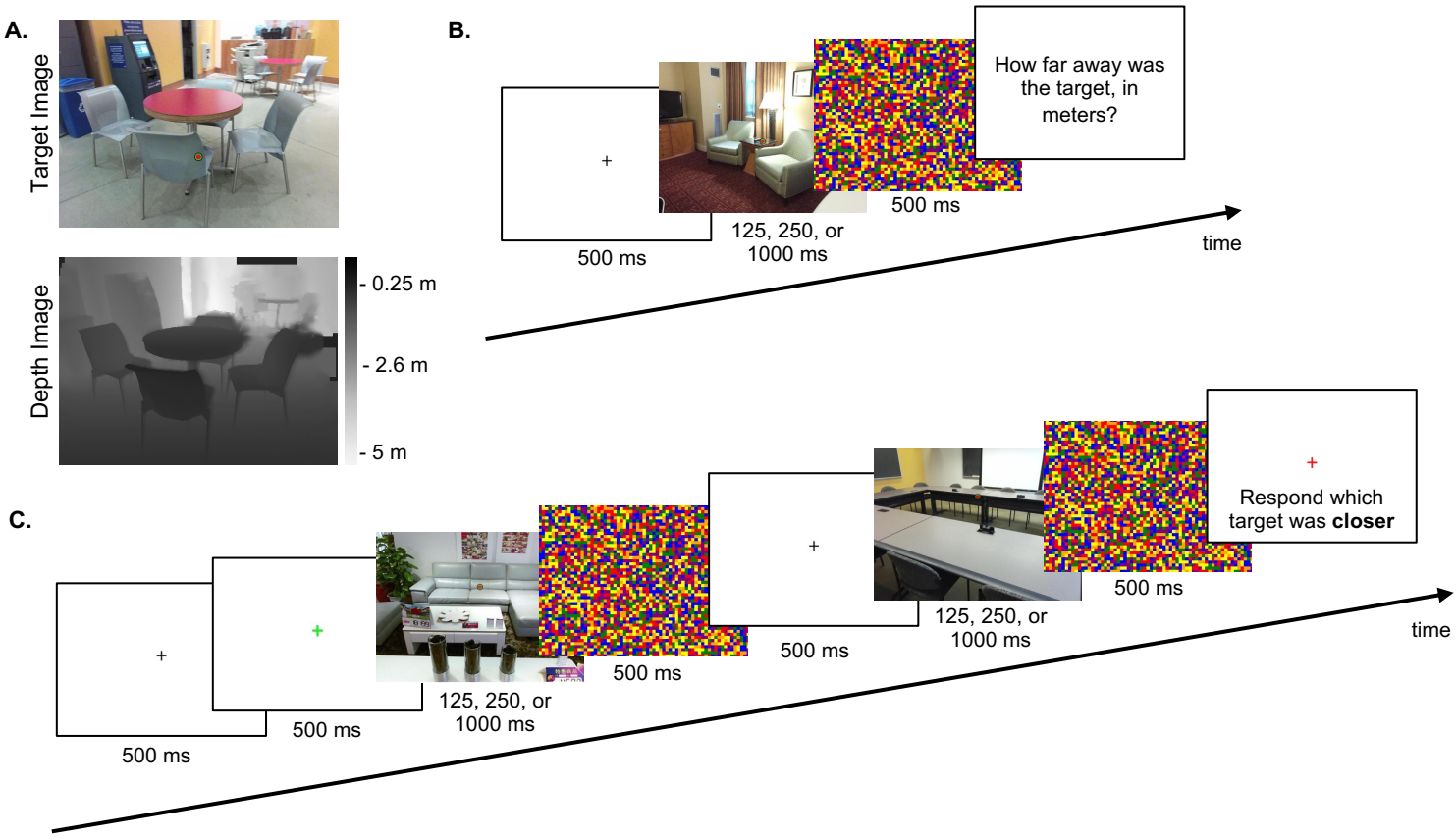

Figure 1. Behavioral Paradigms. A) Sample target image and corresponding SUN-RGBD depth map. B) Verbal estimate paradigm. Scenes with embedded targets were presented for a variable duration followed by a mask. C) Distance discrimination paradigm. Participants were presented with two scenes for the same duration and responded which target was closer.

