Model Appendix

We present the model used to illustrate our account of idealization. This model is a simplified version of the one presented in Sakamoto, et al, (2008).

The model represents categories as Gaussians that give the mean and standard deviation along each dimension for which the category is defined. When an object is encountered, each cluster *i* produces an activation a,

$$a_i = \frac{1}{s_i\sqrt{2\pi}} e^{-\frac{dij^2}{2si^2}} \tag{1}$$

which is a function of the distance *d*,

$$d_{ij} = \sqrt{\sum_m (x_{jm} - x_{im})^2} \tag{2}$$

of the item x_j to the mean x of cluster *i* on each dimension *m*, and the standard deviation of the cluster, s_i. The means of each cluster are learned via gradient descent on error:

$$E = \frac{1}{2}(t_i - a_i)^2 \tag{3}$$

$$\Delta x_{im} = -\varepsilon\frac{\partial E}{\partial M} = \varepsilon(t_i - a_i)\frac{x_{jm} - x_{im}}{s_i^3\sqrt{2\pi}} e^{-\frac{dij^2}{2si^2}} \tag{4}$$

Where ε is the learning rate and t_i is the training value to cluster *i*, equal to α if the stimulus is in category *i*, and 0 otherwise.

In the present application, ε and s_i are estimated as free parameters. Parameter values were attained by fitting the average final cluster position attained from 10,000 runs of the model to the contrast dimension reconstruction results in each of the unidimensional conditions in both experiments (see Table A1 & A2). α was set to the maximum value of a_i (i.e., at d_{ij}=0) given the fitted value of the s_i parameter.

One surprising outcome that was not predicted by the model is the negative distortion found on the non-contrast dimension in the unidimensional conditions. The model, as well as conventional family resemblance accounts of typicality, would expect non-contrasted dimensions to be structured around the true category mean, and not around points closer to other categories. Research on memory for stimulus magnitude, however, suggests that there are occasions in which stimulus representations can be pulled toward the grand mean (Huttenlocher, Hedges & Vevea, 2000; Sailor & Miriam, 2005). This would likely be the case if subjects viewed the values on non-contrasted dimensions as coming from a single distribution.

Table A1.

Reconstruction results for each condition in Experiments 1 & 2

Experiment 1		Reconstruction (%ile)
	Political Categories Contrast	3.392
	Political Categories Non-contrast	-1.976
	Energy-source Contrast	3.061
	Energy-source Non-contrast	-2.604
Experiment 1		Reconstruction (mm)
	Unidimensional Contrast	1.700
	Unidimensional Non-contrast	-0.8081

Table A2.

Learning rate ε and standard deviation s_i parameters for Experiments 1 & 2

	ε	s_i
Experiment 1 (Energy-source)	94500	45
Experiment 1 (Political Categories)	104750	45
Experiment 2	415	15

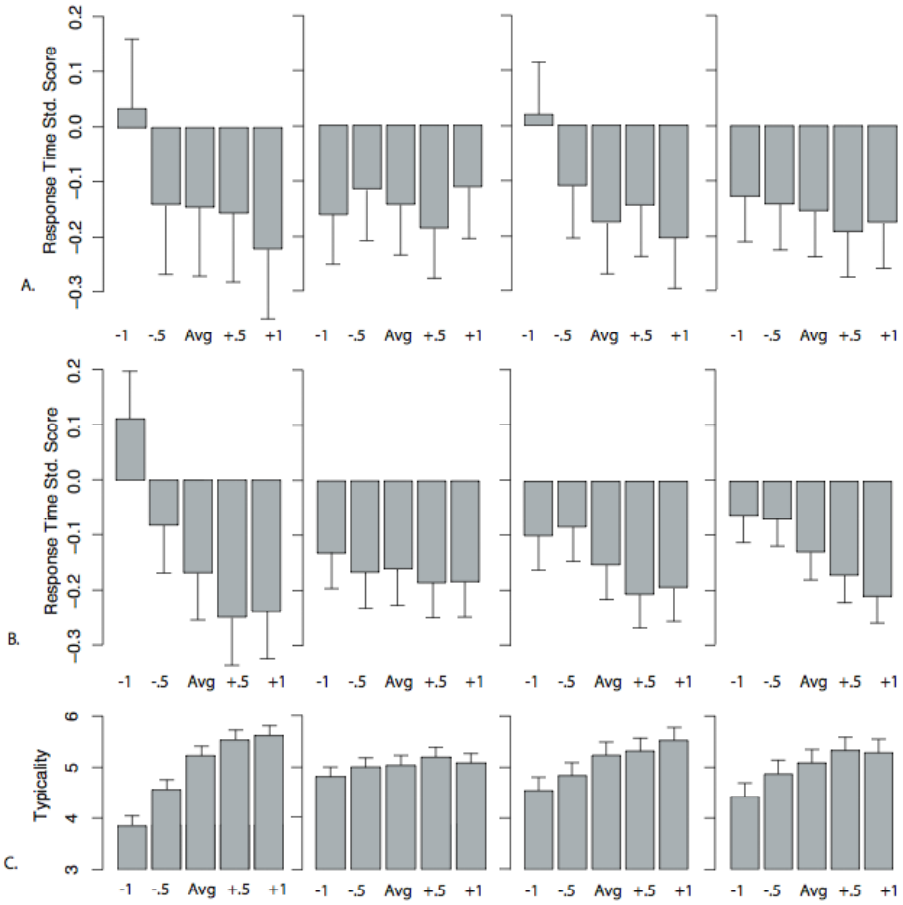

Figure S1. Response time (A) results for Experiment 1, and response time (B) and typicality results (C) for Experiment 2. Each bar represents the within subjects mean (proportion correct/response time/typicality) for stimuli at a particular distance (in z-scores) from the mean stimulus along a stimulus dimension. Positive values indicate ideal stimuli, whereas negative values indicate stimuli located away from the mean and toward the contrasting category.