Rewards are categories.

Erik J. Peterson
Dept. of Psychology
Colorado State University
Fort Collins, CO

Chapter 3 – fMRI analyses

An acquisition

Data Details. fMRI data was acquired at the Intermountain Neuroimaging Consortium (INC) facility located at the University of Colorado at Boulder on a Siemens Allegra 3T (whole body) scanner. All 18 right-handed participants were pre-screened for the typical fMRI exclusion factors (e.g. metal implants, mental disorders, etc). High resolution anatomical data were acquired as a T1-weighted structural image, MPRAGE sequence, at 1x1x1 mm, (256x156x192) with a TR of 2530 ms, and TE of 1.64 ms, with a flip angle of 7°. All functional (i.e. BOLD) data was acquired with T2-weighted echo-planar imaging (EPI), at 2.29 x 2.29 x 4.00 mm (96 x 96 x 26), with a TR of 1500 ms, a TE a 25 ms, a flip angle of 75° and a FOV of 220 mm.

A total of 4 sets of functional data were acquired. The first was of the "refresher" for part 1 of the behavioral training (p**??**), spanning 241 volumes. The second and third covered part 2 of the stimulus-responses learning task (again see p**??**), which was divided up into 2 (nearly) even sets so that participants need not be active for more the 10 or so minutes. Thes sets lasted 390 and 394 volumes respectively. The fourth acquisition covered a scan, that featured repeated examples from both reward categories in a random order. The intent of this scan was to isolate rewarding activity outside the primary task. This localizer was not in the end useful (see p8).

Preprocessed (model) food. Following DICOM to nifiti-1 conversion using dicom2nii (`http://www.mccauslandcenter.sc.edu/mricro/mricron/dcm2nii.html`), each dataset was subjected to the following preprocessing pipeline carried out in SPM8's batch mode (`http://www.fil.ion.ucl.ac.uk/spm/software/spm8/`). For complete code see, `https://github.com/andsoandso/fmri/tree/master/catreward/spm_m`. Anatomical data was first segmented into white and grey matter regions (?, ?). Based on these segments, the parameters necessary for normalization into T1 MNI-352 (1 mm) space were calculated. Normalization has two steps. The first is a Bayesian 12-parameter affine transformation (?, ?). The second is a set of nonlinear deformations, using a 1127 parameter discrete cosine transform (?, ?). Anatomical data was then resampled from 1.27 to 1.00 mm^3 using fourth degree β-splines and finally, using the parameters above, normalized into MNI space.

First movement regressors for all volumes of the functional data were calculated (?, ?). No participant moved more than 1.5 mm. Functional data was then slice-time corrected, using slice 13 (the middle slice from the descending acquisition) as the reference, followed by co-registeration with the pre-processed (native-space) anatomical data, and resampling into 3 mm^3 voxels again using fourth degree β-splines (?, ?). Functional data was then normalized into MNI space using the anatomically-derived parameters above. Finally, the functional data was spatially smoothed using a 6 mm FWHM Gaussian, though a copy of the unsmoothed data was retained for the ROI analyses (described on p8). Each voxel's time course was also low-pass filtered using finite impulse response model, with a cutoff at 0.008 Hz, prior to regression analysis (?, ?). For all whole-brain analyses, the movement regressors were entered into as covariates thus accounting for any head movement. Given the large spatial averages needed for the ROI analyses these analyses weren't motion corrected.

The best of all possible signals. In fMRI and in time-series analysis in general there is an intrinsic trade-off between detecting a signal in the presence of noise and estimating the shape of that signal (Dale, 1999; Birn, Cox, & Bandettini, 2002; Liu, 2004). One way to optimize over both these conflicting objectives is to manipulate the trial order, inside a rapid event-related design (Miezin, Maccotta, Ollinger, Petersen, & Buckner, 2000). One state-of-the-art method for optimizing the trial order is a genetic algorithm which uses two (weighted) loss functions, one for signal detection and one for time-course estimation (Wager & Nichols, 2003). Kao, Mandal, Lazar, & Stufken, 2009, improved on Wager's (2003) design, adding in psychological considerations, and greatly improving execution speed and documentation. As a result, Kao *et al's* (2009) method was used to optimize trial orders for part 1 and 2 of the behavioral task (p**??**), along with the reward category localizer scan (p1).

Mobs of Blobs

All statistical parametric maps (below) were derived from a Random Effects analysis (RFX, or "second-level" in SPM8 jargon), multiple comparison corrected assuming Gaussian Random Fields using the Family Wise Error Rate (FWE) at the $p < 0.05$ level, with a minimum cluster size of 4 voxels (?, ?).

Whole brain activity for the stimulus-response learning portion of the behavioral experiment (i.e. part 2, p**??**) was examined first by comparing all trials to the baseline (rest) condition. This data is presented in two ways. First is the typical statistically thresholded contrast image. The contrast map showed significant (t(15) = 6.59, $p < 0.05$) bilateral activity in the cerebellum, insula and anterior cingulate (Figure 1). Second is a transparent overlay of the raw t-values, which confirms that observed significant effects were robust and widespread in their respective regions, but also allows for the analysis of overall and subthreshold patterns of activity. These raw data suggest near threshold levels of activity in the head of the caudate, ventrolmedial, dorsal lateral frontal cortices as well as (weaker) activity in the occipital lobe (Figure **??**). And indeed in a two-way ANOVA looking at that interaction between gains and losses, significance clusters were observed in head and body of caudate, insula, posterior and anterior cingulate with the posterior activation extending into the precuneus, as well as in dorsal lateral (i.e middle frontal) PFC, and in ventral medial PFC (Figure 3; $F(1, 270) = 30.76$, $p < 0.05$). When trials with gains and losses were examined separately compared to rest, both resulted in activity in the same areas as in the combined condition (not shown).

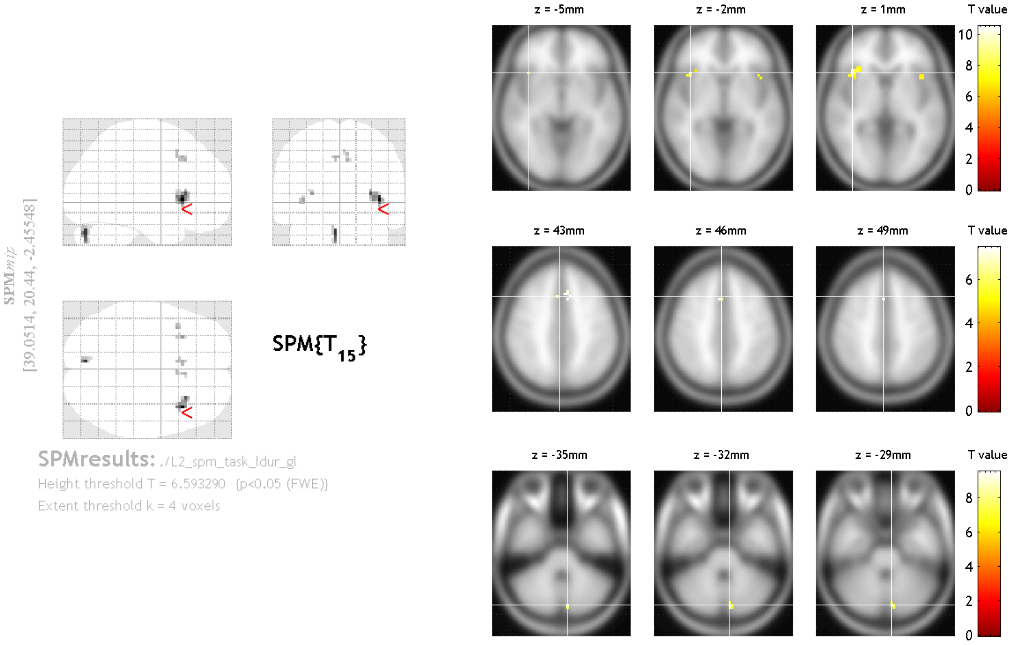

Figure 1. Statistical parametric map for all trials in the stimulus-response learning task (i.e. part 2, p**??**), compared to the rest period. *Left* is a glass brain, showing all significant clusters mapped down to 3 two dimensional representations. *Right* is a set of axial slices highlighting strong areas of activity overlaid onto the T1 MNI-352 template. *Z* is the height of the axial slice in MNI space.

