Neurons in Orbitofrontal Cortex Encode Economic Value

By Camillo Padoa-Schioppa and John A. Assad

Supplementary Information

Supplementary Methods

Subjects, setup, and experimental design

One male ("V", 9.5 Kg) and one female ("L", 6.3 Kg) rhesus monkey (*Macaca mulatta*) participated in the experiments. Neither animal had been used for previous studies. During the experiments, monkeys sat in a monkey chair in a darkened room. Their head was restrained, and their eye position was monitored continuously using a scleral eye-coil system[1] (River-bend Instruments). A computer monitor was placed 57 cm in front of the monkeys. The behavioral task was controlled by custom-written software through a computer interface (ITC-18, Instrutech Corporation).

Each trial began with the appearance of a fixation point (0.2° of visual angle) placed in the center of the monitor. The monkey directed gaze to the fixation point. After 1.5 s, two sets of squares appeared on the left and on the right sides of the fixation point (*offer*). The color of the squares indicated the juice type, and the number of squares indicated the juice amount. After the *offer*, the monkey maintained fixation for a 1-2 s randomly variable delay, until a *go* signal. The *go* was signaled by the appearance of two saccade targets (0.2° of visual angle) placed 7° to the left and to the right of the center fixation point. After the *go*, the monkey had 1.5 s to indicate its choice by making an eye movement towards one of the saccade targets. The monkey then maintained fixation onto the chosen target for an additional 0.75 s before *juice* delivery. The trial was aborted if the monkey broke center fixation at any time before the *go* signal, or after target acquisition but before *juice* delivery. Trials were separated by a 1.5 s inter-trial interval. Center fixation was imposed within 1°.

Sets of 0-10 squares indicating the offers were located around the (initially not visible) saccade targets, within 4.2° of visual angle. The side of each individual square subtended 1.05°. The spatial configuration of a given set of squares (for example, 3 squares on the right) remained the same throughout the experiments. The same color was associated with any given juice type throughout the experiments.

We used the following juices, listed in roughly decreasing order of preference (associated colors in parenthesis): high-sugar lemon Kool-Aid (dark yellow), grape juice (bright green), fruit punch (magenta), apple juice (diluted to 1/2 with water, dark green), cranberry juice (diluted to 1/3 or 1/4 with water, pink), water (white), milk (red), peppermint tea (bright blue), tea (light brown), low-sugar agua frescas Kool-Aid (light red), low-sugar tamarind Kool-Aid (dark brown), slightly salted water (0.60 g/l or 0.65 g/l concentration, light gray).

Recording sessions lasted typically 200-500 trials. The number of trials per offer type presented in each trial block was determined at the beginning of each session, and we generally tried to have more trials for offer types near the animal's (foreseen) indifference point. Typically, this resulted in 20-65 trials for each offer type in each recording session. The left/right configuration of each offer type was pseudo-randomized and counterbalanced within each trial block. Each day, monkeys completed 2-8 sessions (typically 5-6), with different pairs of juices used in different sessions.

For the current experiments, we wanted to have a stable relative value within any recording session. Indeed, the relative value was generally stable (though 3% of sessions were excluded from the analysis for unstable behavior). However, juice values could vary considerably from session to session within each day. In particular, early in the day monkeys were thirstier, and relative values tended to differ less from unity. Later in the day, monkeys generally became more selective and the relative value of less sweetened juices decreased. As a result, some variability was also present in the relative value measured for any given pair of juices in different days. For any given juice pair, the range of variability was typically twofold or less. The order of preference of different juices (i.e., their ranking) remained fairly stable across sessions and across days for both monkeys (though it was not exactly the same for the two monkeys).

Training lasted 6-8 months. During the training, monkeys initially indicated their choice by moving a bidirectional lever by hand. For most of the training, we used only water offered in variables amounts (e.g., 1 drop versus 2 drops). Monkeys spent most of the training learning non-specific aspects of the task, for example that objects on the monitor carried information about the to-be-received liquid, that a hand movement towards the right was associated to the visual stimulus placed on the right, and, later in training, that fixation should be maintained during the

delay. Monkeys were familiarized with each juice separately for 1-3 single-juice sessions. The most specific aspect of the task, namely the trade-off between juice type and juice quantity, did not require training, as monkeys showed it spontaneously and immediately. This was consistent with our previous observations in capuchins using solid foods in full sight[2]. For the current experiments, we wanted to have relative values away from unity. At the same time, we wanted relative values to lie within the tested range of offer types, so that, if the less preferred juice was offered in sufficiently large amount, the monkey would choose it. Prior to recordings, we tested numerous juice types to find suitable pairs.

Juices were delivered through a three-line juice tube. Each juice line was controlled by a separate solenoid valve. We routinely calibrated the three juice lines individually so that the valve-opening times corresponded to the desired multiple of juice quantum. During recordings, we used quanta of 80 µl and 65 µl for monkeys V and L, respectively.

Surgery and recordings

Under general anesthesia and aseptic surgery, we implanted a head-restraining device and a recording chamber on the skull of the monkeys, and a scleral eye coil[1]. In both monkeys we used a large oval custom-made chamber (main axes 50x30 mm). The chambers were centered on stereotaxic coordinates (A30, L0), with the longer axis parallel to a coronal plane. Thus, chambers covered the frontal lobes bilaterally. Following surgery, monkeys were given antibiotics (cefazolin, 20 mg/kg) and analgesics (buprenorphine, 0.005 mg/kg; flunixin, 1 mg/kg) for 3 days. Throughout the experiments, we strictly followed the NIH Guide for the Care and Use of Laboratory Animals and the guidelines of the Harvard Medical School Standing Committee on Animals.

For recordings, we used tungsten electrodes (125-µm diameter, 5±1 MΩ initial impedance, Frederick Haer & Co.). Electrodes were advanced with custom-built motorized micro-drives. The micro-drives, consisting of a titanium housing containing a 0-80 trapped screw, were anchored to a Teflon grid placed on the chamber. A brushless miniature precision gear motor (5.8-mm outer diameter, Micro Precision Systems) drove the trapped screw and was controlled remotely. The system allowed up to 0.5 µm of depth resolution. Typically four (and up to six) electrodes were used each day. We usually advanced electrodes by pairs (one motor for two electrodes), with the two electrodes placed 1 mm apart. Electrical signals from each channel were amplified and band-pass filtered (custom-designed miniature amplifiers, low frequency cutoff 400 Hz, high frequency cutoff 6 kHz) and recorded at 20 kHz by a dedicated processor (Power 1401, Cambridge Electronic Design). Action potentials were detected online by threshold crossing and waveforms were saved to disk for subsequent analysis. Spike sorting was performed semi-manually using commercially available software (Spike 2, version 5, Cambridge Electronic Design). We routinely used multiple algorithms, including template matching, clustering on waveform measurements, and principal component analysis.

Structural MRI scans (1-mm sections) were obtained for both monkeys before and after placing the recording chambers. Scans were performed in a 3.0 T magnet (General Electric) and three-dimensional reconstruction was performed off-line (Slicer-3D, www.slicer.org).

Recordings areas

Recordings were performed on the left hemisphere of monkey V and on the right hemisphere of monkey L. For their sulcal pattern, both hemispheres were classified as of type II of Chiavaras and Petrides[3].