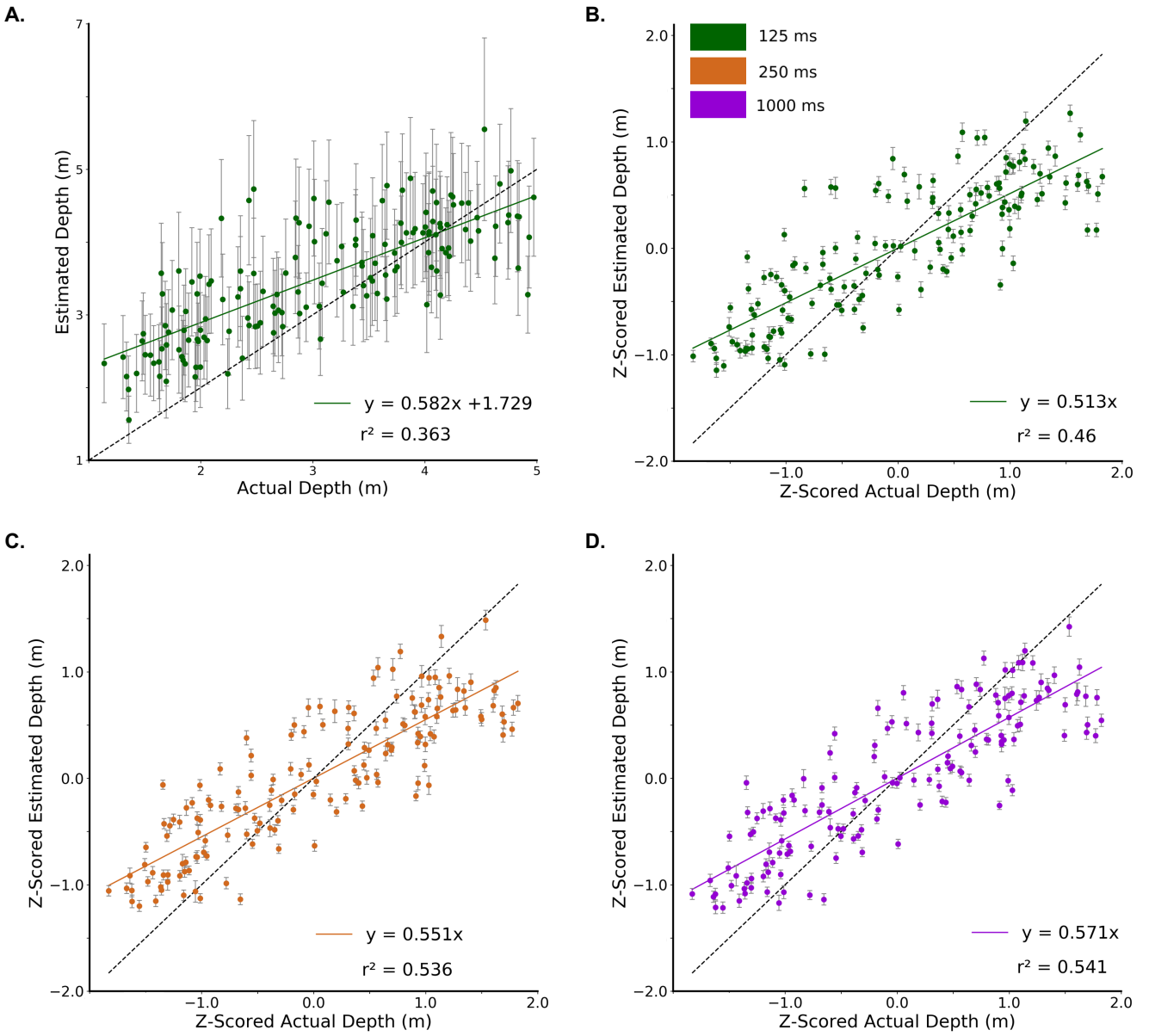

Figure 2. Verbal Estimate Linear Models. A) Raw data from the verbal estimate task for the 125 ms viewing duration condition. Participants are independent between viewing duration conditions and all error bars in plots are between subjects standard error. B) Z-scored data at 125ms. C) Z-scored data at 250ms. D) Z-scored data at 1000ms.

