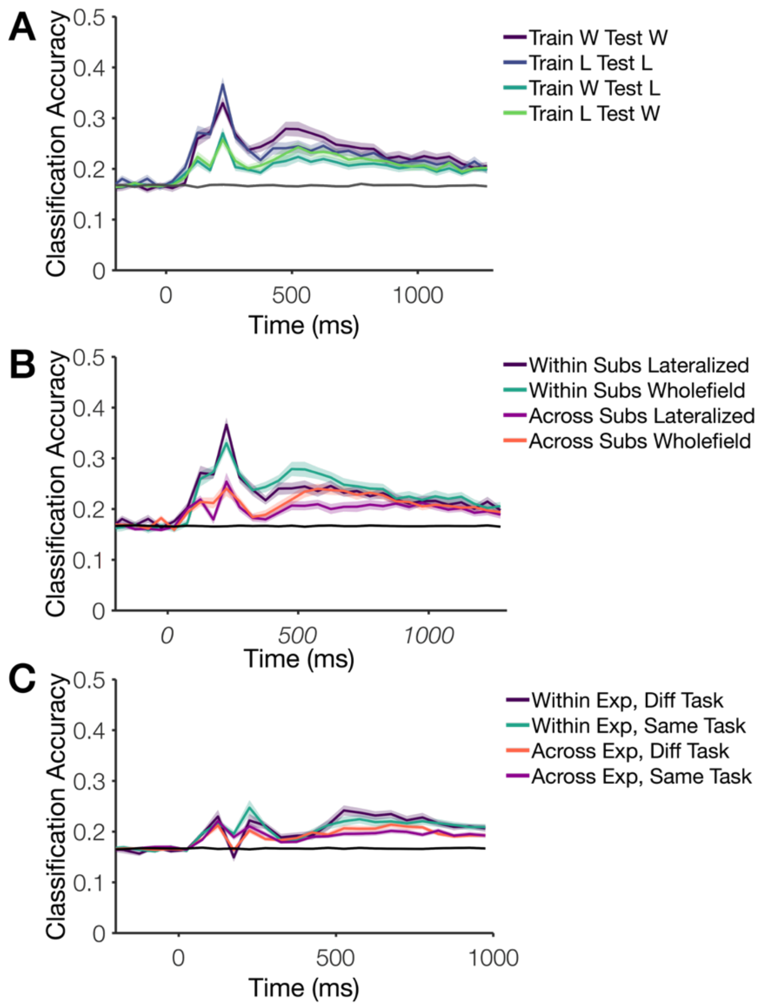

Figure S4. All combinations for cross-task, -subject, and -experiment analyses. (A) Training within and across tasks (within subjects) in Experiment 2A. W is short for the wholefield task, and L is short for the lateralized task. In the main text, the line "within task" is the average of "Train L Test L" and "Train W Test W" lines; the "across tasks" line is the average of the "Train L Test W" and the "Train W Test L" lines. (B) Training within and across subjects in Experiment 2A. We performed the within and across subjects analysis separately for the lateralized and wholefield tasks. In the main text, the line "Within Subs" is the average of the lines "Within Subs Lateralized" and "Within Subs Wholefield" in this figure; the line "Across Subs" is the average of the lines "Across Subs Wholefield" and "Across Subs Lateralized" in this figure. (C) Training within and across experiments (2A and 2B). We performed the within and across experiments analysis separately while also training/testing within or across tasks. In the main text, the line "Within Exp" is the average of the lines "Within Exp, Diff Task" and "Within Exp, Same Task" in this figure; the line "Across Exps." is the average of the lines "Across Exp, Same Task" and "Across Exp, Diff Task" in this figure.

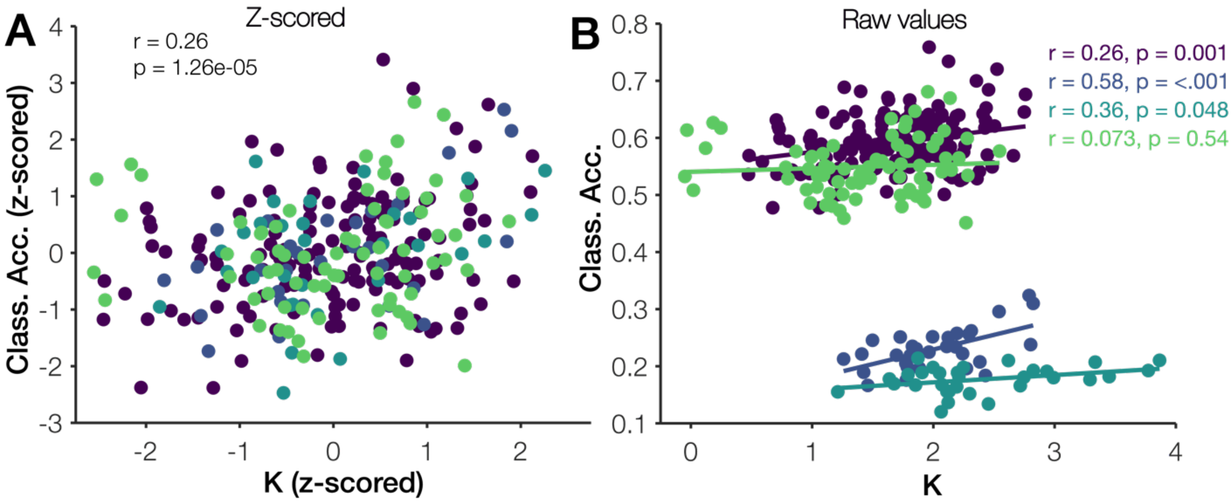

Figure S5. Individual differences for all unique subjects (no subjects excluded for poor performance, > 2 S.D.'s below group average). (A) Combined correlation using z-scored behavior and classification accuracy data. (B) Individual experiment correlations, including all unique subjects.