The position of the chamber allowed us to reach OFC through nearly straight dorso-ventral penetrations. Because the chambers were large, we were able to record from an extended region of OFC. Based on exploratory recordings, we identified in both monkeys a region of interest where neuronal responses were modulated by the offer type in our task. In monkey V, the region of interest was centered in stereotaxic coordinates (A32.5, L–9.0) and extended for 6 mm rostro-caudally and 5 mm medio-laterally. In monkey L, the region of interest was centered in stereotaxic coordinates (A33.5, L8.5) and extended for 6 mm rostro-caudally and 2 mm medio-laterally. Based on the MRI scans and on the sequence of gray and white matter encountered during electrode penetrations, we located the regions of interest in the lateral bank of the medial orbital sulcus and in the medial part of the posterior orbital gyrus. Comparing our reconstruction to the architectonic subdivision of Carmichael and Price[4,5], we tentatively identified our regions of interest with area 13m.

All the neurons described here were recorded from the regions of interest. Apart from this criterion and apart from imposing good and stable electrical isolation, neurons were not otherwise selected prior to data

collection or during the analysis. We did not attempt to identify the cortical layer of recordings. However, because we reached OFC dorso-ventrally from the white matter, we presume that the lower layers are largely represented in our dataset.

Analysis of choice pattern and 3-way ANOVA

Unless otherwise specified, we always assume linear value functions. Indicating with V(x) the value of x, the linear assumption means that a multiple q of a unit volume of juice X has a value V(qX) = q V(X).

Data are analyzed in Matlab (PC version 6.5, Math-Works). In each session, we fit the choice pattern with a sigmoid function, which for our data generally provides an excellent fit (R^2>0.95). The flex of the sigmoid determines the relative value n* such that V(A) = n* V(B). From this equation, we can compute the value of any amount of juice A and B, up to a scaling factor. For example, expressing values in units of V(B), the value of b drops of juice B is b, and the value of a drops of juice A is n*a. Because the analysis of neuronal responses in relation to the juice values is based on linear regressions (see below), the results do not depend on the particular units used to express value.

For quantitative analyses, neuronal firing is averaged over the following time windows: pre-*offer* (0.5 s before the *offer*, a control time window); post-*offer* (0.5 s after the *offer*); late delay (0.5-1.0 s after the *offer*); pre-*go* (0.5 s before the *go*); reaction time (RT; from the *go* to the saccade); pre-*juice* (0.5 s before the *juice*); and post-*juice* (0.5 s after the *juice*).

We first analyze single-trial data from each neuron with a 3-way ANOVA (factors: [position of juice A] x [movement direction] x [offer type]), separately in each time window. We impose a significance level of p<0.001. We choose a relatively conservative threshold because of the large number of responses analyzed (931 neurons x 7 time windows). Because factors [position of juice A] and [movement direction] are rarely significant, we collapse data across these two factors in all subsequent analyses. We also use the results of the ANOVA as a screening criterion for subsequent analyses, which (unless otherwise indicated) we restrict to neurons and time windows for which the factor [offer type] yields a significant effect.

We divide trials into "trial types" based on the offer type and the choice. For example, a monkey facing the offer type 3B:1A can choose either 1A or 3B, corresponding to the two trial types (3B:1A, 1A) and (3B:1A, 3B). In principle, there are twice as many trial types as offer types in any given session. However, many trial types are "empty," because for most offer types the choice of the monkey is univocal (for example, in a given session, a monkey offered 3B:1A might always choose 3B). In subsequent analyses, we include only trial types with two or more trials.

Neuronal activity is averaged across trials separately for each trial type. Hereafter, the term "response" refers to the average activity of one neuron in one time window, as a function of the different trial types. Note that, for sake of clarity, in figures 2d and 3a-e of the main text we grouped trials by offer type, not by trial type.

Neuro-econometric analysis: defined variables and correlation matrix

What variables are encoded in OFC? As illustrated in the main text (figure 3), many responses seem qualitatively related to the variables *chosen value*, *value A offered*, *value B offered*, and *taste*. For a quantitative analysis, however, we want to consider alternative hypotheses, namely other variables with which neuronal responses could a priori correlate. For example, we want to test the hypothesis that some responses might encode the *other value* (i.e., the value of the non-chosen juice), or the *total value*, or the *value A chosen*; etc. In addition, we want to test variables that might capture the decision process, such as the value difference *(chosen–other) value* and the value ratio (*other/chosen) value*. Most importantly, we want to distinguish between variables related to *value* and variables related to physical properties of the juice/stimulus, such as *quantity*, *volume*, *number*, etc. These physical properties are all proportional to each other, and we collectively refer to them with the variable *number*.[(a)] To put *value* and *number* on equal

[a] In our experiments, *value* is subjective but operationally-defined. The neuronal correlates of *value* and *number* cannot be dissociated easily using a single good (e.g., a single juice) because in that case the two variables are inextricably inter-related (assuming a linear value function, *value* and *number* are in fact proportional to each other). Introducing uncertainty does not help, because the confusion remains between *expected value* and *expected number*. However, *value* and *number* can be dissociated using two different juices, insofar as the relative value of the two juices (i.e., the ratio V(A)/V(B)) differs from unity. For example, in a session in which 1A=3B, a smaller *number* (say 2A) may have higher *value* than a larger *number* (say 4B). Of course, a correlation between *value* and *number* always remains (after all, a large quantity of any given juice is always more valuable than a small quantity of that same juice). For this reason dissociating definitively the neuronal correlates of *value* and *number* ultimately requires the "neuro-econometric" analysis described below.

footing in the analysis, we want to test in the number domain every variable tested in the value domain. In summary, we want to analyze the 19 variables defined in figure s1.

These 19 variables are often highly correlated with each other. To estimate the typical correlation between any two variables X and Y, we proceed as follows. For each session, we compute the correlation coefficient:

$$\underset{session}{\rho}(X,Y) = \vec{x}\cdot\vec{y}\Big/\sqrt{\vec{x}^2\cdot\vec{y}^2}$$

where $\vec{x}$ and $\vec{y}$ are vectors of values taken by variables X and Y for different trial types. So defined, the correlation coefficient varies between –1 and +1. Most informative in this context is the absolute value, which we compute and average across sessions:

$$\rho(X,Y) = \left\langle \left| \underset{session}{\rho}(X,Y) \right| \right\rangle_{sessions}$$

Repeating for all pairs of variables, we obtain a symmetric matrix ρ of elements $\rho(X,Y)$ that vary between 0 and 1. Hereafter, we refer to ρ as the "correlation matrix."

The 19 variables defined in figure s1 cast a wide net. But ultimately we would like to test whether *few* variables can capture the neuronal activity of OFC, and identify those that best do so. To achieve this goal, we proceed in two steps. First, we assume that individual OFC responses encode each only one variable and that value functions are linear. We perform independent linear regressions of each response on each variable, and we apply methods of variable selection. We thus identify *value A offered*, *value B offered*, *chosen value*, and *A|B chosen* as the variables that best account for the dataset. These variables explain well the large majority of OFC responses. Second, having identified for each response the variable encoded "at the first order," we relax the two initial assumptions. We test with standard multi-regression methods whether adding a second variable or a quadratic value term improves the regression, and we observe that for the large majority of responses this is not the case. In other words, both the "one response-one variable" assumption and the assumption of linear value functions are, by and large, warranted. We conclude that indeed OFC responses encode the variables identified by the variable-selection procedures.

The next sections and the Supplementary Results detail the methods and results of these analyses.