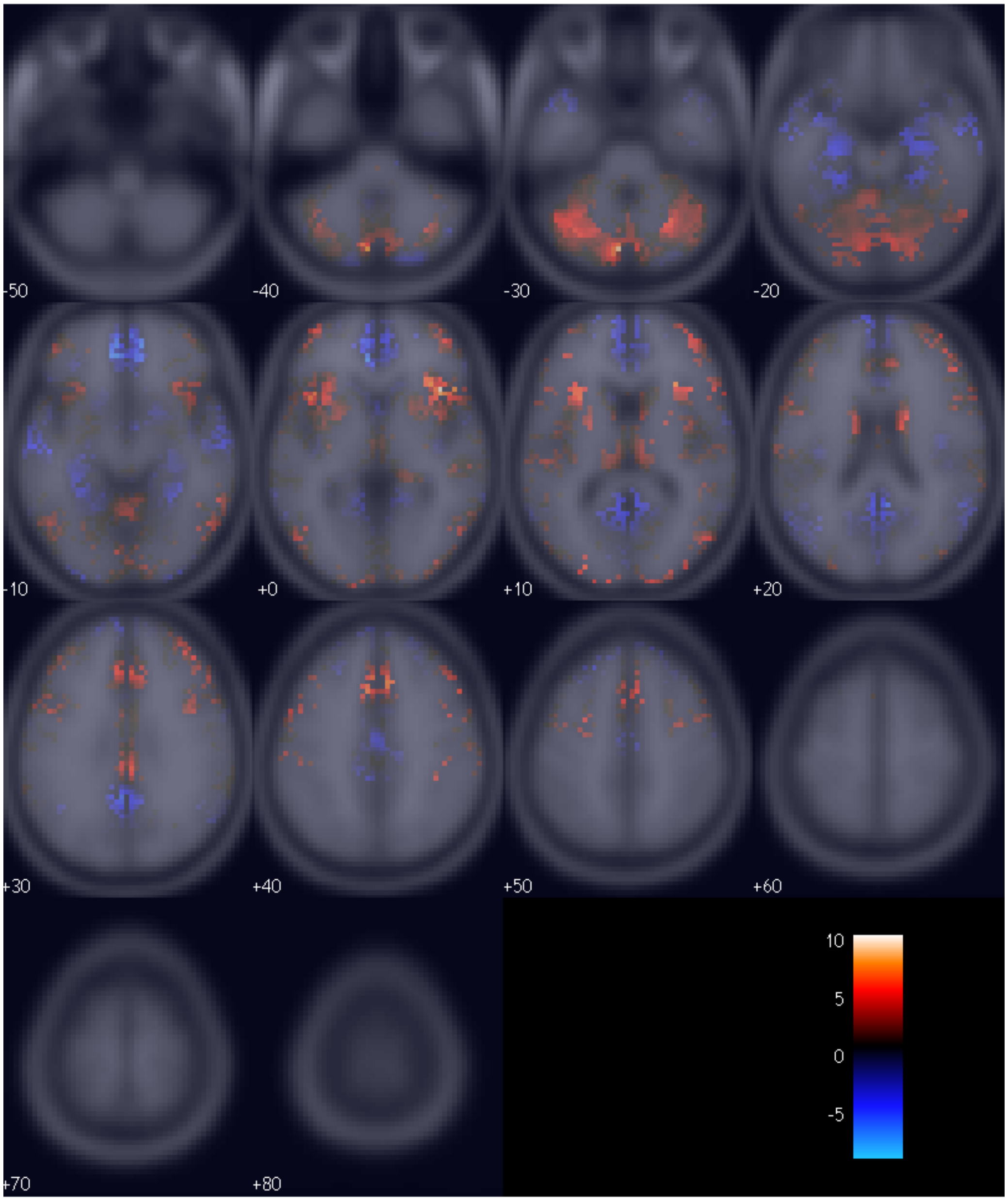

Figure 2. (Raw, that is unthresholded, *t*-values for all trials in the stimulus-response learning task (i.e. part 2), compared to the rest period, overlaid onto the T1 MNI-352 template. Each number is the height of the axial slice in MNI space.

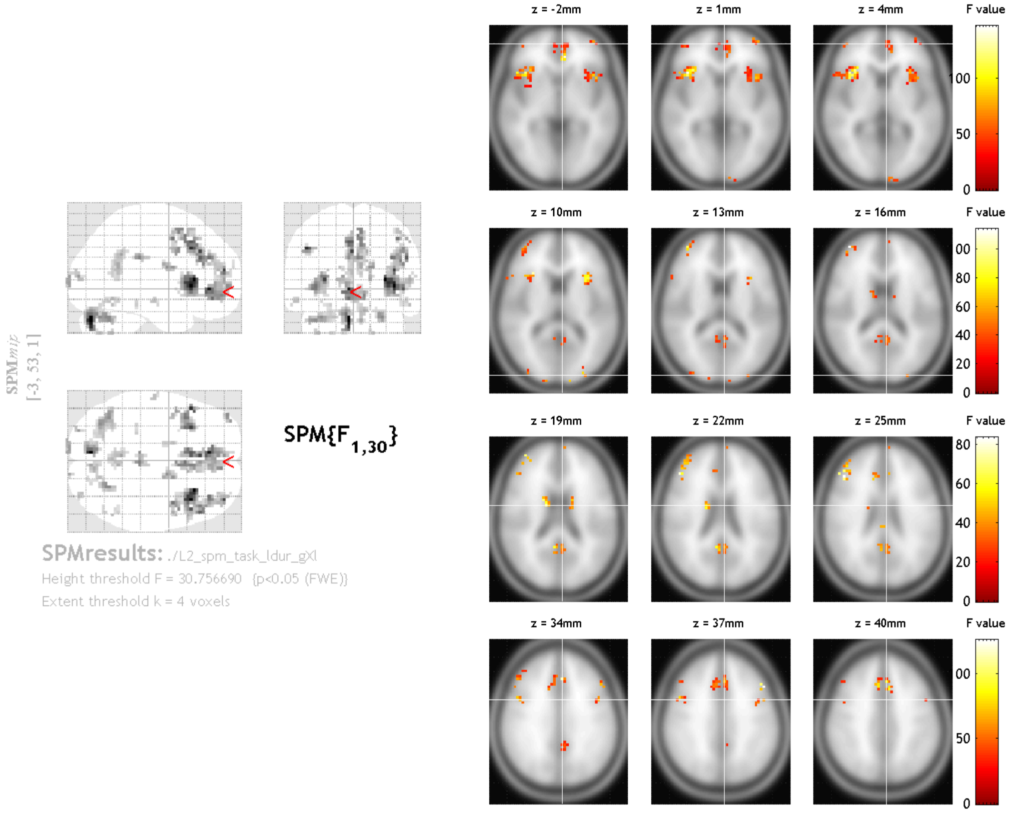

Figure 3. Statistical parametric map for all trials in the stimulus-response learning task (i.e. part 2) examining the interaction between gains and losses. *Left* is a glass brain, showing all significant clusters mapped down to 3 two dimensional representations. *Right* is a set of axial slices highlighting strong areas of activity overlaid onto the T1 MNI-352 template. *Z* is the height of the axial slice in MNI space.

Regions and models

The right chunks. Following whole-brain analysis, regions of interest were selected using two methods, that were later compared. The first employed only regions from the Harvard-Oxford probabilistic anatomical atlas, using the 50% cutoff (?, ?). The second combined anatomical regions with functional clusters isolated using both sets of data collected during the second half of part 1 and from the reward category localizer. Analyses showed the clustered regions and entire anatomical regions displayed very similar model-fits. So to limit the complexity of later analyses, and to increase power, functional analyses were dropped in favor of the larger anatomical regions. Anatomical regions of interest were selected *a priori* based on previous studies of reinforcement and category learning (see the *Introduction for a review*). Left and right subcortical regions of interest were the dorsal caudate, ventral striatum/nucleus accumbens, hippocampus, and amygdala. Bilateral cortical areas were the middle frontal cortex (i.e. dorsal lateral PFC), superior frontal cortex (which contains ventral medial PFC), orbital frontal cortex, anterior and posterior cingulate (ACC and PCC for short).

A Way To(o) Many. In total there 6 models under evaluation – the three kinds of similarity adjustment, ("none", "exp", and "gauss"), with two possible reward codes ("acc" and "gl"). With the two terms of interest (i.e. value and the reward prediction error) that is 12 comparisons. The are also a number of *a priori* confounds to our signals of interest: the similarity metrics, the reward codes, and the grating parameters. Bringing the total to 23. However the models are not nested[1] and so are not amenable to F-tests, the common statistical way to compare model fits. Further complicating the issue is the fact that each of the models is covariate, if not collinear, with the others. To top it off, none of the three are statistically independent; Reinforcement learning can viewed as a regression of the reward code onto behavioral choices. All these factors combined would make statistical testing difficult, to say the least. But fortunately finding *the* best model is not the goal.

The latest recordings of phasic (i.e. reward prediction) activity in the VTA/SNc suggests a complicated reward and prediction error coding scheme (see p**??**), wherein several separate sets of calculations may be carried out independently (?, ?, ?, ?). The observed BOLD signal is then an aggregate of these many activities. It is likely then that more than one of the models is correct, making null hypothesis significance tests an incorrect choice. Model selection is the right choice.

Model selection is the process finding a *family* of models/variates that best predict a given dataset (?, ?), with most techniques trying to wisely balance parsimony with increasing fit (i.e. solving the bias versus variance dilemma (?, ?)). Unfortunately most model selection techniques require assumptions the models cannot meet (e.g. statistical independence). The few that can tend to be complex recent statistical inventions. And rather than navigate the those troubled and unproven waters, I took a simpler approach, simply examining each model independently and ranking them.

[1] Often defined by whether or not tow models can be made identical by adding or subtracting parameters (?, ?)

A score (AIC, Akaike Information Criterion (?, ?)) was assigned to each of the models/codes for every participant and region of interest. The absolute AIC score across participants is not however meaningful. Only the relative values are of interest (?, ?). Individual's scores were normalized and ranked by subtracting from each from the best (lowest) score (?, ?). The normalized set was then transformed to Akaike Weights, a way to easily compare the conditional probabilities of each model being true (?, ?). The Akaike Weights were then averaged across participants for each model and region of interest.

Information on Information. AIC is a measure of loss; how much information is lost by substituting the model for the true distribution, i.e. the data. The lower the AIC score, the better the model. Unlike both the null hypothesis tests, and Bayesian measures, AIC based methods do not seek to find *a* truth, but instead serve to rank models. AIC offers then only relative insight, and is unable to make any claims about absolute significance. Significance is a separate question, one I'll return to later. Besides this limitation, AIC has some significant advantages. Five are reviewed below.

One, unlike maximum-likelihood AIC is designed to a be parsimonious score. It penalizes for additional parameters. It may therefore choose a worse model (as measured by likelihood or mean squared error) over a better but more complex one. This is the essence of Occam's razor[2].