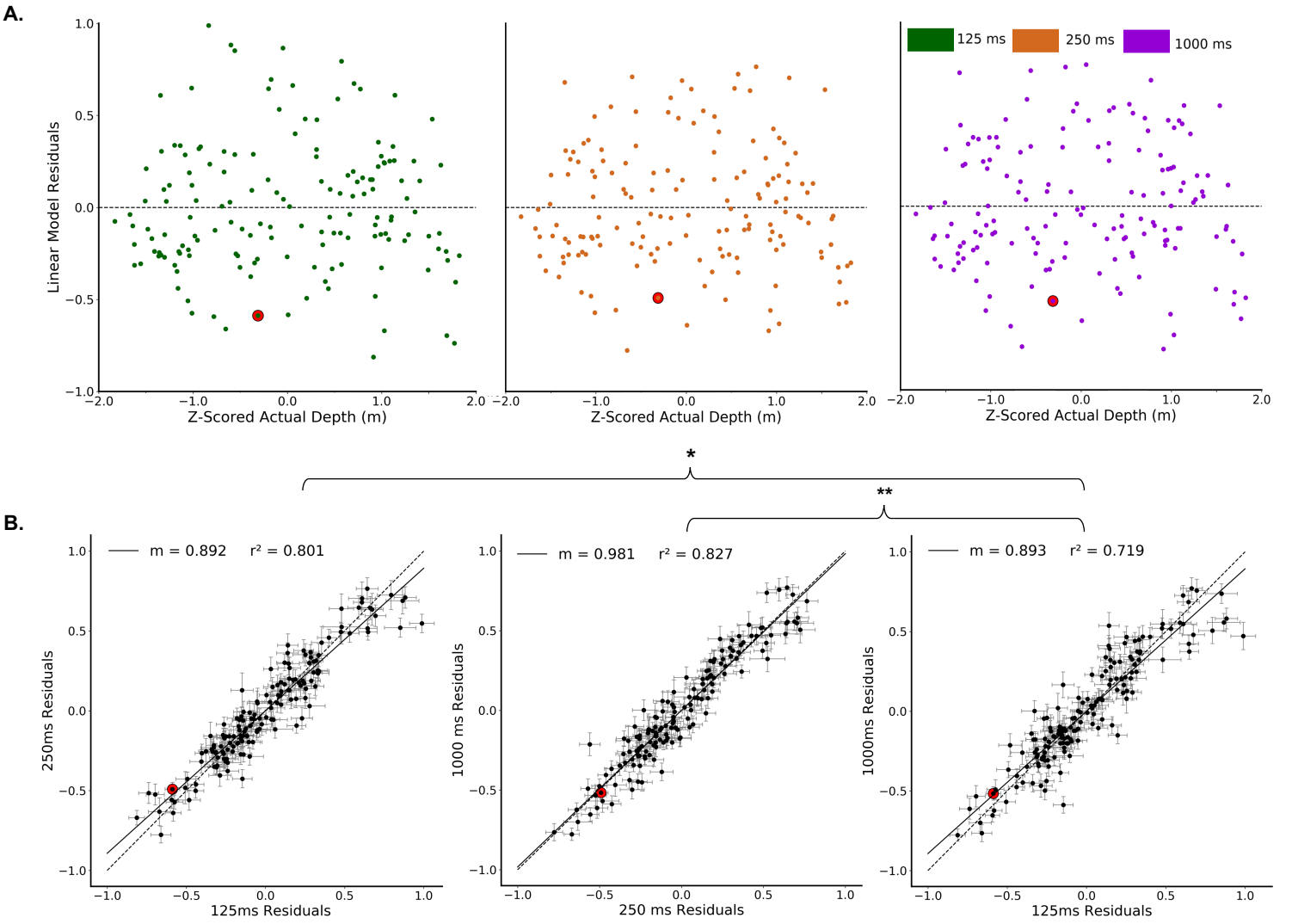

Figure 3. Correlation between verbal estimate regression residuals in Figure 2. A) Verbal estimate residuals (difference between the model prediction and the average estimate for that stimulus) at each viewing duration plotted against z-scored actual depth. The residual for a particular stimulus at each viewing duration is highlighted in red. B) Correlation between all combination of verbal estimate regression residuals at the three viewing durations (125, 250, 1000ms). The residual correlation for a particular stimulus is highlighted in red on all plots. (* indicates $p < 0.05$, ** indicates $p < 0.01$)