Neuro-econometric analysis: variable selection

The variable-selection analysis is based on the assumption that individual responses encode each (at most) one variable. For this analysis, we regress each response separately on each of the 19 variables. Each regression provides a slope and the corresponding R^2. We consider a variable capable of "explaining" a given response if the regression slope is significantly different from zero ($p<0.05$), and we conventionally set $R^2 = 0$ for variables that do not explain the response. In general, any given response can be explained by multiple variables. Largely, this is because variables are often highly correlated with each other (for example, this is the case for variables *chosen value*, *total value* and *chosen number*), a situation referred to as multi-collinearity[6,7].

In the presence of multi-collinearity, it is often possible to identify a small subset of variables that account for much of the data. However, the problem of identifying the appropriate subset is in principle not trivial. For our analysis we adapt two methods of variable selection routinely used in the case of multi-linear regressions, namely the "stepwise selection" method and the "best-subset" method[6,7]. Notably, our situation differs from that typically found in multi-linear regression, where any single response depends on multiple variables. Each of our responses is sampled in few data points (typically 8-10 trial types), many fewer than the number of variables we want to test (19 variables). However, we can capitalize on the large number of responses available (a total of 1379 responses pass the ANOVA screening criterion) and identify the most relevant variables through a population analysis (see below).

In addition to the 19 original variables, we define two "collapsed" variables *value A|B offered* and *value A|B chosen*, as follows. The variable *value A|B offered* is taken to explain a given response if at least one of the two variables *value A offered* and *value B offered* explains the response. The R^2 of the collapsed variable is equal to the higher R^2 obtained from of the two original variables (and equal to zero if none of the original variables explains the response). Similarly, the variable *value A|B chosen* is defined by collapsing variables *value A chosen* and *value B chosen*.

Neuro-econometric analysis: second order encoding

To explore the possibility that individual OFC responses might encode a mixture of variables, we proceed formally as follows[8].

Consider a response encoding at the first order the variable X with $R^2 = R_X^2$ (i.e., a response explained by X better than by any other selected variable, with $R_X^2 > 0$). To establish whether adding a second variable Y to the regression provides a significantly better account, we compute

$$F_{Y|X} = (n-3) * (R_{XY}^2 - R_X^2) / (1 - R_{XY}^2)$$

In the equation, R_X^2 is obtained from the linear regression on X only; R_{XY}^2 is obtained from the bi-linear regression on X and Y; and n is the number of trial types (data points in the regression). We compute $F_{Y|X}$ for each of the variables Y we wish to test as potentially encoded at the second order, and we take the maximum $F = \max\{F_{Y|X}\}$. The degrees of freedom of F are 1 for the numerator and $n-3$ for the denominator. We then set a threshold F^* corresponding to a desired p* = 0.01 (we choose this threshold because residuals are not pre-screened and because we test each response with multiple potential second order variables). If F passes the criterion, we identify the second order variable encoded by the response. If F does not pass the criterion, we conclude that the response does not encode any second order variable (at least among those tested).

For second order encoding, we test some of the variables tested for first order encoding. In addition, we test value-encoding responses with quadratic value terms. Hence, with this analysis we scrutinize the validity of the two assumptions underlying the variable selection procedures, namely that individual OFC responses encode each only one variable and that value functions are linear.

Analysis of time course

To analyze the time course by which neurons in OFC encode variables *value A offered*, *value B offered*, *chosen value*, and *A|B chosen*, we define 50-ms non-overlapping time bins, aligning trials separately at the time of the *offer* and at the time of *juice* delivery. For each 50-ms time bin, we subject each cell in the population of N_{tot} = 931 neurons to independent linear regressions on the four selected variables. We assign a neuron to a variable if the regression slope differs significantly from zero in that time bin. If more than one variable has a non-zero slope (a rare case), we assign the neuron to the variable with highest R^2. For this analysis, we do not screen neuronal data prior to regression, because small time bins result in spike counts mostly equal to 0 or 1. For this reason, we choose a slightly more conservative threshold (p<0.01) to identify regression slopes as significantly non-zero. In each time bin, the number of neurons expected to be assigned by chance to each class is N_{chance} = p* N_{tot} = 9.31. For each time bin, we then compute:

$$N_{value\ A|B\ offered} = N_{value\ A\ offered} + N_{value\ B\ offered} - N_{chance}$$

Supplementary Results

Neuronal database, linear regressions, and correlation matrix

We recorded the activity of 931 neurons (375 from monkey V; 556 from monkey L). Figure s2 reports the number of cells for which each main factor in the ANOVA has a significant effect. In particular, the factor [offer type] yields a significant effect in at least one time window for 505 (54%) neurons in total (211 = 56% from monkey V; 294 = 53% from monkey L). We refer to these as "task-related" cells. Figure s3 shows the average neuronal activity separately for task-related cells and for "other" (i.e., not task-related) cells. It can be observed that the average activity profile of task-related cells peaks early after the *offer*, slowly decays during the delay, and has two secondary peaks before and after *juice* delivery. In contrast, the average activity profile of other cells is lower and essentially flat throughout the trial.

A total of 1379 responses are significantly modulated by [offer value] in the ANOVA (656 from monkey V; 723 from monkey L). We further analyze these responses in relation to the variables defined in figure s1. In total, 1227/1379 (89%) responses are explained by at least one of the 19 variables (567/656 = 86% from monkey V; 660/723 = 91% from monkey L). Methods for variable selection are applied to this dataset.

Figure s4 illustrates the results of individual linear regressions obtained for one particular response (i.e., the activity of one neuron in one time window as a function of the different trial types). The two top left panels show the behavioral choice pattern and the neuronal response plotted with respect to offer type. In the other panels, the same neuronal response is plotted against each of the 19 variables. A blue regression line indicates that the slope is significantly non-zero (p<0.05); if so, the respective R^2 is indicated in the top left corner of the panel. In the top left panels, the activity of the cell as a function of the offer type has a U-shaped profile, qualitatively similar to that

hypothesized for a response encoding the *chosen value*. In fact, the variable *chosen value* explains the response (R^2 = 0.90). However, the response is also explained by several other variables, for example *total value* (R^2 = 0.72), *chosen number* (R^2= 0.50), etc.

The fact that multiple variables can explain the same response is not surprising, since the variables are in general not independent. This point is illustrated in figure s5, which shows the correlation matrix. In the figure, different shades of gray represent different values of correlation (ρ) with black corresponding to $\rho = 1$ and white corresponding to $\rho = 0$. A small white circle or cross indicates $\rho > 0.8$. For example, the variables *chosen value* and *total value* are highly correlated with each other and with other variables (e.g., *total number*, *max number*, and *chosen number*)—a case of multi-collinearity.

Ideally, to provide a concise description of the dataset, we would like to achieve two goals: first, to assign each response to only one variable; second, to identify a small subset of variables that can explain as many responses as possible. Clearly, achieving the second goal helps to achieve the first. But how can we select the "right" subset of variables? In the following sections, we first make a qualitative case for variable selection and then we illustrate the results obtained with statistically principled procedures.