Two, it fits with the process of science. When designing an experiment it is

[2] Famously and pithily expressed as, "Entities are not to be multiplied beyond necessity".

rare that there are only two possible outcomes, instead typically there are several competing hypothesis, some of which may not be mutually exclusive. AIC's focus on relative differences, and evidential weights, meshes perfectly with the reality of multiple working hypotheses (?, ?).

Three, truth can remain elusive. A common alternative to AIC is BIC, the Bayesian Information Criterion. Like AIC, BIC is derived from the log-likelihood of a model, however its derivation requires a rather strict (and often unrealistic) assumption – that the true model is among the candidates (?, ?). And while it may be philosophically debatable whether any mathematical model can *completely* describe reality, in this study it is I know my models are incomplete. As, one, the human reinforcement learning literature contains several recent theoretically unaccounted for findings and, two, there are theoretical developments I do not include here to keep the models tractable (see the *Introduction* for a review).

Four, AIC values are easily interpretable once they're transformed to Akaike Likelihoods or Weights[3]. The likelihood is simply the likelihood the model is correct (based on the information loss associated with it), while the Akaike Weights are just normalized likelihoods. As the Weights sum to one, the conditional likelihood of one model compared to another is just the ratio of their weights (?, ?). For example, the conditional likelihood of model A over model B is just w_A/w_B. That is, the likelihoods and Akaike Weights are intrinsically measures of effect size (?, ?, ?).

[3] Likelihood for model k among K working hypotheses/models is given by $L_k = e^{-0.5(AIC_k - min_K(AIC))}$, which is then normalized, becoming an Akaike Weight by $w_k = L_k / \sum\limits_{k=1}^{K} L_k$ (?, ?).

Despite the fact that it is often used to express the likelihood of correctly rejecting the null hypothesis, the p value is not a measure of effect, as p is contingent not just on effect size but on sample number.

Five, AIC has a history with models of categorization. ?, ?, ?, among several others, used AIC to compare behavioral results to several alternative models of categorization.

F-Them. AIC ranks offer no information on significance, in the familiar null hypothesis sense, or on the absolute fit of the model. I addressed both of these in a series of F-tests run prior to AIC analysis. These (fixed-effect, across participant) ominbus tests asked whether the total set of regression parameters for each linear model (described below) could explain the BOLD time series better than chance, i.e could the null hypothesis (of 0) be rejected. Keeping with recommendations of ?, ?, ?, who argue that as AIC and significance tests are so dissimilar that direct comparison/interaction between them them will be at best misleading, the models are not discarded based on significance. All models are retained, and later AIC ranked. The F-tests are a separate measure whose results are integrated during interpretation, not during model selection.

Code, BOLD, and Models.. A total of 23 models were compared for each of the 16 regions of interest for each of the 16 subjects, 5888 comparisons in total. Each of the models is described below. In general, a time-series (e.g the reward prediction error for each trial or the similarity for that trial's outcome) was convolved with a "canonical" haemodynamic response function, a mixture of gamma functions that serves as a parsimonious estimate of the (instantaneous) BOLD response (?, ?). The convolved series was then low-pass filtered, matching the treatment of the BOLD data (p2). Each convolved and filtered model was then regressed onto the BOLD response for each participant's region of interest, retaining all parameters and fit measures inside subject-level HDF5 files (`http://www.hdfgroup.org/HDF5/`).

No available fMRI analysis package returns AIC scores (or measures that could be converted to such) and none allow for the efficient (i.e programmatic) analysis of many competing computational models. So I created a roi-focused fMRI data analysis tool in Python (v2.7.1) to meet those two needs. This module, simply named "roi", has since been release under the BSD license and is available for download at `https://github.com/andsoandso/roi`. It relies on the nibabel library to read the nifiti-1 files (v1.2.0; `http://nipy.org/nibabel`), nitime for timeseries analysis, (v0.4; `http://nipy.sourceforge.net/nitime/`) Numpy for generic numerical work (v1.6.1; `http://numpy.scipy.org/`), with the GLS function from the scikits.statsmodels module handling the regerssions (v0.40; `http://statsmodels.sourceforge.net/`). Model-to-BOLD fit parameters, as well as other useful metadata, was then extracted and stored in text files suitable for importing into R (v2.15.1; `http://www.r-project.org/`). All plotting and model ranking (as well as the *F*-tests) were carried out in R. For complete BSD licensed code see, `https://github.com/andsoandso/fmri/tree/master/catreward/roi/results`.

Our Kinds of Models. To simplify visualization and analysis, each of the models was classified into one of 5 families. Family one, denoted “boxcar”, was identical to that first used in the whole-brain analysis (p4) – all trials versus the rest condition. This is a univariate time-series that predicts no trial-specific effects; No matter the task the brain, thus the BOLD response, just flicks on then off. It serves as a useful standard against which to compare the model-based regressors. The next two families were controls (i.e. *a priori* covariates), with the similarity metrics and grating parameters grouped into one family (“control_similarity”) and the reward codes (both raw and similarity adjusted) into the other (“control_reward”). The fourth family was all the reward prediction errors (“rpe”). The fifth was the value estimates (“value”).

Table 1:: All models, their designations (Codes), families, and descriptions.

Number	Code	Family	Description
1	0_1	boxcar	The simplest model, a univariate analysis of all conditions.
2	acc	control_reward	Behavioral accuracy.
3	acc_exp	control_reward	Behavioral accuracy, diminished by (exponential) similarity.
4	acc_gauss	control_reward	Behavioral accuracy, diminished by (gaussian) similarity.
5	gl	control_reward	Gains and losses.

6	gl_exp	control_reward	Gains and losses, diminished by (exponential) similarity.
7	gl_gauss	control_reward	Gains and losses, diminished by (gaussian) similarity.
8	rpe_acc	rpe	Reward prediction error - derived from accuracy.
9	rpe_acc_exp	rpe	Reward prediction error - derived from accuracy diminished by (exponential) similarity.
10	rpe_acc_gauss	rpe	Reward prediction error - derived from accuracy diminished by (gaussian) similarity.
11	value_acc	value	Value - derived from accuracy.
12	value_acc_exp	value	Value - derived from accuracy diminished by (exponential) similarity.
13	value_acc_guass	value	Value - derived from accuracy diminished by (gaussian) similarity.
14	rpe_gl	rpe	Reward prediction error - derived from gains and loses.

15	rpe_gl_exp	rpe	Reward prediction error - derived from gains and losses diminished by (exponential) similarity.
16	rpe_gl_gauss	rpe	Reward prediction error - derived from gains and losses diminished by (gaussian) similarity.
17	value_gl	value	Value - derived from gains and losses.
18	value_gl_exp	value	Value - derived from gains and losses diminished by (exponential) similarity.
19	value_gl_gauss	value	Value - derived from gains and losses diminished by (gaussian) similarity.
20	exp	control_similarity	Outcome similarity (exponential).
21	gauss	control_similarity	Outcome similarity (gaussian).
22	angle	control_similarity	Grating angle parameter.
23	width	control_similarity	Grating width parameter.

Model Results

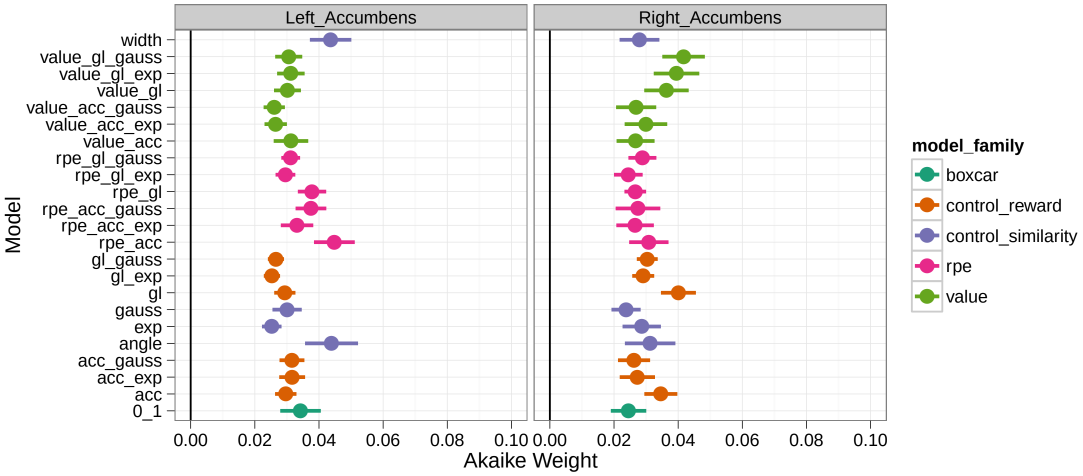

Figure 4. Nucleus Accumbens (left and right) – Akaike Weights for all models. Colors indicate model family (see p12 for details). Bars represent standard errors.

Under Cortical. Discuss the subcortical results first....

On that thinkin' sheet. Now discuss the cortical results....