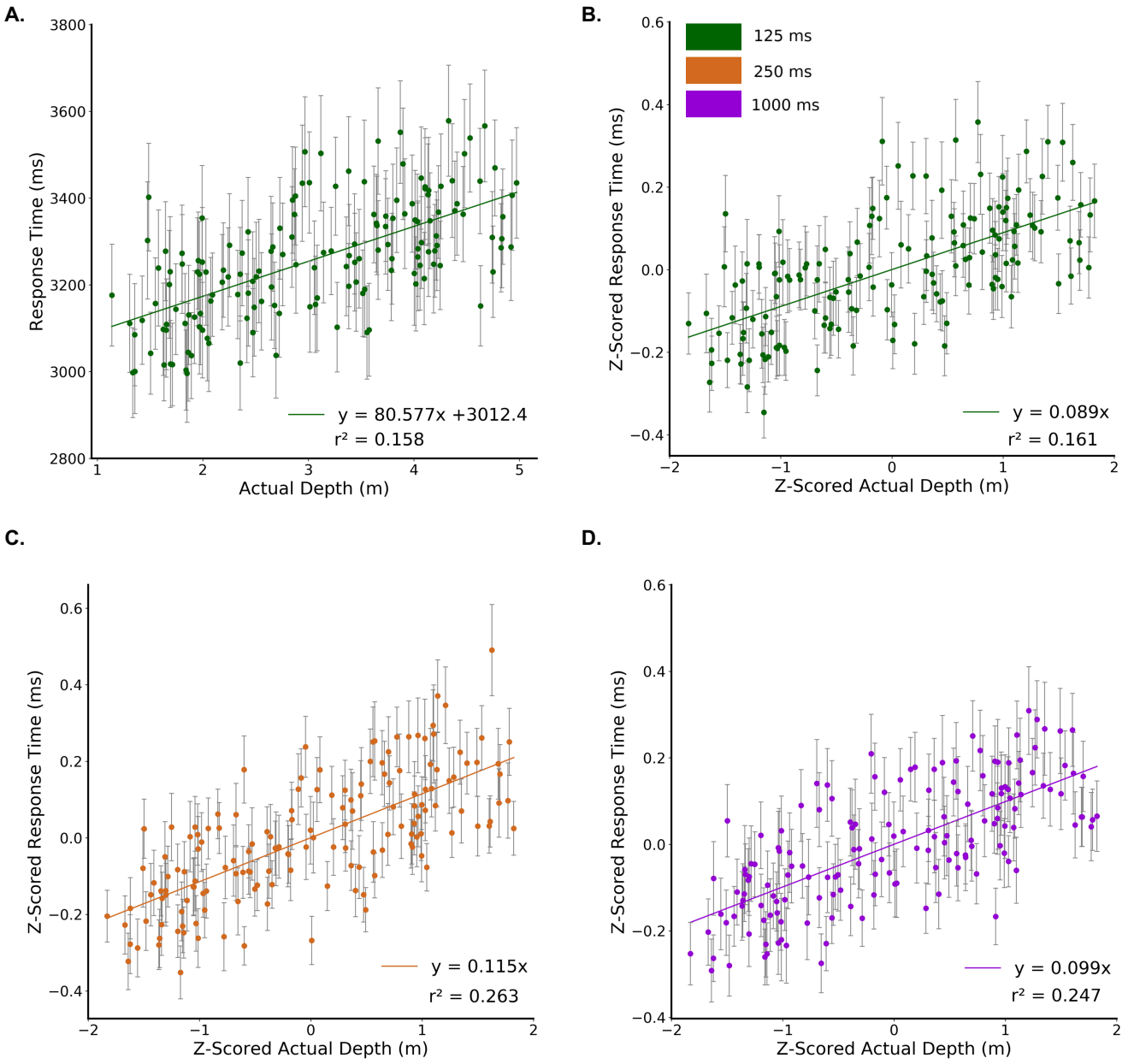

Figure 4. Verbal Estimate Response Time Linear Models. A) Raw response time from the verbal estimate task for the 125 ms viewing duration condition. Participants are independent between viewing duration conditions and all error bars in plots are between subjects standard error. C) Z-scored response time at 125ms. B) Z-scored response time at 250ms. D) Z-scored response time at 1000ms.

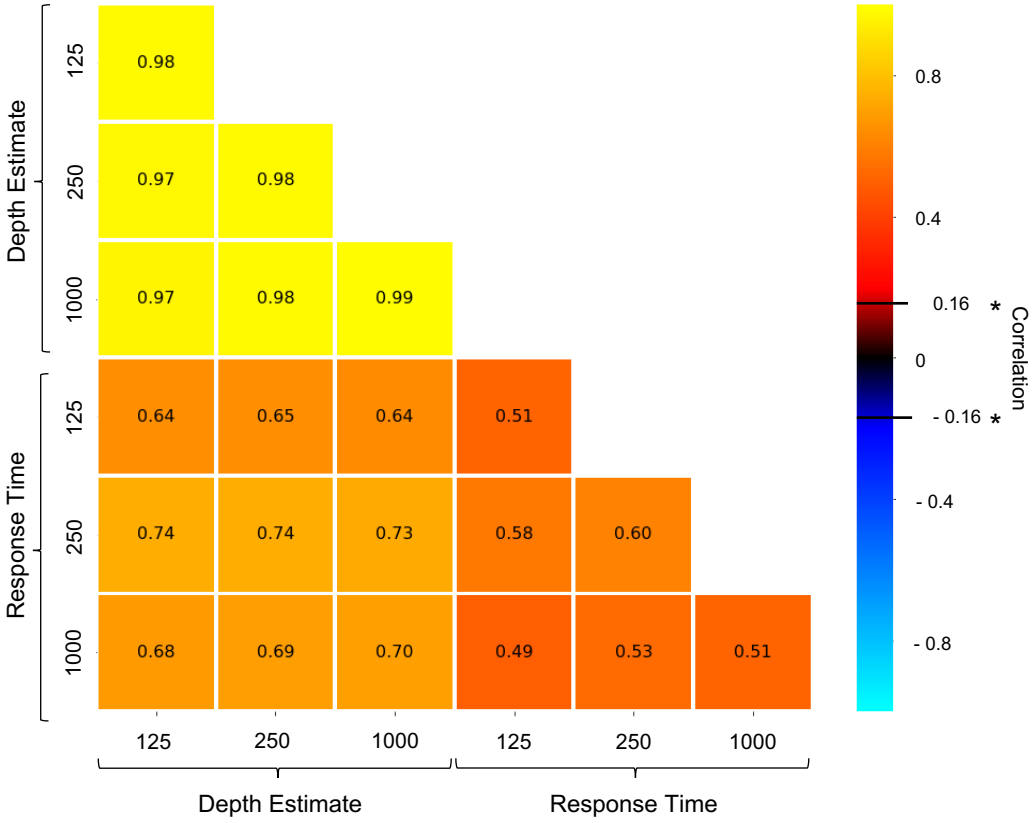

Figure 5. Verbal Estimate Average Correlation Matrix. Data were randomly split in half and correlated against itself. This was repeated 10,000 times resulting in this average correlation matrix. Given a sample size of 156 images and a significance level of 0.05, r >= 0.16 are significant. Within task outcome correlations are significant within and between viewing durations, demonstrating strong reliability between independent groups of participants. Significant correlations between task outcome (response time, depth estimate) and viewing durations were also observed, indicating stimulus-level generalization.