Variable selection: Qualitative analysis

Consider the response analyzed in figure s4. Our initial hypothesis that U-shaped responses might reflect the *chosen value* is bolstered by the fact that the variable *chosen value* provides the best fit (i.e., the highest R^2). However, at this level of analysis, it would be unreasonable to rule out other variables (for example *total value*) that provide a slightly lower R^2. A potentially powerful approach is to consider the entire population of responses. For example, we might discover that whenever (or most often when) both variables *chosen value* and *total value* provide a non-zero slope, the variable *chosen value* provides a slightly higher R^2. If that were the case, we could conclude that neuronal responses genuinely encode the *chosen value*, and that the variable *total value* has no additional explanatory power. This argument is obviously complicated by the fact that, as we observed qualitatively (figure 3, main text), not all responses in our dataset encode the same variable. In addition, variables have more than just pairwise correlation and cannot be simply considered two at the time. So we must analyze the entire population of responses and all the variables at once.

Figure s6 illustrates two complementary ways to derive a population analysis from the individual linear regressions. In figure s6a, we compute for each time window and for each variable the number of responses explained. For example (top left), in the post-*offer* time window, the variable *total value* explains 130 responses. Because, as we noted, more than one variable may explain any given response, any given response may appear in multiple bins in this plot. Some trends emerge clearly. First, as indicated by the ANOVA, more responses are modulated by the trial type in early time windows (post-*offer* and late delay) and late time windows (pre-*juice* and post-*juice*), as compared to the peri-movement time windows (pre-*go* and RT). Second, focusing on the post-*offer* time window we observe that a group of variables explain a large number of responses. For example, the variables *total value*, *chosen value*, *total number*, *max number*, *chosen number*, *(max–min) number*, and *value B offered* all explain more than 100 responses. Inspection of figure s5 reveals that these variables are highly inter-correlated. So the picture in figure s6a clearly contains a high degree of redundancy, which we may hope to resolve with further analysis. Third, in addition to these variables, which are prevalent in the post-*offer*, pre-*juice* and post-*juice* time windows, there is another group of variables, including *(chosen–other) number*, *A|B chosen*, and *value B chosen*, which are common in the pre-*juice* and post-*juice* time windows, but not in earlier time windows. Again, inspection of figure s5 reveals that these variables are highly inter-correlated, a redundancy that we may hope to resolve with further analysis.

A complementary way to derive a population analysis from the individual regressions is to consider for each response only the variable providing the best fit (i.e., the highest R^2). By doing so, we essentially force each response in only one bin. Figure s6b shows the result of this analysis. Note that much of the redundancy of figure s6a seems naturally resolved. For example, many more responses are best fit by the variable *chosen value* than by any of the variables *total value*, *total number*, etc. that are highly correlated with *chosen value*. Likewise, many more responses are best fit by the variable *A|B chosen* than by either *(chosen–other) number*, or *value B chosen*. Thus, although the population picture obviously remains complex, figure s6b suggests that fewer variables than the 19 initially considered may be sufficient to account for most responses.

In some sense, the pictures presented in figures s5a and s5b represent two opposite and extreme ways to analyze the population of responses. On the one hand, in figure s6a we forgo the possibility of ranking the quality of the linear fits (i.e., the R^2) and, by doing so, we give up the possibility of a concise description. In fact, it could be argued that we spuriously added redundancy by including "unreasonable" variables in the analysis. And, indeed, one could easily come up with more variables that would further complicate the picture, to little or no advantage. On the other hand, in figure s6b we ignore that a given variable may sometimes provide a satisfactory explanation for a given response, even if another variable provides a better fit. That is clearly excessive: because neuronal data are noisy, responses that genuinely encode a given variable (say *chosen value*) might at times be best fit by another, highly correlated variable (say *total value*). In other words, it is likely that some responses that would genuinely belong to one bin in figure s6b "spilled over" to another bin because of noise. So here too it could be argued that we spuriously added complexity to the picture by including "unreasonable" variables in the analysis. And again, one could certainly further complicate the picture by adding more variables, to little or no advantage.

In conclusion, an ideal account of the dataset would on the one hand make use of the information of figure s6b, namely that some variables (say *chosen value*) do consistently better than others (say *total value*) in fitting the population of neuronal responses. On the other hand, an ideal account would also use the information embedded in figure s6a, namely that one variable might provide a satisfactory (albeit suboptimal) explanation for many responses. But how can we identify a small subset of variables to explain a large number of responses? This problem closely resembles that of variable selection encountered in multi-linear regressions in the presence of multi-collinearity. Methods commonly used for multi-linear regressions[6,7] can be adapted to our case.

Variable selection: Stepwise method

One simple way to handle our problem is to select variables one at the time. First, we select the variable that provides the highest number of best fits (the darkest bin in figure s6b). In our case, this variable is *A|B chosen*. We explore the entire dataset and we remove all the responses that can be explained by this variable. Then we iterate the procedure by selecting a second variable, then a third variable, etc. We continue the procedure as long as any newly selected variable explains at least a certain percentage (for example 5%) of otherwise unexplained responses. At the end of the procedure, we classify responses based on the R^2 (i.e., we assign responses explained by more than one selected variable to the variable with highest R^2). This method is called "stepwise selection."[6,7 (b)]

We apply the stepwise selection method to our dataset. In figure s7, panels on the left represent the population of unexplained responses at different iterations of the procedure. Tables of best fit are shown, so that the top panel is the same as shown in figure s6b. For each iteration, a red asterisk indicates the selected variable; blue dots indicate variables excluded by the 5% selection criterion. The first four iterations select variables *A|B chosen*, *chosen value*, *value A offered*, and *value B offered*, and no other variable reaches the 5% selection criterion in subsequent iterations.

In the analysis shown in figure s7, we keep separate the 19 variables defined in figure s1. However, it could be argued that in some cases different variables really represent the same *class* of response. For example, the distinction between *value A offered* and *value B offered* is somewhat arbitrary, because one particular juice may be labeled "A" in a given session and "B" in another session. Furthermore, if a selection procedure identified only one of the two variables, the result would be difficult to interpret. Thus it is probably more correct to combine the variables *value A offered* and *value B offered* using the "collapsed" variable *value A|B offered* (see Supplementary Methods). Likewise, we can use the other collapsed variables *value A|B chosen*. This leaves us with 17 variables. We re-analyze our dataset using collapsed variables. The stepwise procedure selects the three variables *value A|B offered*, *chosen value*, and *A|B chosen*, confirming our previous result.

In summary, the stepwise method applied to our dataset selects three variables: *chosen value*, *value A|B offered*, and *A|B chosen*. These three variables explain

[b] The stepwise selection method is used with the following *caveat*. The "marginal explanatory power" of one particular selected variable X is the number of responses that are explained by X and that are not explained by any other selected variable. The 5% selection criterion sets a threshold on the marginal explanatory power of *newly* selected variables. However, the marginal explanatory power of a selected variable X generally drops over iterations, because variables selected after X may explain some of the responses previously explained only by X. Eventually, it is possible that a selected variable X only explains a small percentage (less than 5%) of otherwise unexplained responses. For a correct procedure, we check *at each iteration* that all of the selected variables actually meet the 5% selection criterion, and we exclude previously selected variables that fail to meet the criterion.

1085 responses, corresponding to 79% of all responses significantly modulated by offer type (1379), and to 88% of responses explained by all 17 variables (1227). In other words, in limiting ourselves to three variables, we "lost" only 12% of responses. Having thus identified the variables encoded in OFC, we classify responses based on the R^2 (i.e., we assign responses explained by multiple selected variables to the variable with highest R^2). Figure s8 (top) summarizes the final result of this classification. Note that, to provide a more intuitive interpretation of the variables, in the main text we renamed variable *value A|B offered* as *offer value* and variable *A|B chosen* as *taste*. The histograms in figure s8 (bottom) illustrate the distribution of R^2 obtained as a result of the final classification. In general, selected variables capture the variability of individual responses remarkably well (mean $R^2 = 0.63$).