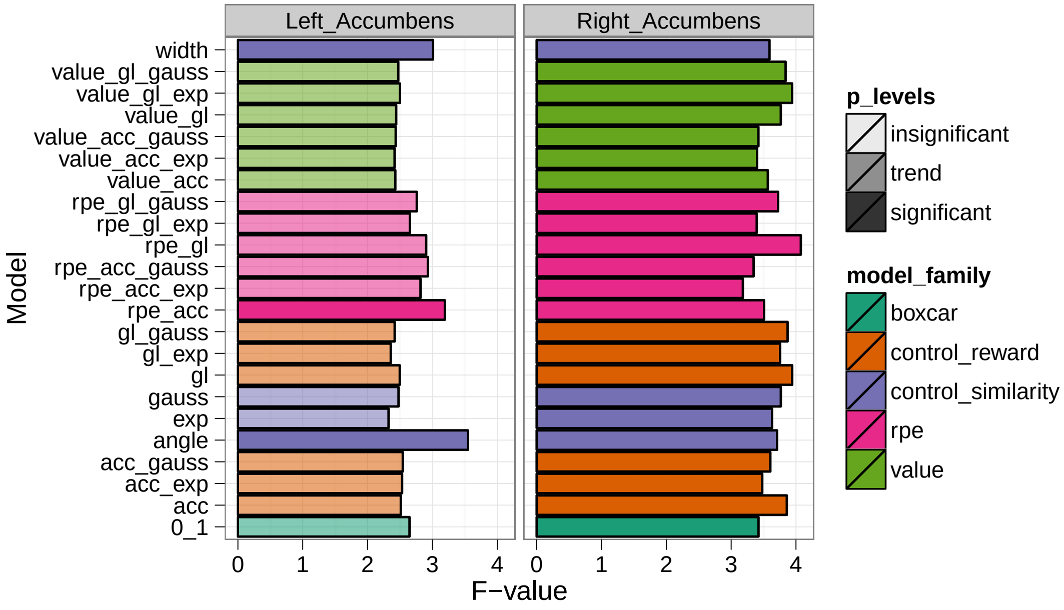

Figure 5. Nucleus Accumbens (left and right) – F-values for all models. Significance is the $p < 0.05$ level, trend is between $p < 0.05$ and 0.10. Colors indicate model family (see p12 for details).

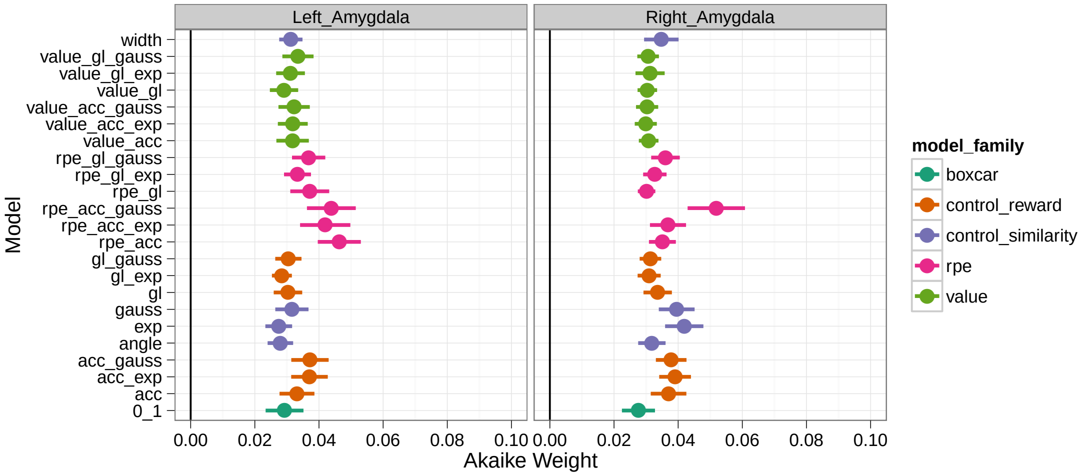

Figure 6. Amygdala (left and right) – Akaike Weights for all models. Colors indicate model family (see p12 for details). Bars represent standard errors.

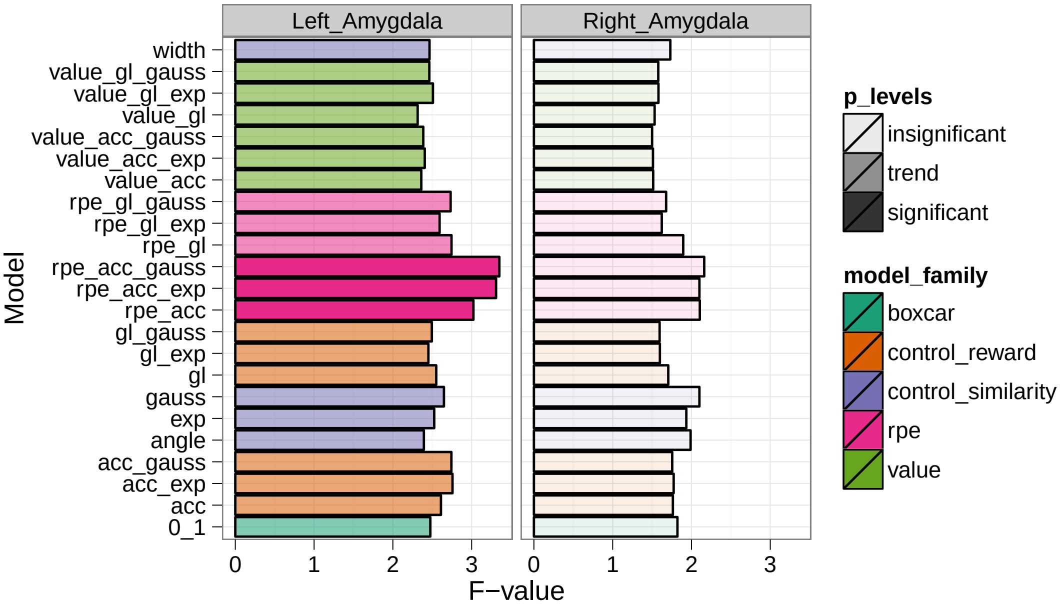

Figure 7. Amygdala (left and right) – F-values for all models. Significance is the $p <$ 0.05 level, trend is between $p <$ 0.05 and 0.10. Colors indicate model family (see p12 for details).

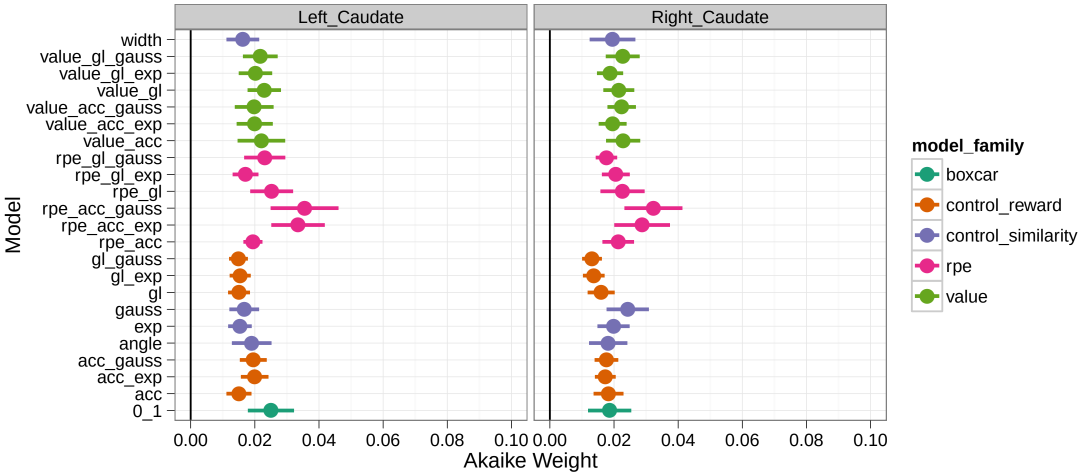

Figure 8. Dorsal Caudate (left and right) – Akaike Weights for all models. Colors indicate model family (see p12 for details). Bars represent standard errors.

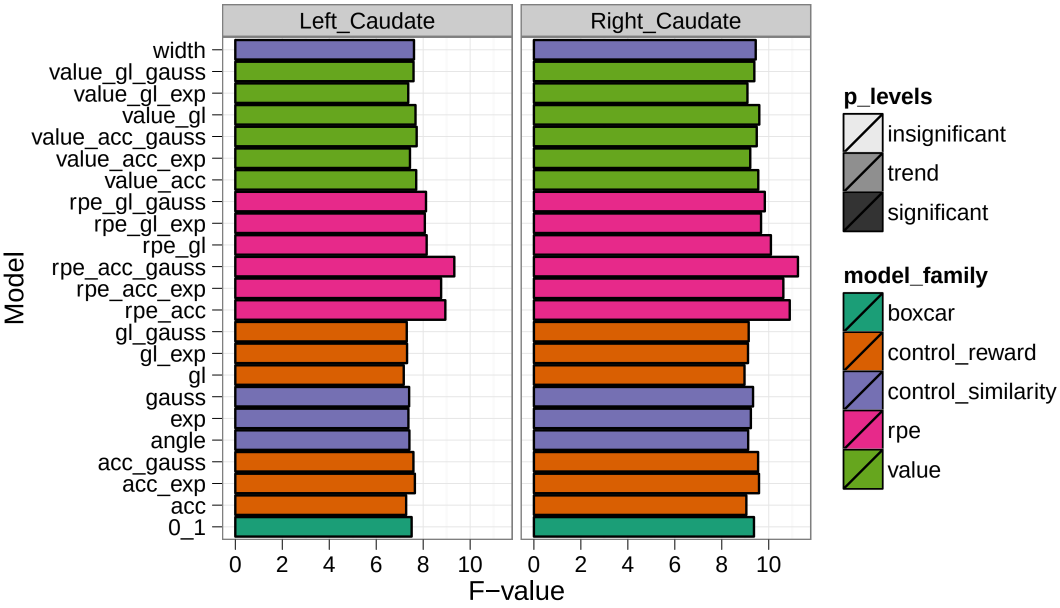

Figure 9. Dorsal Caudate (left and right) – F-values for all models. Significance is the $p < 0.05$ level, trend is between $p < 0.05$ and 0.10. Colors indicate model family (see p12 for details).