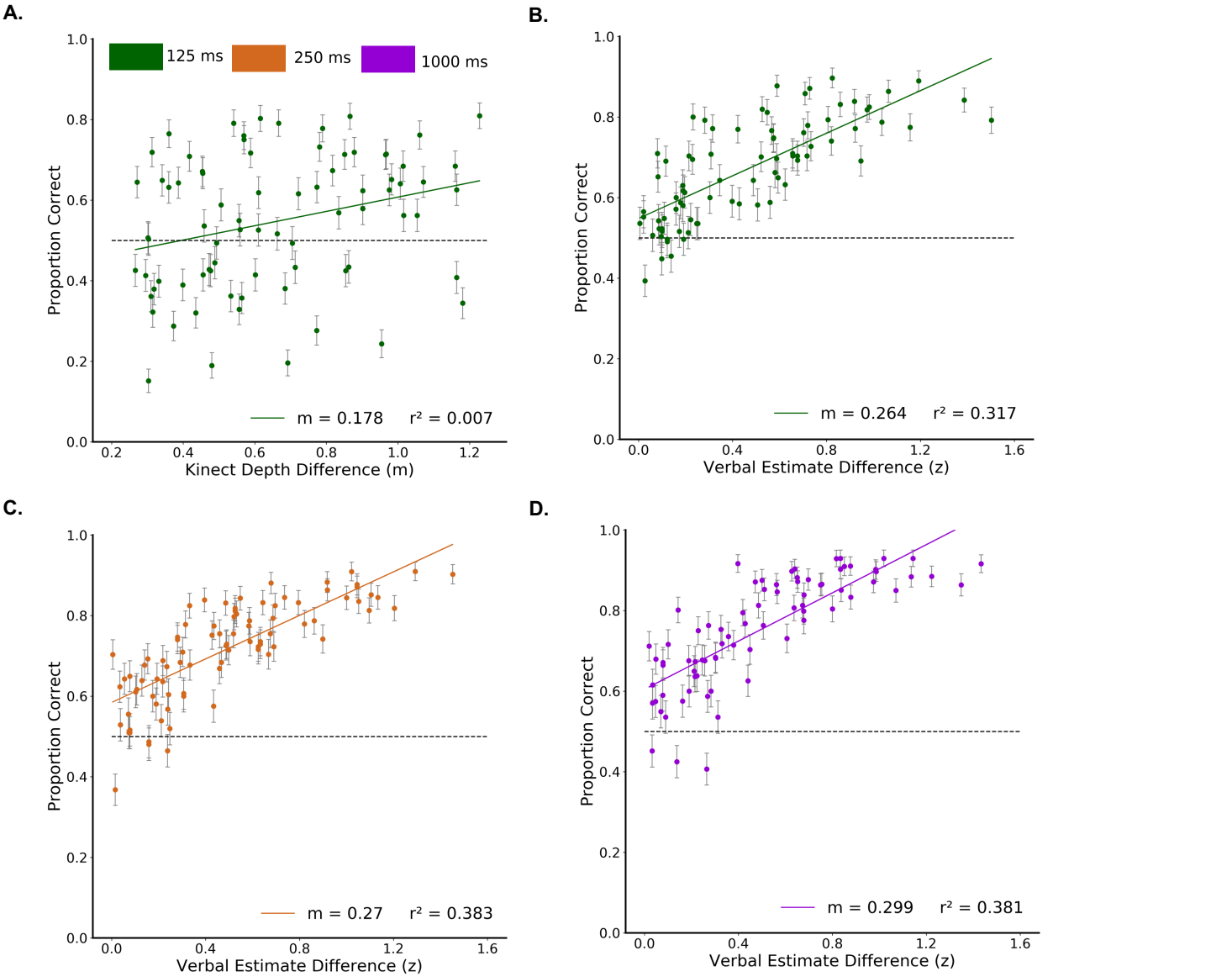

Figure 6. Discrimination Accuracy Results. A) Linear model of discrimination accuracy as a function of the Kinect depth difference between targets for the 125ms viewing duration condition. B) Linear models of discrimination accuracy as a function of verbal estimates from independent participants at 125ms. The verbal estimate difference is calculated as the absolute difference between verbal estimates for images within a discrimination trial divided by their joint variance, effectively resulting in a z-score of the verbal estimate difference. The verbal estimate and discrimination task were comprised of independent groups of participants. C) Accuracy as a function of verbal estimates at 250ms D) Accuracy as a function of verbal estimates at 1000ms D) Discrimination accuracy as a function of the Kinect (bars with colored outline) and as a function of differences in verbal estimates from independent participants (* indicates $p<0.05$; ** $p<0.005$; *** $p<0.0005$)

