Supplemental Materials

**Figure S-1**: Chart detailing subject exclusions

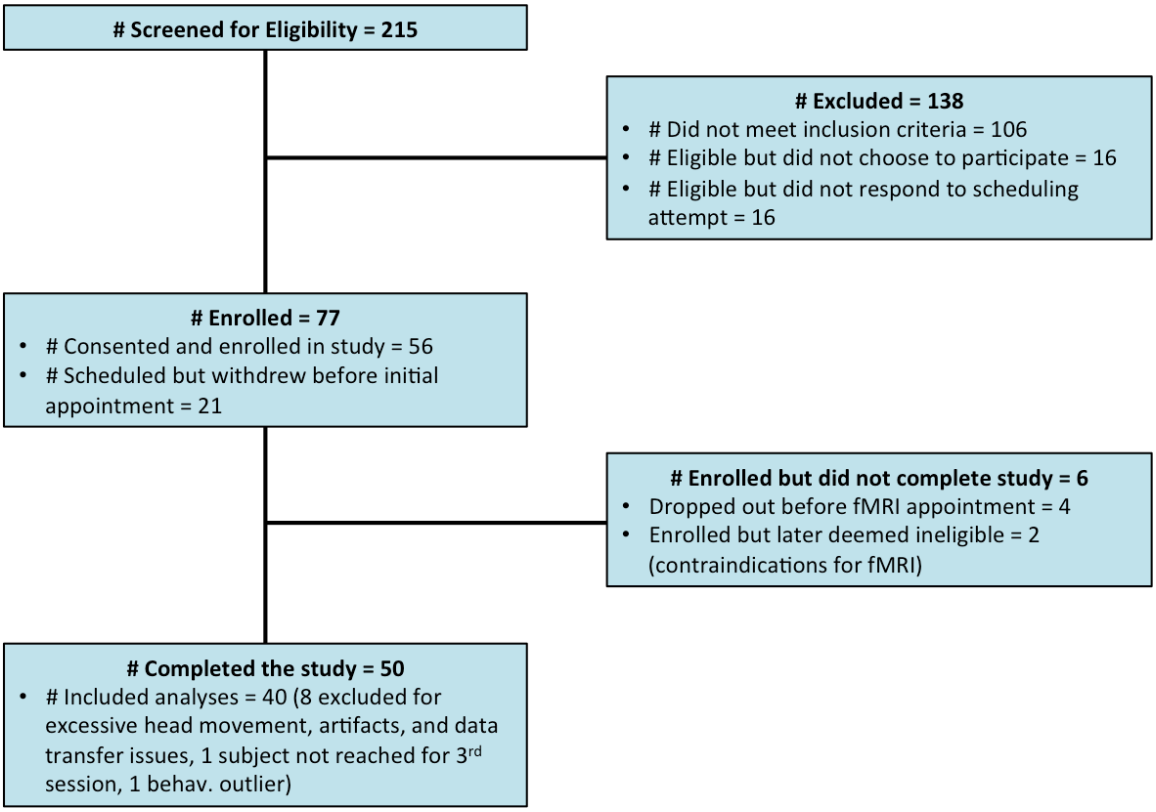


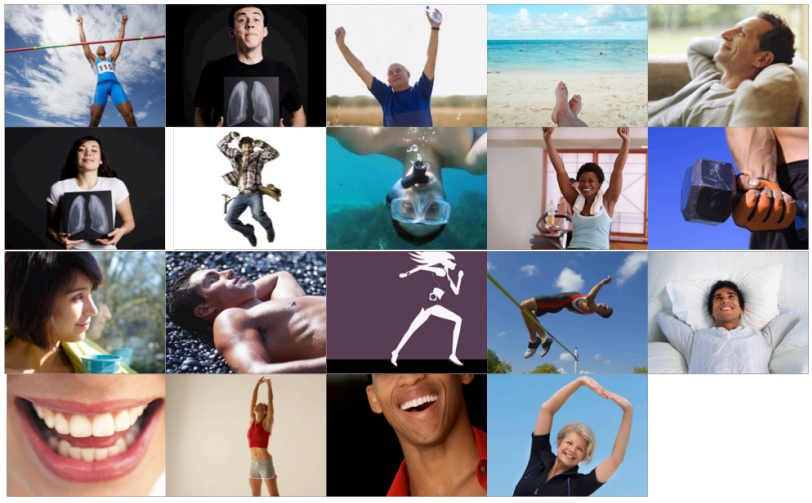

**Figure S-2:** All stimuli used in the experiment are displayed below. During the experiment each images was presented with the text above the image "Stop Smoking. Start Living."

**Positive / Social**

**Negative / Social**

**Negative / Non-Social**

**Positive / Non-Social**



**Supplementary Analysis S-1 -** Subjective "Pleasantness" ratings as valence

After the scan, subjects rated the pleasantness of each image. To complement our "category" model of valence, we also created a more continuous model of valence by using these subject-specific pleasantness ratings as models in place of the binary valence model.