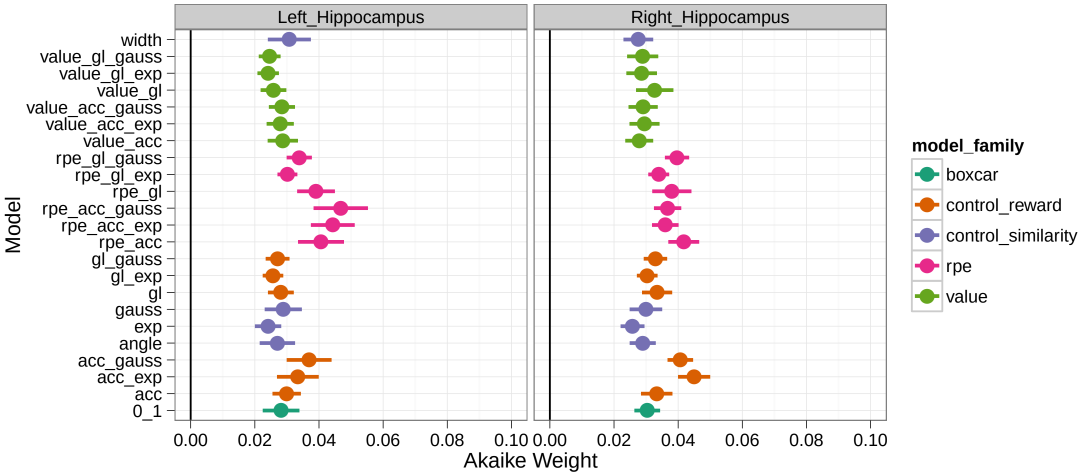

Figure 10. Hippocampus (left and right) – Akaike Weights for all models. Colors indicate model family (see p12 for details). Bars represent standard errors.

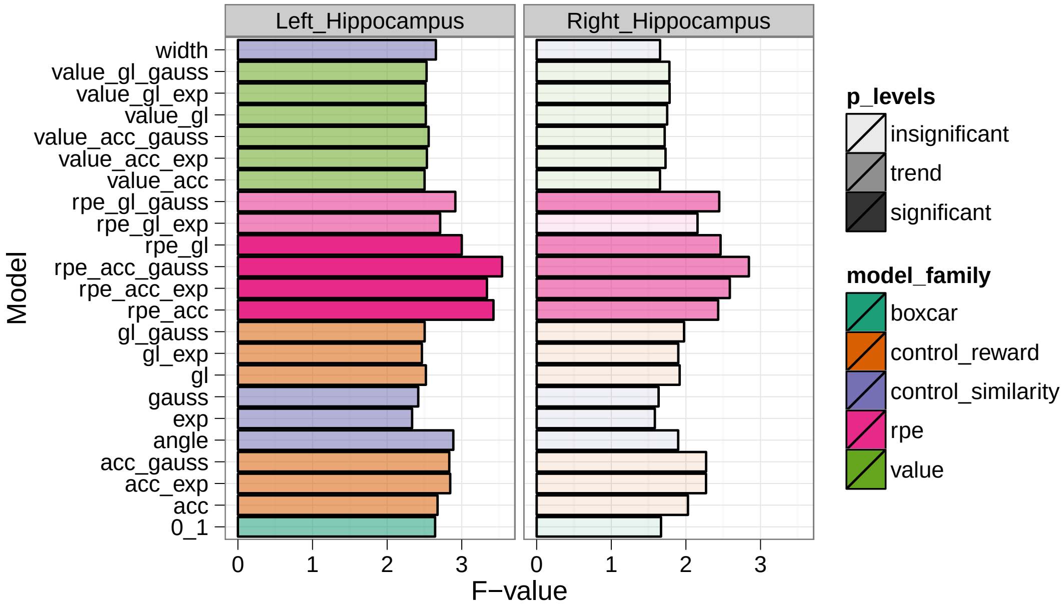

Figure 11. Hippocampus (left and right) – F-values for all models. Significance is the $p <$ 0.05 level, trend is between $p <$ 0.05 and 0.10. Colors indicate model family (see p12 for details).

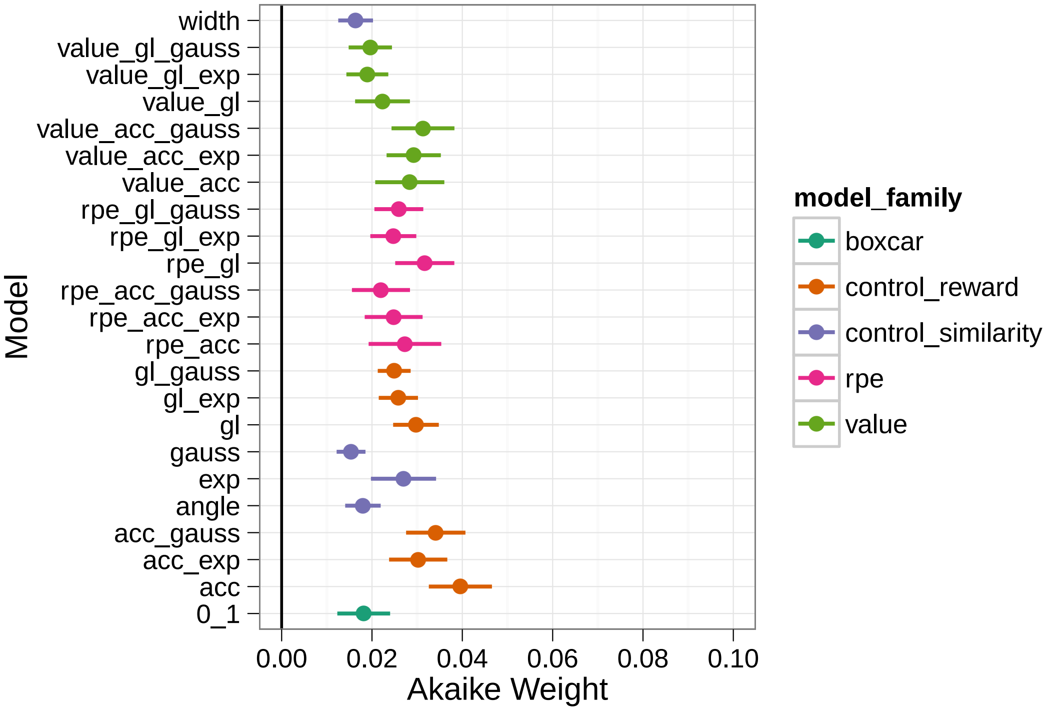

Figure 12. Anterior Cingulate Gyrus – Akaike Weights for all models. Colors indicate model family (see p12 for details). Bars represent standard errors.

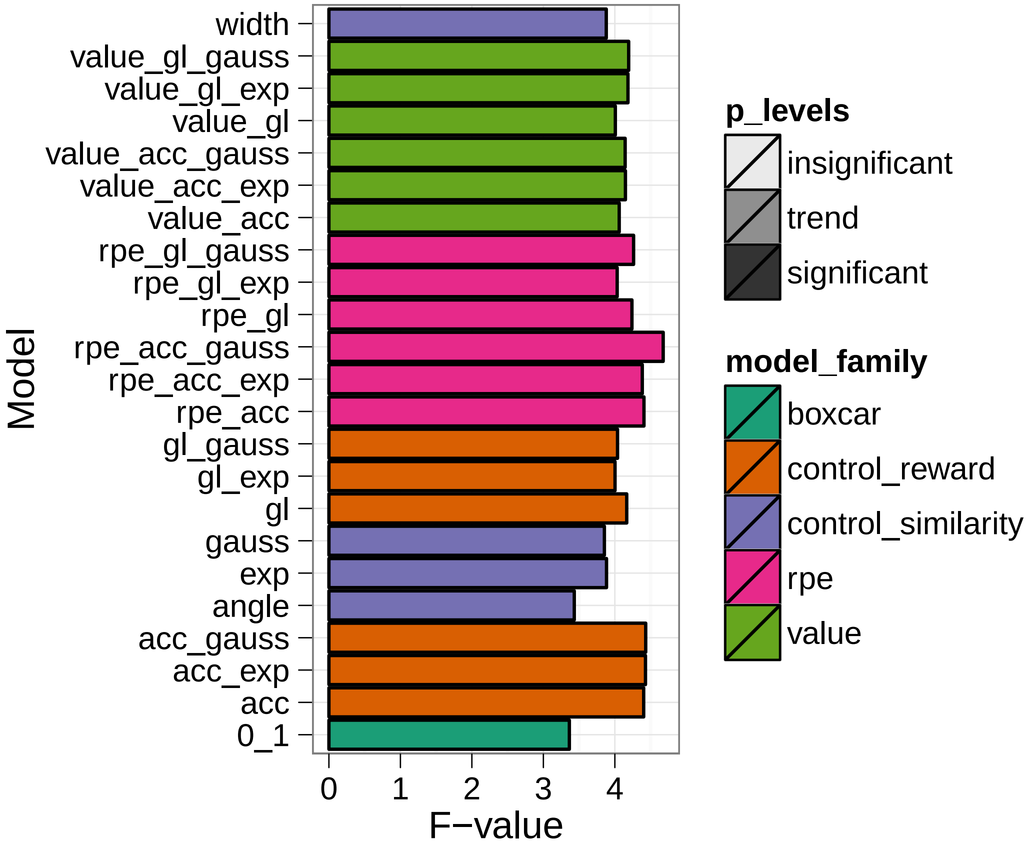

Figure 13. Anterior Cingulate Gyrus – F-values for all models. Significance is the $p < 0.05$ level, trend is between $p < 0.05$ and 0.10. Colors indicate model family (see p12 for details).