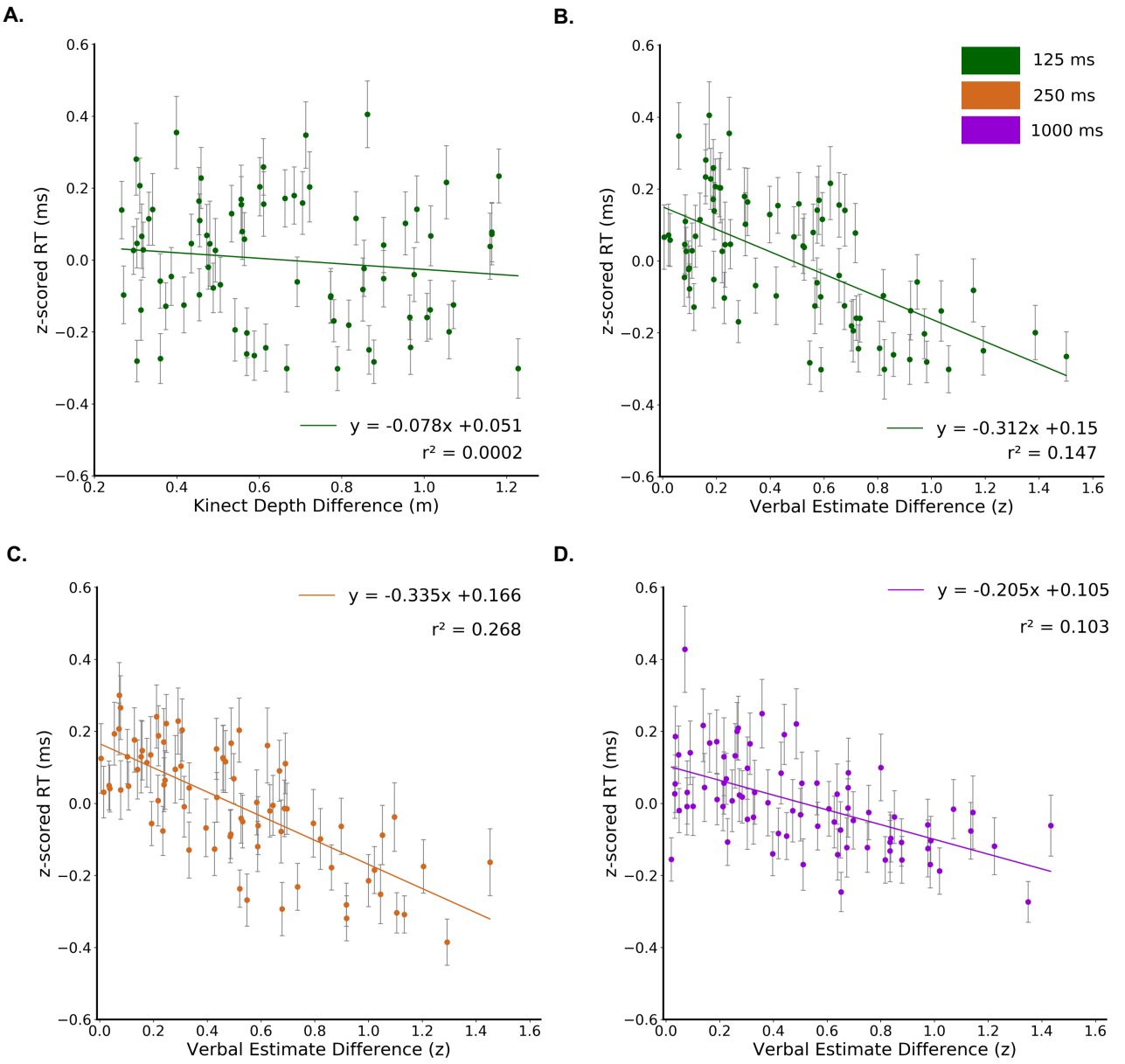

Figure 7. Discrimination Response Time Results. A) Linear model of discrimination response time as a function of the Kinect depth difference between targets for the 125ms viewing duration condition. B) Linear model of discrimination response time as a function of verbal estimates from independent participants at 125ms. The verbal estimate difference is calculated as the absolute difference between verbal estimates for images within a discrimination trial divided by their joint variance, effectively resulting in a z-score of the verbal estimate difference. The verbal estimate and discrimination task were comprised of independent groups of participants. C) 250ms and D) 1000ms

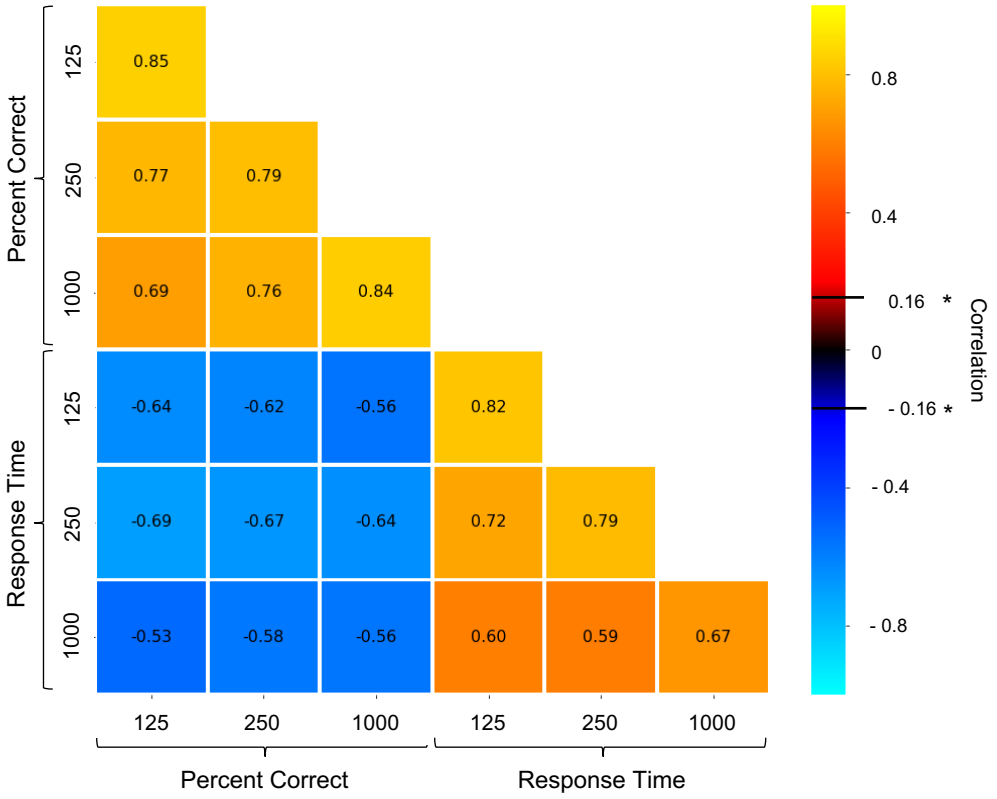

Figure 8. Discrimination Average Correlation Matrix. Data were randomly split in half and correlated against itself. This was repeated 10,000 times resulting in this average correlation matrix. Given a sample size of 156 images and a significance level of 0.05, r >= 0.16 are significant. Within task outcome correlations are significant within and between viewing durations, demonstrating strong reliability between independent groups of participants. Significant correlations between task outcome (response time, percent correct) and viewing durations were also observed, indicating stimulus-level generalization.