In the following sections, we show that alternative procedures for variable selection yield essentially the same results.

Variable selection: Best-subset method

One potential limitation of the stepwise method for variable selection is that the results obtained may be "path-dependent," and the method is not guaranteed to select the best possible subset of variables. This goal can be achieved using the "best-subset" method[6,7].

The idea of the best-subset method is to compute for each possible subset of d variables the corresponding number of responses explained; to identify the best subset of d variables as the subset that explains the maximum number of responses; and to repeat the procedure for $d = 1, 2, 3, \ldots$ If $n(d)$ is the number of responses explained as a function of d, the number d^* of variables necessary to characterize the population can be determined either by an "elbow" in the function $n(d)$, or with a threshold criterion (e.g., imposing that at least 85% of responses be explained).

Figure s9 shows the results of analyzing our dataset with the best-subset method. In the top panel, the x-axis represents the number d of variables included in each subset, and the y-axis represents the percentage $n(d)$ of responses explained by the best subset. The bottom table indicates the variables included in the best subset for various d. Several points should be noted. First, although $n(d)$ does not present a clear "elbow," selecting only three variables seems reasonable, as additional variables add limited explanatory power. Second, the three variables identified by the stepwise selection method, namely *chosen value*, *value A|B offered*, and *A|B chosen*, are indeed the best possible subset of three variables. Third, these three variables are also included in the best subset of four variables and in the best subset of five variables. This is not necessarily expected, because in general the best subset of d+1 variables might not include all or any of the variables included the best subset of d variables. The fact that the three variables that make up the best subset for $d = 3$ are also included in the best subset for $d = 4$ and for $d = 5$ is a sign of robustness of the result.

In conclusion, the variable-selection analysis using the best-subset method confirms the results obtained with the stepwise selection method. In both cases, the three selected variables are *chosen value*, *value A|B offered*, and *A|B chosen*.[c, d]

Variable selection: Over-fitting and post hoc analysis

Although it guarantees optimality, the best-subset procedure does not provide a measure of reliability of the result. For example, in our case, the method indicates that, among the subsets of three variables, *chosen value*, *value A|B offered*, and *A|B chosen* explain the highest number of responses, but we do not know how well these three variables do compared to the possible alternatives. In other words, we might have a problem of over-fitting. In multi-linear regressions, the standard way to handle this problem is to repeat measures and to analyze multiple datasets[6,7] (see below). But in addition to repeating measures, our case also lends itself to an informative post hoc analysis.

In the post hoc analysis, we want to test selected variables against highly correlated but discarded variables. For example, the best-subset method selects the variable *chosen value*. We want to establish whether selecting this variable is *significantly* better than selecting the variable *total value*, which is highly

[c] The stepwise selection method and the best-subset method differ in how different time windows are analyzed. Using the stepwise method, responses from different time windows are analyzed separately but in parallel. Variables are selected for the number of best fits by time window, but variables can then explain responses in any time window. In contrast, using the best-subset method, responses from different time windows are pooled together. Keeping time windows separate may or may not be viewed as a desirable feature. In any case, it is reassuring that the two methods provide converging results.

[d] One important advantage of the best-subset method is that results do not depend on "irrelevant" variables. That is to say, including in the analysis variables that turn out to have little explanatory power has no effect on the outcome. This ensures that we don't "add noise" to the procedure by testing "unreasonable" variables.

correlated with *chosen value* and is therefore a presumably "challenging" alternative. That *chosen value*, and not *total value*, is part of the best subset ultimately means that *chosen value* has a higher marginal explanatory power: disregarding responses explained by either *value A|B offered* or *A|B chosen*, the number of responses explained by *chosen value* and not by *total value* is greater than the number of responses explained by *total value* and not by *chosen value*. To establish whether this inequality is true in a statistical sense, we can formally proceed as follows.

Consider the sub-population of responses explained by *chosen value* (variable X) and/or by *total value* (variable Y), and not explained by any other variable in the best subset (i.e., *value A|B offered* and *A|B chosen*). Of this sub-population, a number nX of responses are explained by X and not by Y; a number nY of responses are explained by Y and not by X; and the other responses are explained by both X and Y. The best-subset method indicates that nX>nY, but it is possible that nX is actually quite close to nY. The problem of determining whether the inequality nX>nY is statistically significant is equivalent to asking how unlikely it is to draw nX heads out of nX+nY tossings of a fair coin, and can be addressed with a simple binomial test. For this particular case, we have nX = 58 and nY = 21, from which we infer that the marginal explanatory power of *chosen value* is significantly higher than that of *total value* ($p<10^{-5}$).

We repeat this procedure for all the pairs of variables (X,Y) for which X is a variable in the best subset and Y is a variable highly correlated to X (i.e., $\rho(X,Y) > 0.8$; see figure s5). Additionally, we test *chosen value* against *chosen number*. The results are presented in figure s10. In these pairwise comparisons, all the variables included in the best subset do significantly better than the challenging alternatives (maximal $p<10^{-5}$). The only exception is that the variable *A|B chosen* fails this test against the variable *value A|B chosen* (p = 0.12). As discussed below, this degree of ambiguity remains largely unresolved in our analysis. For all other aspects, the post hoc analysis confirms our previous conclusions.

Variable selection: Other procedures and conclusions

We employed a number of alternative procedures to select variables. Essentially, they all confirm the results described in previous sections. For example, in the spirit of repeating measures, we find that data from the two monkeys analyzed separately yield statistically indistinguishable results. Another variant consists in weighting the contribution of each response to the explanatory power of any variable with the corresponding R^2. The variable selection analyses using this alternative metrics yields results identical to the ones obtained with the binary "explained/not-explained" metrics. Finally, correcting regressions for unequal variance also provides identical results.[e] Using collapsed variables, all these procedures consistently indicate that the best subset includes *value A|B offered* and *chosen value*, and either *A|B chosen* or *value A|B chosen*, with these two latter variables having statistically indistinguishable marginal explanatory power. Using non-collapsed variables, all the procedures provide statistically consistent results. However, using non-collapsed variables, we find that the marginal explanatory power of *A|B chosen* is significantly higher than that of either *value A chosen* or *value B chosen*.

Although the issue between *A|B chosen* and *value A|B chosen* remains largely unresolved, two arguments seem to favor the former variable. First, in the analysis of non-collapsed variables, *A|B chosen* does significantly better than either *value A chosen* or *value B chosen* taken alone Second (and partly related), *A|B chosen*, with only one intrinsic degree of freedom (i.e., the relative value of the two juices), is more parsimonious than *value A|B chosen*, which has two intrinsic degrees of freedom (i.e., the relative value, and which one of the two juices is coded for). In summary, although these arguments are largely heuristic, it seems preferable to summarize our data in terms of the variable *A|B chosen*.[f]

[e] The least-squares method, used here for all regressions, is based on two assumptions: gaussianity (i.e., data must come from gaussian distributions) and homoscedasticity (i.e., the gaussians must all have the same variance). Cortical single-trial spike counts are known to violate both these assumptions [ref 9]. Responses analyzed here (i.e., averages across trials) approximately satisfy gaussianity (for the central limit theorem). However, responses generally do not satisfy homoscedasticity. In cases of unequal variance (heteroscedasticity), the least squares method can be corrected by weighting the residual associated to each data point m_i with a term proportional to $1/\sigma_i$, where σ_i is the standard deviation of the distribution from which m_i is drawn [ref 8]. In our case, m_i is the response measured for a particular trial type and we can estimate σ_i with the standard error. Notably, $1/\sigma_i$ is proportional to the square root of the number of trials. This suggests that, in our case, it might be preferable *not* to correct for unequal variance, because trial types close to the indifference point, which are in many respects the most informative, are also those for which we have fewer trials (because monkeys "divide" their choices between the two juice types). For this reason, we generally use uncorrected linear regressions in all our analyses.