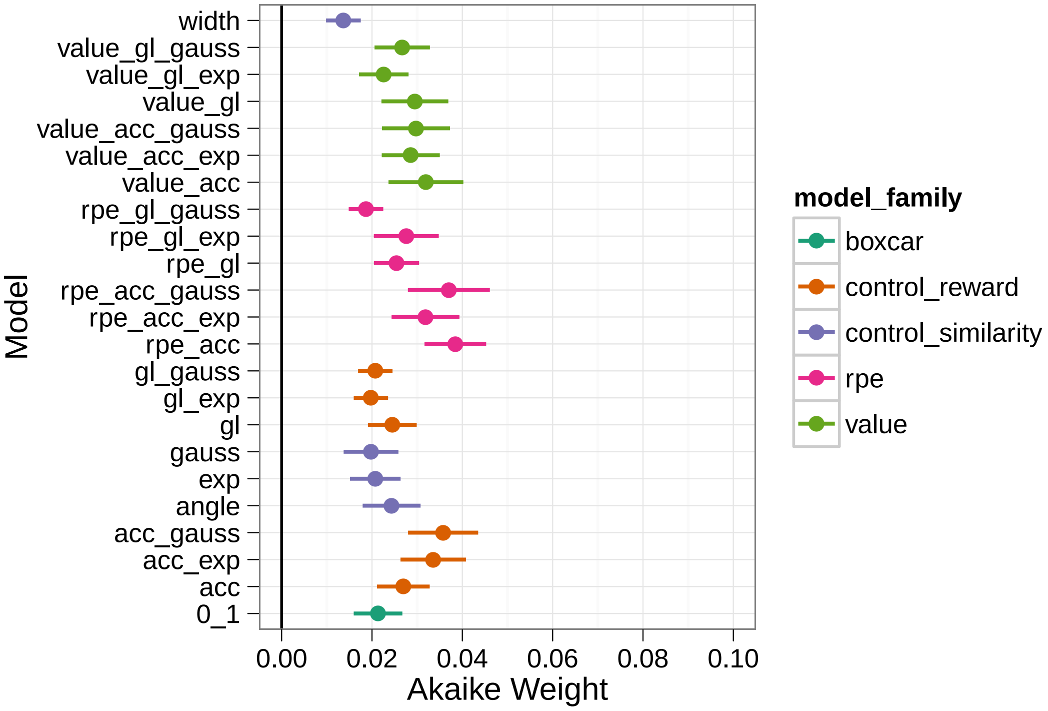

Figure 14. Posterior Cingulate Gyrus – Akaike Weights for all models. Colors indicate model family (see p12 for details). Bars represent standard errors.

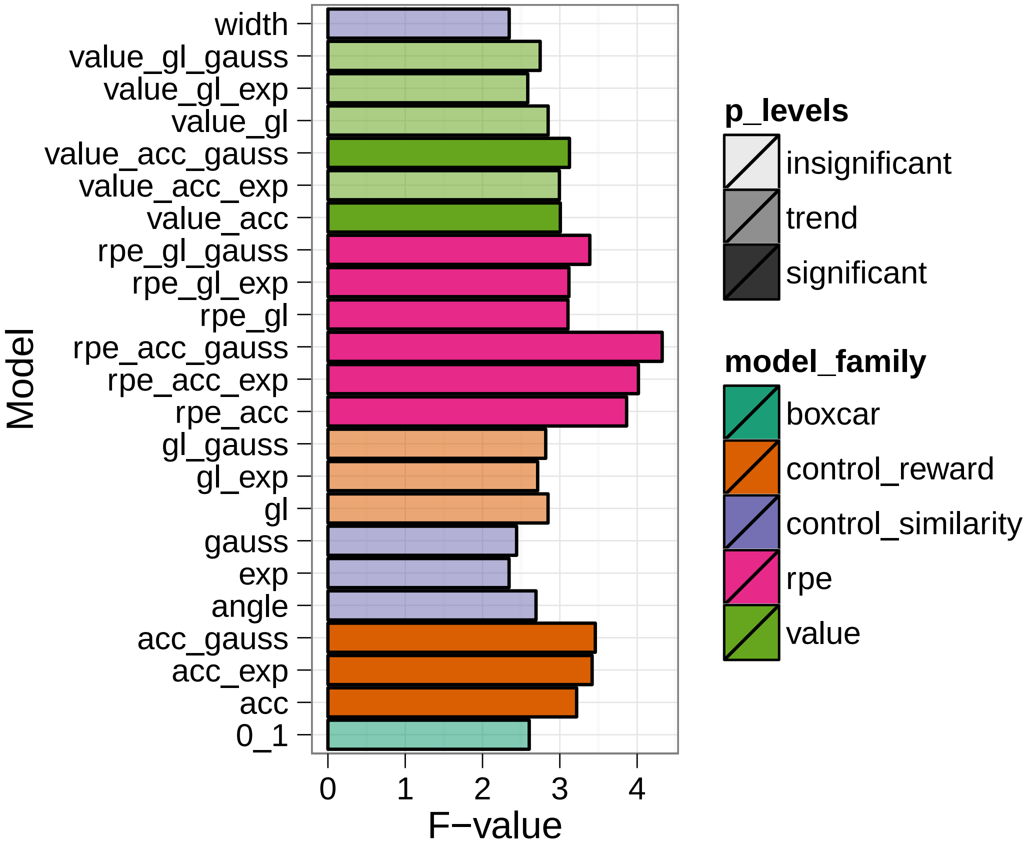

Figure 15. Posterior Cingulate Gyrus – F-values for all models. Significance is the $p < 0.05$ level, trend is between $p < 0.05$ and 0.10. Colors indicate model family (see p12 for details).

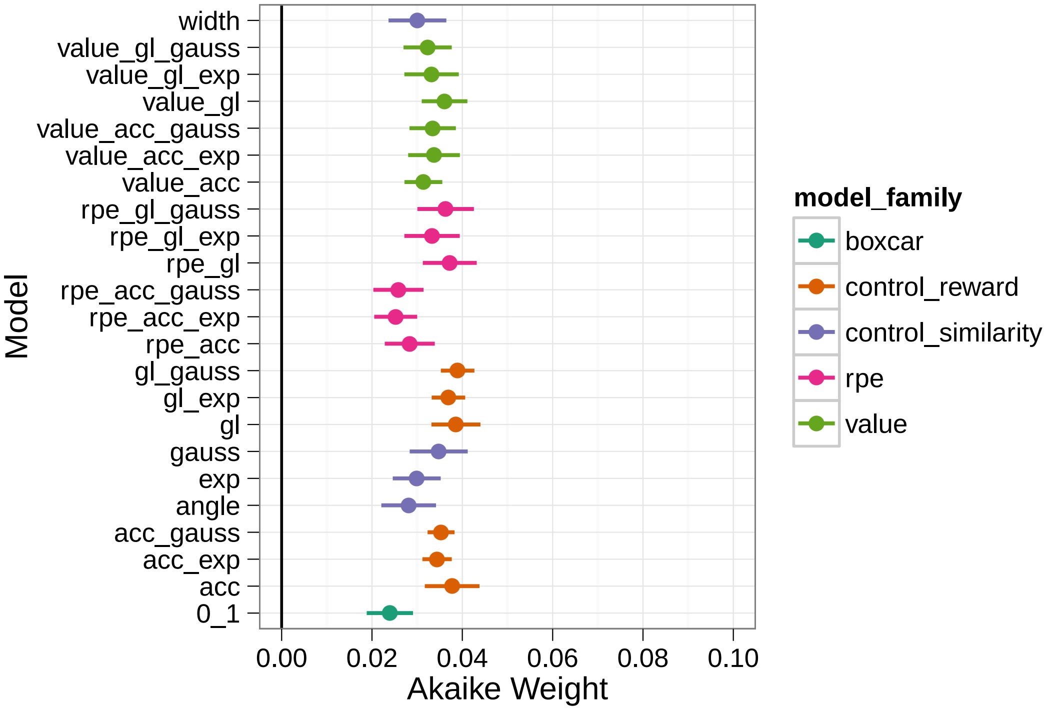

Figure 16. Frontal (ventral) medial PFC – Akaike Weights for all models. Colors indicate model family (see p12 for details). Bars represent standard errors.

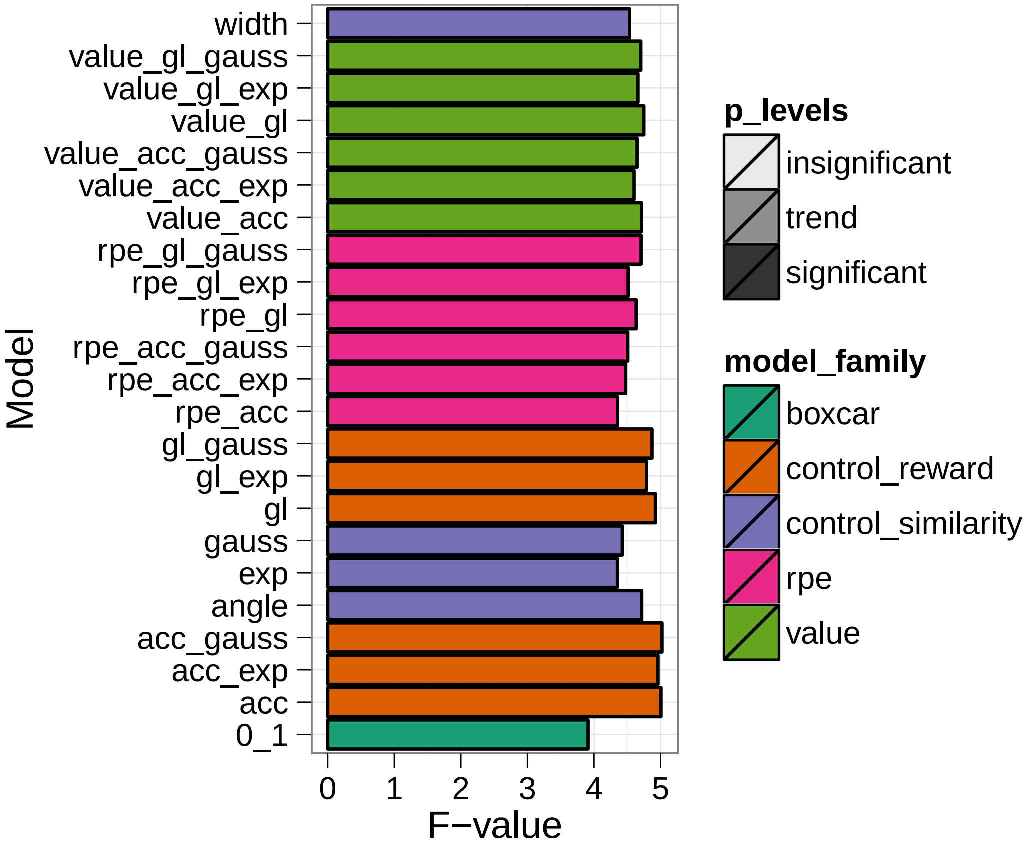

Figure 17. Frontal (ventral) medial PFC – F-values for all models. Significance is the $p < 0.05$ level, trend is between $p < 0.05$ and 0.10. Colors indicate model family (see p12 for details).