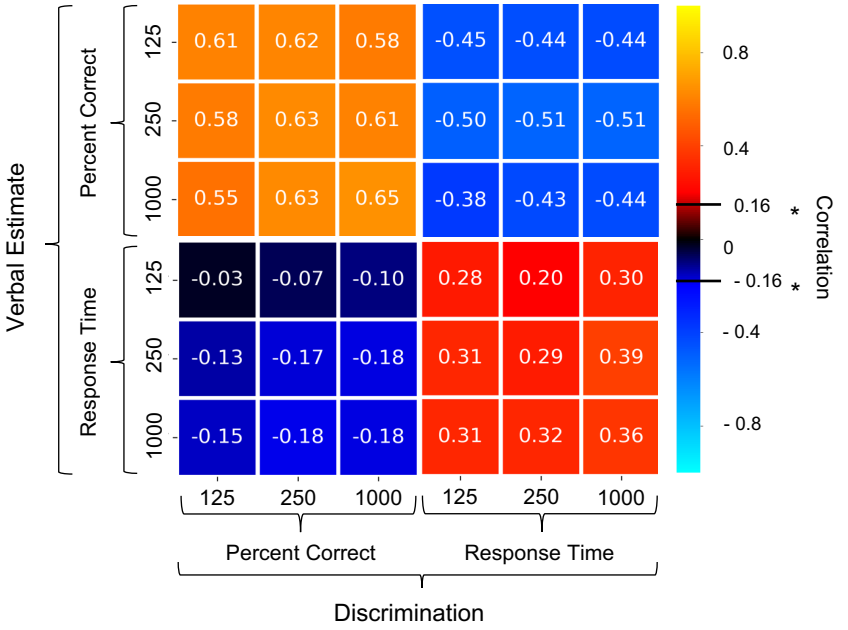

Figure 9. Average Correlation Matrix (Task x Outcome Measure x Viewing Duration). Data were randomly split in half and correlated against itself. This was repeated 10,000 times resulting in this average correlation matrix. Given a sample size of 156 images and a significance level of 0.05, r >= 0.16 are significant. This analysis was conducted in the 'discrimination space', meaning that comparisons between tasks were made in the context of a discrimination trial. Between task correlations are presented here. Significant correlations were observed in accuracy and response time between tasks (discrimination, verbal estimate) within and between viewing durations. Impressively, significant correlations were also found between tasks, outcome measure, and viewing duration (top right 3x3 matrix). These results indicate systematic errors and biases for individual stimuli between independent groups of participants.

A.

B.