[f] Consistent with the results of a recent study of gustatory responses in this brain area (Pritchard et al., 2005), in our final classification, we observe *A|B chosen* (i.e., *taste*) responses in

In conclusion, having tested a large number of hypotheses with a variety of variable selection procedures, we identified *value A|B offered*, *chosen value*, and *A|B chosen* as the three variables encoded by OFC responses. These three variables describe our dataset well, and significantly better than challenging alternatives.

Second order encoding

The variable selection analysis presented in previous sections rests on two key assumptions: the assumption that OFC responses encode each only one variable and the assumption that value functions are linear. We now put these two working hypotheses under scrutiny.

For second order encoding, we test variables selected at the first order (that is, we test whether responses encode mixtures of selected variables). We also test the two variables *value A chosen* and *value B chosen* that were excluded last in the variable selection analysis, as well as the variable *chosen number*. Finally, for responses encoding at the first order *value A offered*, *value B offered*, and *chosen value*, we test the corresponding quadratic term *(value A offered)*^{2}, *(value B offered)*^{2}, and *(chosen value)*^{2}. We extend this analysis to the population of 1085 responses classified at the first order (figure s8) and we pool together responses from different time windows. The results are presented in figure s11.

In the two tables, rows indicate first order variables and columns indicate second order variables. The rightmost column indicates responses that do not encode any second-order variable. The top table (figure s11a) reports the number of responses. The bottom table (figure s11b) reports the same data as percentages (normalized by the number of responses explained by each first-order variable). In general, we observe that the vast majority of responses do not encode second order variables, independently of the variable encoded at the first order. Over the entire population, 837/1085 (77%) responses do not encode second order variables. If we exclude from this analysis the quadratic value terms, we find that 890/1085 (82%) responses do not encode second order variables. We interpret this result as a substantial justification for the one response-one variable assumption. Similarly, for most value-encoding responses, adding a quadratic term does not improve the regression significantly. Repeating the analysis with only quadratic terms as potential second order variables, we find that in 757/817 (93%) cases quadratic terms fail to provide a statistically appreciable gain. Again, we interpret this result as a justification for the assumption of linear value functions.[g]

In summary, the analysis of second order encoding provides *post facto* support for the two main assumptions underlying the variable selection analysis, namely the one response-one variable assumption and the assumption of linear value functions. This result concludes our "neuro-econometric" analysis.

Relationship between slope ratio and relative value

During the experiments, we used a large number of juice pairs. In each session we measured the relative value n* from the behavioral pattern of choice, through a sigmoid fit. Naturally, we measured different relative values for different juice pairs. For example, monkeys generally had a mild preference for grape juice over fruit punch (typically, $n^*<2$), but a strong preference for grape juice over peppermint tea (typically, $n^*\geq 3$). In addition, the relative value of *any given juice pair* could vary from session to session and from day to day. For example, the relative value of ½ apple juice over peppermint tea varied over many days in the range $n^*\in(1.5,3)$. If U-shaped responses indeed encode the *chosen value*, we should expect them to reflect this variability. The following analysis confirms this prediction.

To avoid any bias, we want to identify U-shaped responses independently of the specific variable that they might encode. We proceed using a bi-linear regression:

$$fr = a_0 + a_A(\#A) + a_B(\#B) \qquad (1)$$

where fr is the firing rate of the neuron, and (#A) and (#B) are, respectively, the number of drops of juice A and juice B chosen by the monkey. (Note that in any given trial either (#A) = 0 or (#B) = 0.) We then define a response to be "U-shaped" if the regression slopes a_A and a_B are both significantly different from zero ($p<0.01$). (For this analysis, we do not pre-screen responses with the ANOVA.)

168/931 (18%) of recorded neurons, most prevalently at the time of juice delivery (i.e., in pre-*juice* and post-*juice* time windows).

[g] Of course, the presence and nature of quadratic value terms is a consequential issue in economic theory, where convexity of the value function plays a fundamental role. The present results indicate that the assumption of linearity is adequate for our dataset and at the relatively coarse level of our analysis. However, the present results do not exclude the possibility that a consistent departure from linearity may emerge upon more refined examination, or in experiments that span a wider range of values.

As illustrated in figure s12 for one particular response, the regression slopes a_A and a_B are in general different from each other. Since both $a_A \neq 0$ and $a_B \neq 0$, Eq.1 can be re-written as follows:

$$fr = a_0 + m\left(k^* (\#A) + (\#B)\right) \qquad (2)$$

with $m = a_B$ and $k^* = a_A/a_B$. The hypothesis that U-shaped responses encode the *chosen value* leads to a simple prediction regarding the slope ratio k*, which should be (statistically) equal to the relative value of the two juices. Critically, k* should co-vary with n*.

At least for the response in figure s12, the relationship $k^* \approx n^*$ holds true. Indeed, from the bi-linear regression we obtain $k^* = 3.0$ (±1.4) (±s.d.), while from the behavioral choice pattern we obtain $n^* = 3.2$. Applying the bi-linear regression criterion to all 931 neurons, we identify 254 U-shaped responses. Figure s13 illustrates the relationship between k* and n* observed for this population. The scatter plot in figure s13a includes data from all sessions, pooling together all pairs of juices. In the plot, the x-axis represents the behaviorally measured relative value n*, the y-axis represents the neuronal slope ratio k*, and both axes are plotted in log scale. Each dot represents one response. Defining "A" as the preferred juice is equivalent to imposing $n^* > 1$. In principle, the slope ratio k* could assume any possible value. However, nearly always $k^* > 1$. Moreover, we observe a significant correlation between k* and n* ($p<10^{-12}$).

Most importantly, the relationship $k^* \approx n^*$ can be observed when the analysis is restricted to responses recorded with individual pairs of juices, as shown in figure s13b for one particular pair of juices (A = ½ apple, B = peppermint tea). During the experiments, We used a total of 25 different juice pairs. The number of U-shaped responses recorded with any given juice pair varied between 1 and 40. Restricting our analysis to the 7 pairs of juices for which we have >10 responses, we test whether the relationship $k^* \approx n^*$ holds true using the regression function:

$$k^* = b_0 + b_1 n^* \qquad (3)$$

For the pair of juices shown in figure s13b, the regression indicates $b_0 = 0.08$ and $b_1 = 1.18$, with 95% confidence intervals $b_0 \in (-0.21, 0.37)$ and $b_1 \in (0.77, 1.59)$. Congruent results are obtained for all 7 pairs of juices. Figure s13c illustrates in particular for different juice pairs (y-axis), the value of the regression slope b_1 (x-axis), together with the 95% confidence intervals. Notably, the values of b_1 are distributed around the red dashed lined corresponding to $b_1 = 1$. Averaging across the 7 juice pairs, we obtain $\langle b_1 \rangle = 1.05 \pm 0.15$ (mean ± s.e.m.). With respect to the intercept, averaging across juice pairs we obtain $\langle b_0 \rangle = -0.13 \pm 0.15$. These results are consistent with the predicted identity $k^* = n^*$.