Specifically, we created a valence model for each subject individual by calculating pairwise Euclidean distances between the pleasantness ratings of each image. These models were correlated with univariate and multivariate brain RDMs and entered into the regression models.

Our results showed that using the pleasantness ratings produced consistent results with the main model (Equation 1): For the univariate regression, this pleasantness variable replicated the valence effect by showing a significant negative correlation with behavior change. For the multivariate regression, only social information showed a positive relationship with behavior change, again replicating the original results:

| **Equation 1: $\triangle B = \beta_1 V + \beta_2 S + \beta_3 H + \varepsilon$** | | | | |
|---|---|---|---|---|
| *(N=40)* | **β** | **S.E.M.** | **t** | **P-value** |
| *Using univariate RDMs* | | | | |
| $\beta_1 V$ (pleasant) | -2.27 | 1.04 | -2.19 | 0.035* |
| $\beta_2 S$ (social consequences) | -5.57 | 3.23 | -1.73 | 0.093 |
| $\beta_3 H$ (health consequences) | -4.73 | 2.76 | -1.72 | 0.094 |
| | | | | |
| *Using multivariate RDMs* | | | | |
| $\beta_1 V$ (pleasant) | -2.40 | 1.95 | -1.23 | 0.228 |
| $\beta_2 S$ (social consequences) | 5.59 | 2.18 | 2.56 | 0.015* |
| $\beta_3 H$ (health consequences) | 2.76 | 2.06 | 1.34 | 0.187 |

These results suggest that our findings are robust across these classification methods (i.e., classification of images by independent raters, as originally reported, or by the subjects' own classifications).

**Supplementary Analysis S-2 -** "This image gives me thoughts about quitting" ratings

Subjects were asked to what extent each image gave them thoughts about quitting. We created subject-specific RDMs out of the Euclidean between pairwise ratings, and then correlated these with subject-specific brain RDMs. Finally, we added these values as a fourth regressor to our main effects model (Equation 1). Here, we've shown that adding the correlation of this subject-specific RDM to the regression does not change the significance of the other neural predictors, but itself shows a negative relationship with smoking behavior, suggesting that multivariate representations of "quitting" are correlated with decreased smoking, suggesting that the degree to which self-MPFC represented motivations to quit, the more smokers reduced their smoking.

| Equation: $\Delta B = \beta_1 V + \beta_2 S + \beta_3 H + \beta 4Q + \varepsilon$ | | | | | |
|---|---|---|---|---|---|
| *(N=40)* | **β** | **Std. Error** | **t-value** | **95% C.I.** | **P-value** |
| *Using univariate RDMs* | | | | | |
| $\beta_1 V$ (valence) | -9.56 | 3.94 | -2.43 | -17.56, -1.56 | 0.021* |
| $\beta_2 S$ (social consequences) | -5.78 | 3.22 | -1.8 | -12.32, 0.75 | 0.081 |
| $\beta_3 H$ (health consequences) | -4.4 | 2.73 | -1.61 | -9.93, 1.14 | 0.116 |
| $\beta_4 Q$ (thoughts about quitting) | -0.24 | 1.48 | -0.16 | -3.25, 2.78 | 0.875 |
| | | | | | |
| *Using multivariate RDMs* | | | | | |
| $\beta_1 V$ (valence) | 5.93 | 3.25 | 1.82 | -0.67, 12.54 | 0.077 |
| $\beta_2 S$ (social consequences) | 6.22 | 2.08 | 3.00 | 2.00, 10.43 | 0.005** |
| $\beta_3 H$ (health consequences) | 3.13 | 1.92 | 1.63 | -0.77, 7.03 | 0.112 |
| $\beta_4 Q$ (thoughts about quitting) | -3.21 | 1.53 | -2.10 | -6.31, -0.11 | 0.043* |

**Supplementary Analysis S-3 -** Group-level content representations in self-MPFC

Using all subjects, group-averaged neural multivariate and univariate RDMs were calculated. Spearman correlations were then run to compare these group-average neural RDMs with the valence, social, and health models respectively.

To determine p-values, RDM item labels were shuffled and correlated to each model RDM across 10,000 iterations, creating a null distribution of Spearman's *r* values. Two-tailed p-values were derived from the location of the true correlation score with this distribution. No type of message content showed group-level representation in MPFC.

| | Spearman's *r* | p-value |
|---|---|---|
| *Univariate RDM* | | |
| valence model RDM | -0.010 | 0.084 |
| social model RDM | -0.010 | 0.848 |
| health model RDM | -0.018 | 0.848 |
| | | |
| *Multivariate RDM* | | |
| valence model RDM | -0.038 | 0.535 |
| social model RDM | 0.009 | 0.901 |
| health model RDM | -0.010 | 0.834 |