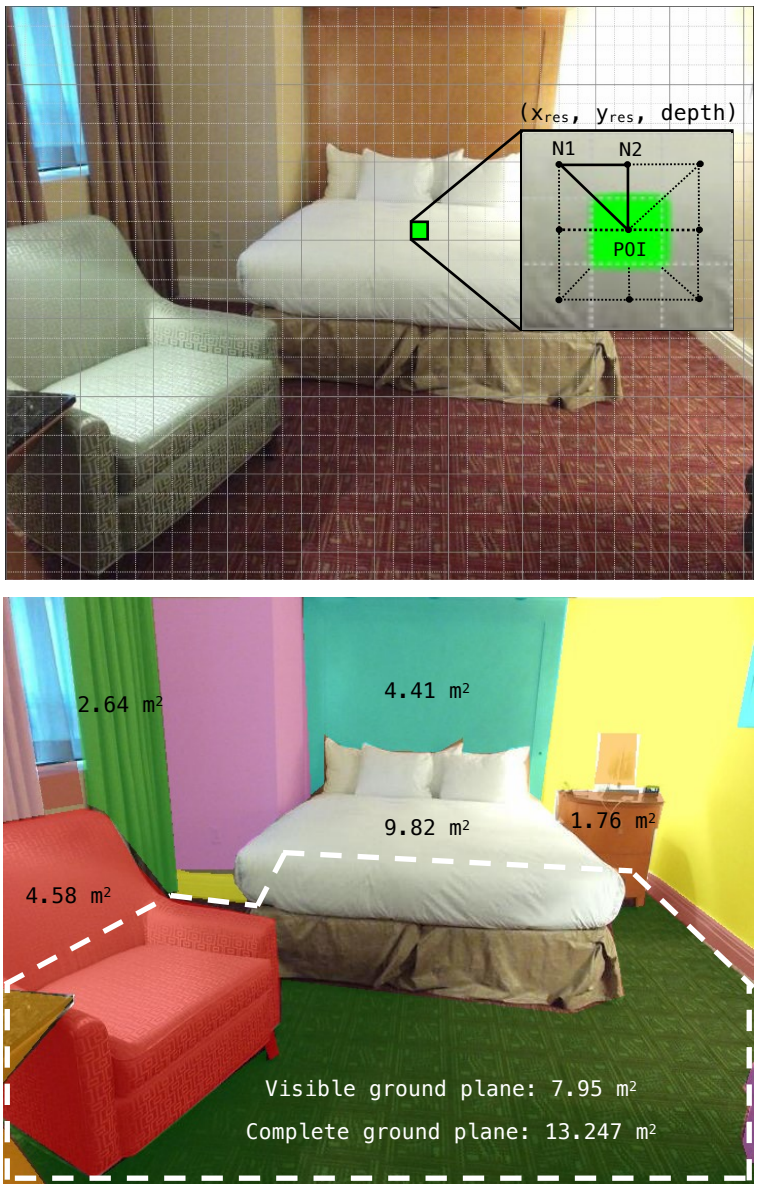

Figure 10. Pixelwise derivation of real-world object surface area. A) Surface area for each pixel with the image (730x530 pixels) is calculated by first deriving its coordinates in 3D space as (x_{res}, y_{res}, depth) where x_{res} is the real-world width of the pixel and y_{res} is the real-world height of the pixel. Euclidean distance between the pixel of interest (POI) and its neighbors is calculated. The area of the triangles defined by the POI and its neighbors is subsequently calculated and summated across all pixels within an object. B) Sample output of the surface area algorithm. The polygon that subtends the complete ground plane (GP) is estimated using scene boundaries. The surface area of the visible ground plane (VGP) is calculated using the method described in A). Accordingly, the surface area of GP is equal to the proportion of pixels belonging to the visible vs. complete ground plane multiplied by the surface area of the VGP.

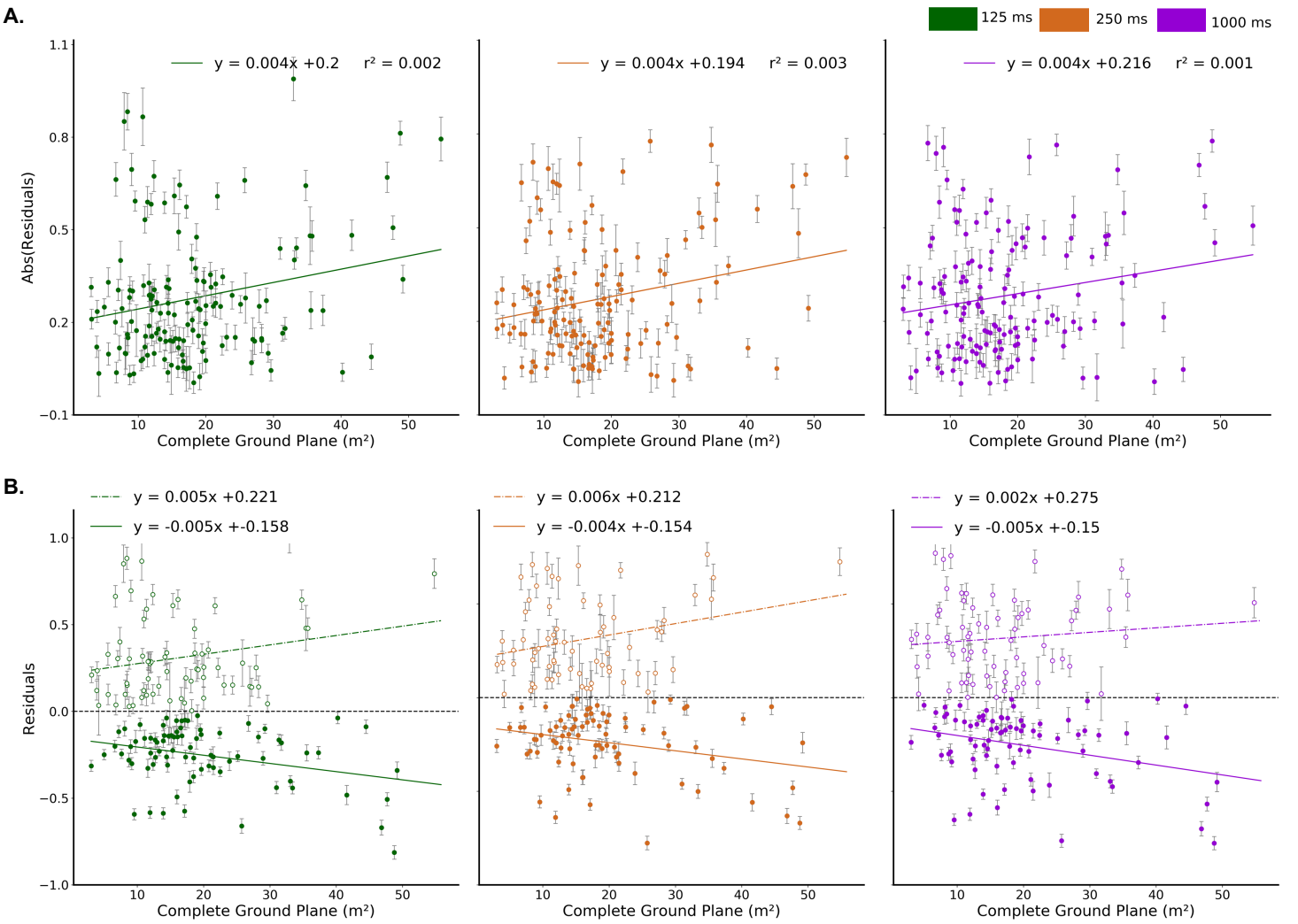

Figure 11. Complete Ground Plane Linear Models. A) Absolute value of the residuals from the verbal estimate linear models against the complete ground plane surface area for each stimulus (plotted for each viewing duration - 125, 250, and 1000 ms). B) Directional residual linear models (i.e., positive and negative residual models) per viewing duration reveal a diverging effect within the residuals wherein the positive residuals increase, and the negative residuals decrease with larger complete ground plane values.