In summary, the neuronal "U" shapes recorded in OFC closely match the behavioral choice pattern on a juice-by-juice and session-by-session basis, a phenomenon naturally captured by the concept of economic value.

Ingredient-based hypothesis

One concern is whether U-shaped responses, which vary with the quantity of both juices A and B chosen by the monkey, may simply encode the quantity of one particular ingredient, or combination of ingredients, present in both juices. By "ingredient," we mean any compound contained in the juice, for example a compound that would elicit a taste response (e.g., water, sugar, citric acid, etc.)[10-13]. The strongest argument against the ingredient-based hypothesis follows from the relationship $k^* \approx n^*$ found in the previous section. To appreciate this point, let us refer to the cartoon shown in figure s14.

To summarize the results of the previous section, we showed that for *any two juices* A and B, in the scatter plot of k* versus n*, data lie on a diagonal line (figure s14, gray line) and cannot be described by a horizontal line. We can now examine different variants of the ingredient-based hypothesis.

The first variant is the hypothesis that U-shaped responses *all* encode the quantity of one particular ingredient. If this is true, given the ingredient and two juices A and B, the two regression slopes a_A and a_B should be proportional to the concentrations ρ_A and ρ_B at which the ingredient is present in the juices. Therefore, the slope ratio k* should be equal to the concentration ratio ρ_A/ρ_B, independently of the relative value n* recorded in the session. Consider for example the ingredient water. Because for any amount of juice the quantity of water is equal to the juice volume, if U-shaped responses all encode the quantity of water, neuronal data should lie on the horizontal line $k^* \approx 1$ (figure s14, blue line), contrary to what we observe. In other words, the sole fact that neuronal data overwhelmingly lie in the quadrant $\{k^*>1, n^*>1\}$ rules out the possibility that U-shaped responses all encode the quantity of water consumed by the monkey. (This is an alternative way to rule out the variable *chosen number*.) Consider now another ingredient, for example sugar, which we may assume to be present in given juices A and B at a certain concentration ratio

ρ_A/ρ_B. If U-shaped responses all encode the quantity of sugar chosen by the monkey, we should observe neuronal data lying on the horizontal line $k^* \approx \rho_A/\rho_B$, independently of the relative value n* recorded in the session. However, our observations contradict this prediction. More generally, no matter what ingredient one may want to test, the same argument can always be made. If OFC responses all encode that particular ingredient, given two juices A and B, neuronal data should always lie on a horizontal line corresponding to the concentration ratio, contrary to our observations. In conclusion, the simple hypothesis that U-shaped responses all encode one single ingredient can be rejected.

A second variant of the ingredient-based hypothesis is that U-shaped responses all encode a particular linear combination of multiple ingredients, such as water, sugar, etc. However, any linear combination of horizontal lines in figure s14 is itself a horizontal line, such as the brown dashed line in the figure. Therefore, if U-shaped responses all encode a linear combination of ingredients, neuronal data should lie on *some* horizontal line, contrary to our observations. Hence, this variant of the ingredient-based hypothesis can be also rejected.

A third variant is the hypothesis that different U-shaped responses encode different ingredients. For example, the linear relationship between k* and n* could be perhaps explained if U-shaped responses with large k* encode, say, sugar (red line), while U-shaped responses with small k* encode another ingredient, say ingredient X (green line). However, this hypothesis can only be true assuming that we happened to record from sugar-encoding neurons on days in which n* was large, and that we happened to record from X-encoding neurons on days in which n* was small—a seemingly impossible coincidence. Thus this variant can also be rejected.

The fourth and last variant of the ingredient-based hypothesis is that different U-shaped responses encode different linear combinations of ingredients. This variant combines the second and third variants, and can be rejected for the same reasons.

Studies of gustatory responses in various areas of the orbitofrontal cortex, particularly those of Rolls and colleagues, found neurons encoding the taste of one particular juice, whose responses decreased following selective satiation of that juice[12,14]. We emphasize that the relationship between k* and n* observed in our data cannot be accounted for by the ingredient-based hypothesis taken in combination with the phenomenon described by Rolls. To appreciate this point, consider again the response shown in figure s12. According to the ingredient-based hypothesis, the response is U-shaped because the activity of this neuron encodes the quantity of some ingredient X present in both juices A and B. (Here we refer to the first variant of the hypothesis, but the argument is valid more generally.) In the context of the ingredient-based hypothesis, the phenomenon described by Rolls can be described in terms of reduced responsiveness, or neuronal de-sensitization. For example, under de-sensitized conditions, the response of the neuron in figure s12 to any given quantity of ingredient X may be reduced by half. Imagine now to record from this same neuron under conditions of reduced responsiveness. Since the activity of the neuron encodes the quantity of X consumed by the monkey, and since X is present in both juices, reduced responsiveness will affect both trials in which the monkey chooses juice A *and* trials in which the monkey chooses juice B. For example, if the neuronal responsiveness is halved, we will observe half-sized responses when the monkey chooses A, and half-sized responses when the monkey chooses B. As a consequence, in figure s12, both slopes a_A and a_B will be halved. But if both slopes are halved, their *ratio* does not change. More generally, if both slopes are scaled by the same factor, their ratio does not change. As quantified in Eq.2, this means that any change in responsiveness to the encoded ingredient affects the scaling parameter m, which multiplies both (#A) and (#B), but does not affect the ratio parameter k*. In summary, even assuming changes in responsiveness of the kind described by Rolls, the ingredient-based hypothesis predicts that k* should be constant and independent of n*, contrary to our observations. Thus the ingredient-based hypothesis cannot be "salvaged" by appealing to changes in neuronal responsiveness a la Rolls.

To conclude, our analysis demonstrates that U-shaped responses do not encode the quantity of any particular ingredient or combination of ingredients, but rather the value the monkey assigns to the juice it chooses to consume.

Conclusions

We showed that OFC responses do not depend on the visuomotor contingencies of the task. Assuming that OFC responses encode each only one variable and that value functions are linear, we showed that OFC responses are best described as encoding *value A|B offered*, *chosen value*, and *A|B chosen*. In the main text, we refer to these variables respectively as *offer*

value, *chosen value*, and *taste*. The explanatory power of these three variables is high (mean $R^2 = 0.63$) and significantly higher than that of challenging alternatives. Conversely, we showed that in the large majority of cases, it is sufficient to assume that OFC responses encode each only one variable, if that variable is one of the three characteristic of this area. Indeed, taking into consideration the possibility that OFC responses encode a mixture of variables generally does not provide a significantly better account. We also showed that for the large majority of OFC responses recorded in our experiment it is adequate to assume a linear value function. In a separate analysis of U-shaped neuronal responses we showed a statistical identity between the slope ratio and the behaviorally measured relative value. For any given pair of juices, the two quantities co-vary on a session-by-session basis, which rules out any ingredient-based hypothesis. Finally, we showed that the timing by which OFC neurons encode *offer value*, *chosen value* and *taste* corresponds well to the mental operations that monkeys presumably undertake during economic choice. Specifically, neurons encoding the *offer value*—an operation necessary to make a choice—are the most prevalent shortly after the *offer* is presented to the monkey. Neurons encoding the *chosen value* are frequently observed during the delay, when the monkey has presumably internally made a choice, but before the choice is actually revealed. Finally, neurons encoding the *taste* of the chosen juice are most prevalent immediately before and after juice delivery. Taken together, these results suggest that neurons in OFC provide a substrate for value assignment during economic choice.