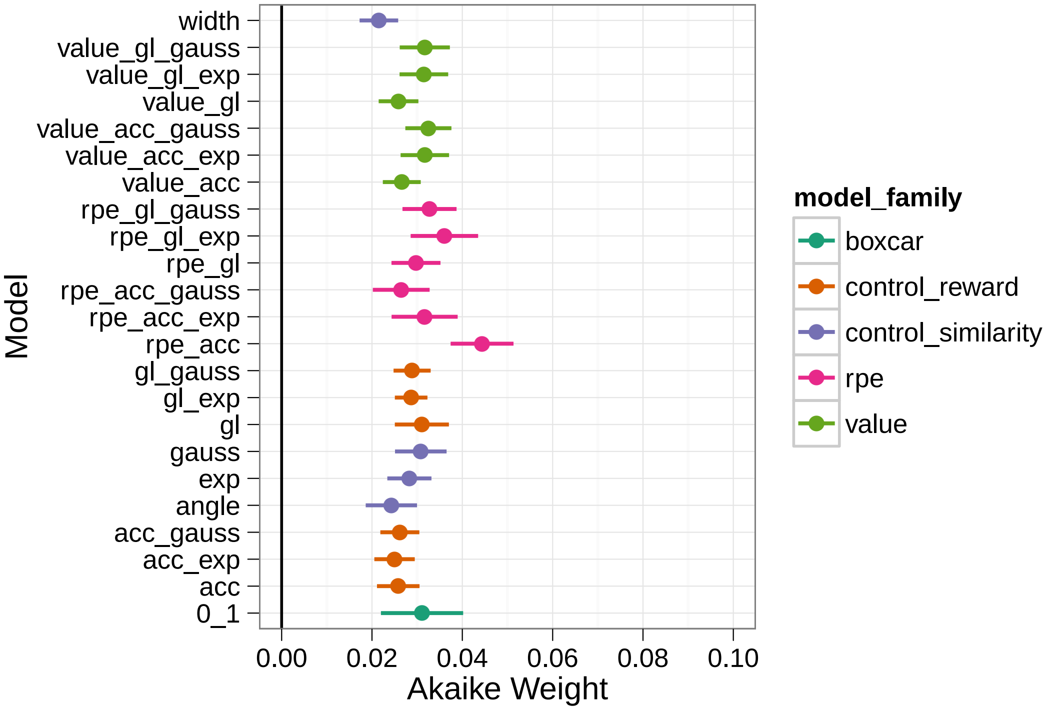

Figure 18. Orbital Frontal Cortex – Akaike Weights for all models. Colors indicate model family (see p12 for details). Bars represent standard errors.

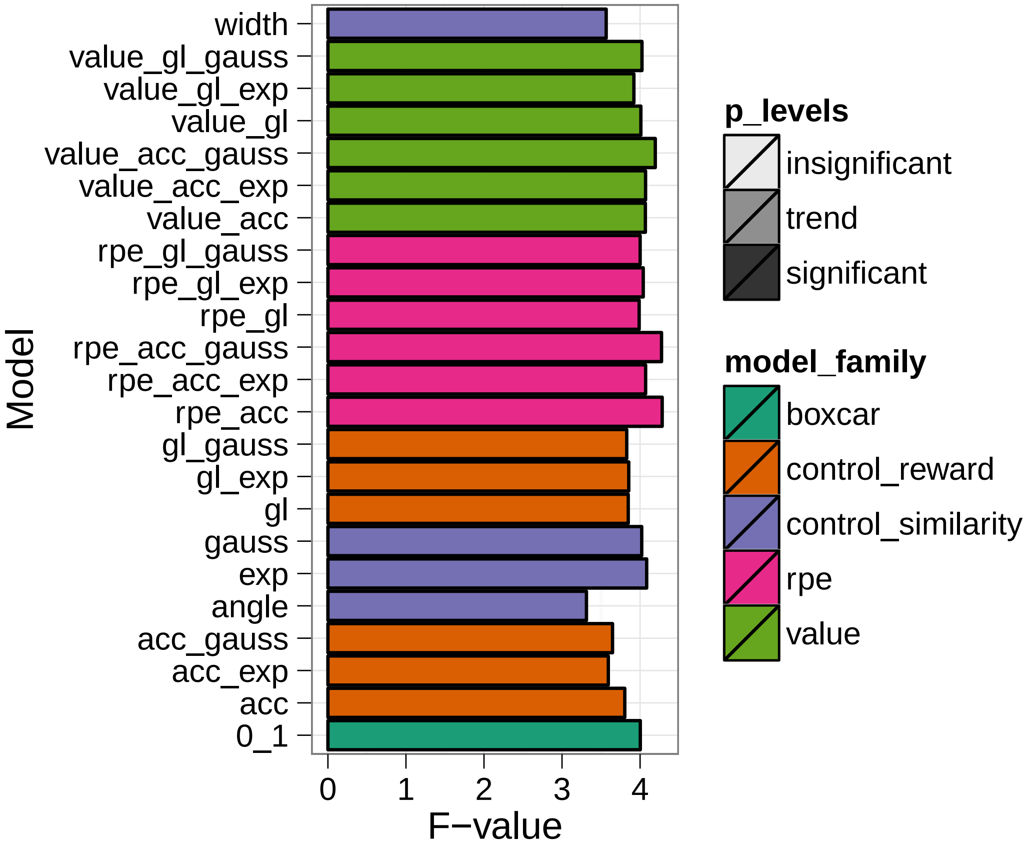

Figure 19. Orbital Frontal Cortex – F-values for all models. Significance is the $p < 0.05$ level, trend is between $p < 0.05$ and 0.10. Colors indicate model family (see p12 for details).

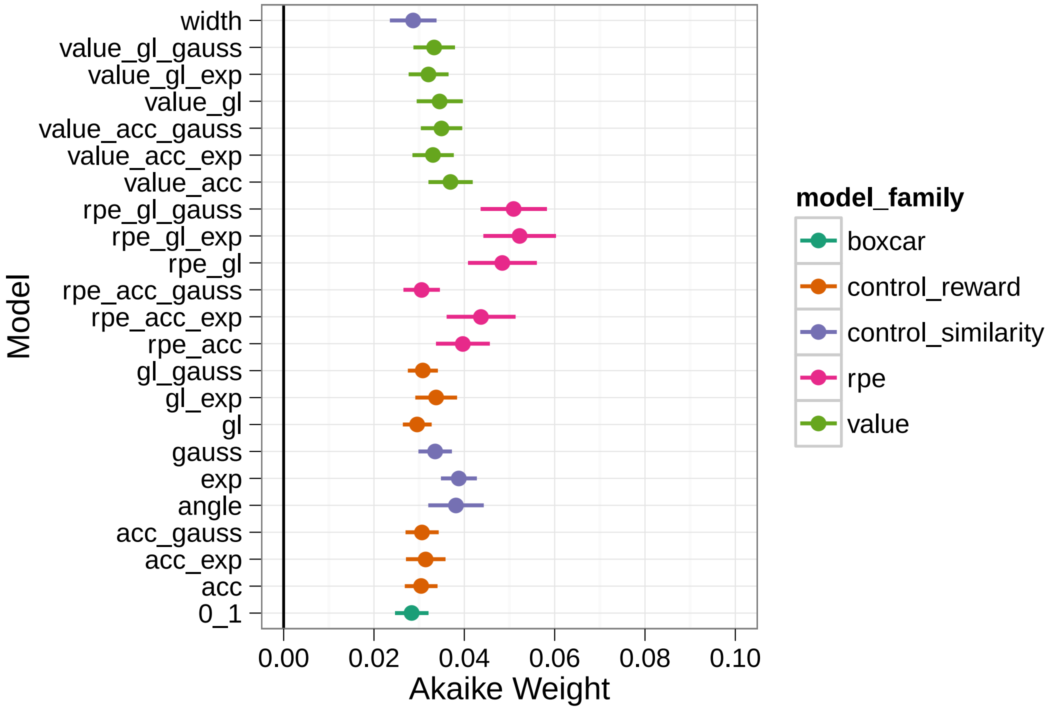

Figure 20. Insula – Akaike Weights for all models. Colors indicate model family (see p12 for details). Bars represent standard errors.

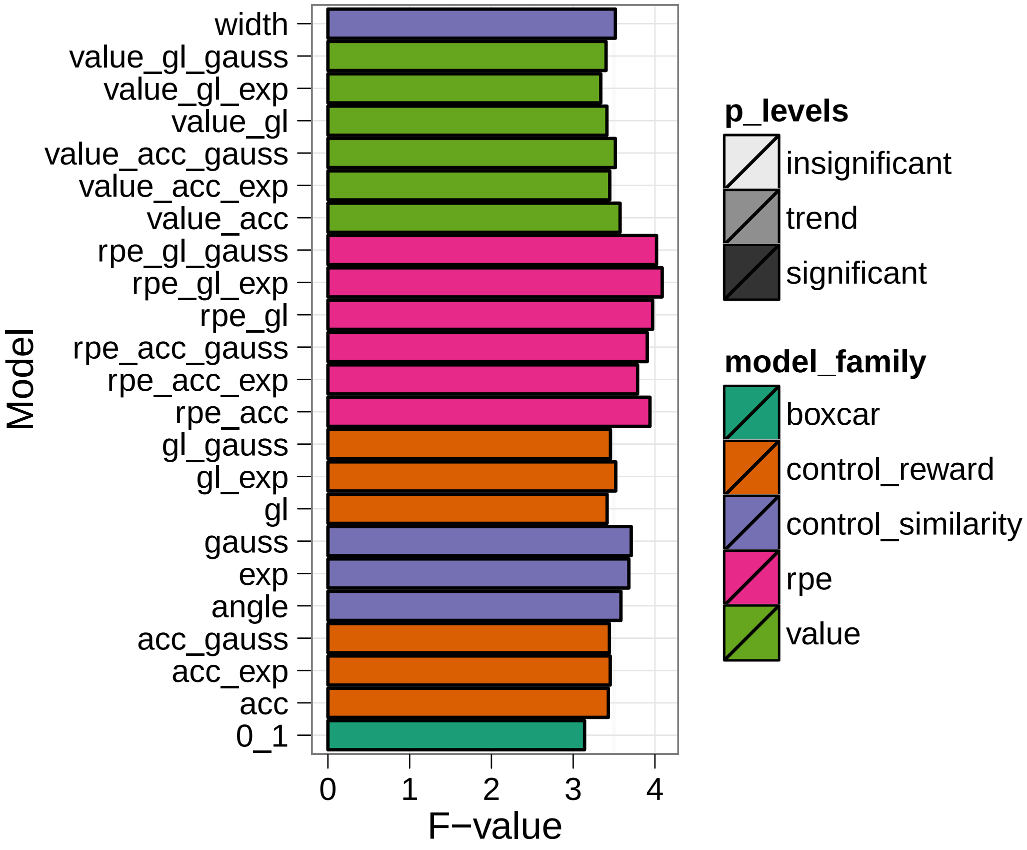

Figure 21. Insula – F-values for all models. Significance is the $p < 0.05$ level, trend is between $p < 0.05$ and 0.10. Colors indicate model family (see p12 for details).

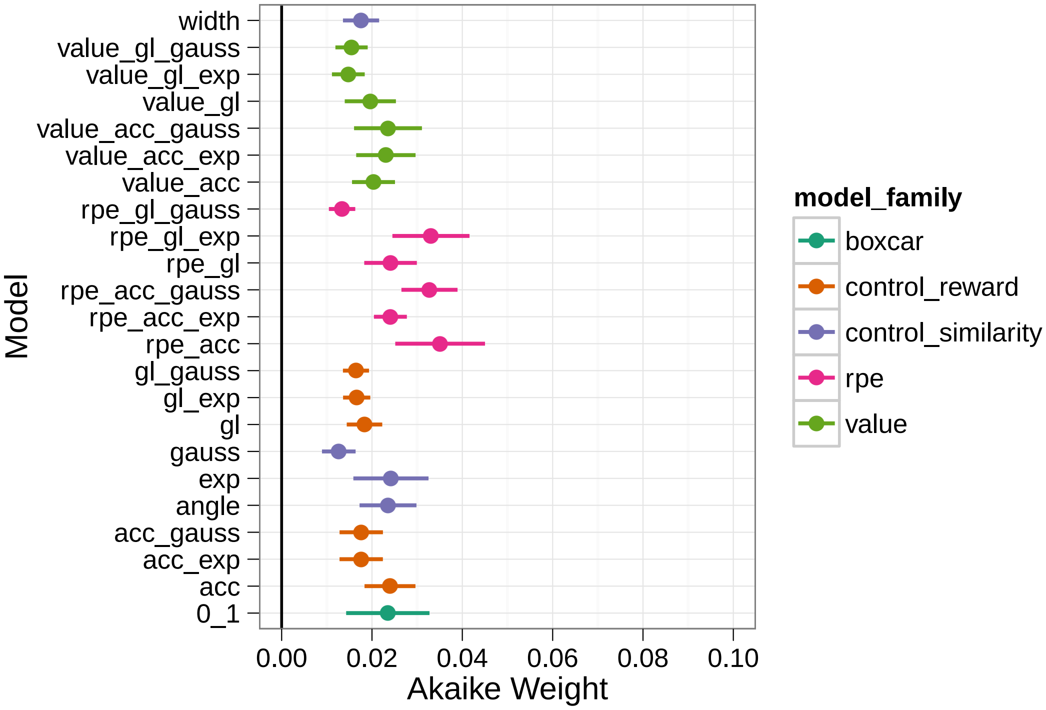

Figure 22. Middle Frontal (dorsal-lateral) PFC – Akaike Weights for all models. Colors indicate model family (see p12 for details). Bars represent standard errors.

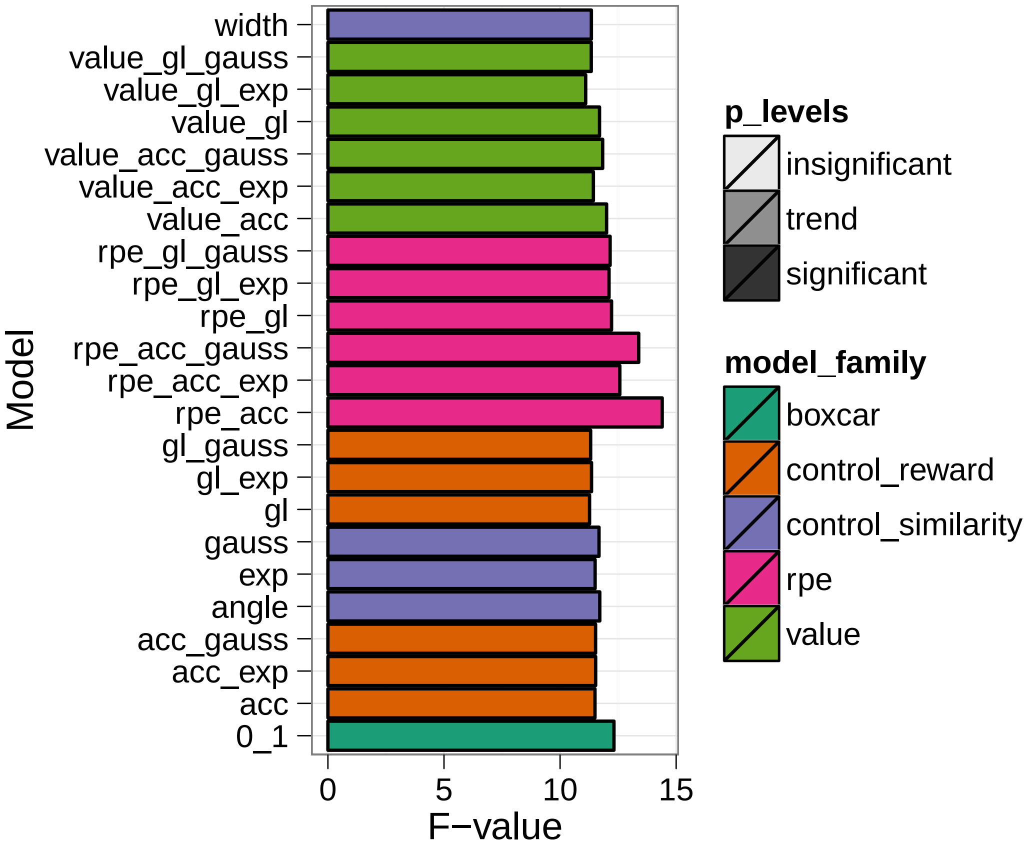

Figure 23. Middle Frontal (dorsal-lateral) PFC – F-values for all models. Significance is the $p < 0.05$ level, trend is between $p < 0.05$ and 0.10. Colors indicate model family (see p12 for details).

References

Birn, R. M., Cox, R. W., & Bandettini, P. A. (2002, Jan). Detection versus estimation in event-related fmri: choosing the optimal stimulus timing. *Neuroimage*, *15*(1), 252–64.

Dale, A. M. (1999, Jan). Optimal experimental design for event-related fmri. *Hum Brain Mapp*, *8*(2-3), 109–14.

Kao, M.-H., Mandal, A., Lazar, N., & Stufken, J. (2009, Feb). Multi-objective optimal experimental designs for event-related fmri studies. *Neuroimage*, *44*(3), 849–56.

Liu, T. T. (2004, Jan). Efficiency, power, and entropy in event-related fmri with multiple trial types. part ii: design of experiments. *Neuroimage*, *21*(1), 401–13.

Miezin, F. M., Maccotta, L., Ollinger, J. M., Petersen, S. E., & Buckner, R. L. (2000, Jun). Characterizing the hemodynamic response: effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing. *Neuroimage*, *11*(6 Pt 1), 735–59.

Wager, T. D., & Nichols, T. E. (2003, Feb). Optimization of experimental design in fmri: a general framework using a genetic algorithm. *Neuroimage*, *18*(2), 293–309.