Supplementary References

1. Judge, S. J., Richmond, B. J. & Chu, F. C. Implantation of magnetic search coils for measurement of eye position: an improved method. *Vision Res* **20**, 535-8 (1980).

2. Padoa-Schioppa, C., Jandolo, L. & Visalberghi, E. Multi-stage mental process for economic choice in capuchins. *Cognition* **99**, B1-B13 (2006).

3. Chiavaras, M. M. & Petrides, M. Orbitofrontal sulci of the human and macaque monkey brain. *J Comp Neurol* **422**, 35-54 (2000).

4. Carmichael, S. T. & Price, J. L. Architectonic subdivision of the orbital and medial prefrontal cortex in the macaque monkey. *J Comp Neurol* **346**, 366-402 (1994).

5. Carmichael, S. T. & Price, J. L. Connectional networks within the orbital and medial prefrontal cortex of macaque monkeys. *J Comp Neurol* **371**, 179-207 (1996).

6. Dunn, O. J. & Clark, V. *Applied statistics: analysis of variance and regression* (Wiley, New York, 1987).

7. Glantz, S. A. & Slinker, B. K. *Primer of applied regression & analysis of variance* (McGraw-Hill, Medical Pub. Division, New York, 2001).

8. Neter, J., Wasserman, W. & Kutner, M. H. *Applied linear statistical models: regression, analysis of variance, and experimental designs* (Irwin, Homewood, IL, 1990).

9. Rieke, F., Warland, D., de Ruyter van Steveninck, R. & Bialek, W. *Spikes: exploring the neural code* (MIT Press, Cambridge, MA, 1997).

10. Scott, T. R. & Plata-Salaman, C. R. Taste in the monkey cortex. *Physiol Behav* **67**, 489-511 (1999).

11. Brand, J. G. in *Tasting and smelling* (eds. Beauchamp, G. K. & Bartoshuk, L.) 1-24 (Academic Press, San Diego, CA, 1997).

12. Pritchard, T. C. et al. Gustatory neural responses in the medial orbitofrontal cortex of the old world monkey. *J Neurosci* **25**, 6047-56 (2005).

13. Zigmond, M. J., Bloom, F. E., Landis, S. C., Roberts, J. L. & Squire, L. R. *Fundamental neuroscience* (Academic Press, San Diego, CA, 1999).

14. Rolls, E. T., Sienkiewicz, Z. J. & Yaxley, S. Hunger modulates the responses to gustatory stimuli of single neurons in the caudolateral orbitofrontal cortex of the macaque monkey. *Eur J Neurosci* **1**, 53-60 (1989).

Figure s1

Name Used in Main Text	Collapsed	Variable Name	Definition
		total value	*chosen value + other value*
chosen value		*chosen value*	Value of the chosen juice
		other value	Value of the non-chosen juice
		(chosen–other) value	*chosen value – other value*
		(other/chosen) value	*(other value) / (chosen value)*
		total number	*max number + min number*
		max number	Maximal offered number
		chosen number	Chosen number
		min number	Minimal offered number
		other number	Non-chosen number
		(max–min) number	*max number – min number*
		(chosen–other) number	*chosen number – other number*
		(min/max) number	*(min number) / (max number)*
		(other/chosen) number	*(other number) / (chosen number)*
offer value	*value A|B offered*	*value A offered*	Value of juice A offered
		value B offered	Value of juice B offered
taste		*A|B chosen*	Binary: 1 if A chosen, 0 if B chosen
	value A|B chosen	*value A chosen*	Value of juice A chosen
		value B chosen	Value of juice B chosen

Figure s1. Tested variables. We tested variables related to value, variables related to number, and variables related to one of the two juices. Note that having assumed a linear value function (see Supplementary Methods), the value of one of the two juices offered or chosen is proportional to the number of drops of juice. The two collapsed variables are indicated in the bottom left of the table. In the main text, we re-labeled the variable *value A|B offered* = *offer value*, and the variable *A|B chosen* = *taste*.

Figure s2

	position juice A	move direction	offer type
pre-*offer*	0	0	1
post-*offer*	18	14	284
late delay	4	6	207
pre-*go*	4	8	133
RT	5	5	93
pre-*juice*	3	14	329
post-*juice*	5	11	332
total	39	58	1379
at least 1	33	46	505

Figure s2. 3-way ANOVA. We recorded the activity of 931 cells. We analyze each cell in each time window with a 3-way ANOVA (factors [position of juice A] x [movement direction] x [offer type]). The three columns in the table indicate the number of cells for which each of the main factors has a significant effect ($p<0.001$). The bottom row indicates the number of cells whose activity pass the ANOVA test in at least one time window. The factor [offer type] is significant for a total of 1379 responses.

Figure s3

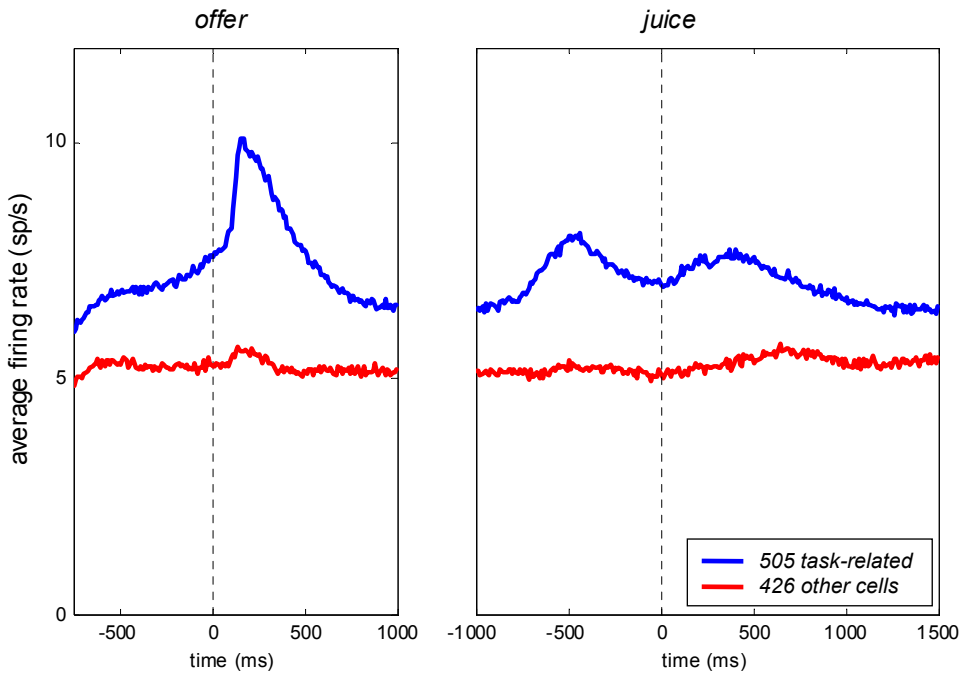


Figure s3. Average activity profile. Neuronal activity was analyzed in 10ms-large, non-overlapping time bins. For each cell, we included in this analysis all trials and all trial types. We then averaged the resulting activity profile across cells. This analysis was done separately for task-related cells (i.e., cells that passed the ANOVA criterion in at least one time window; blue color) and for other cells (i.e., cells that did not pass the ANOVA criterion in any time window; red color).

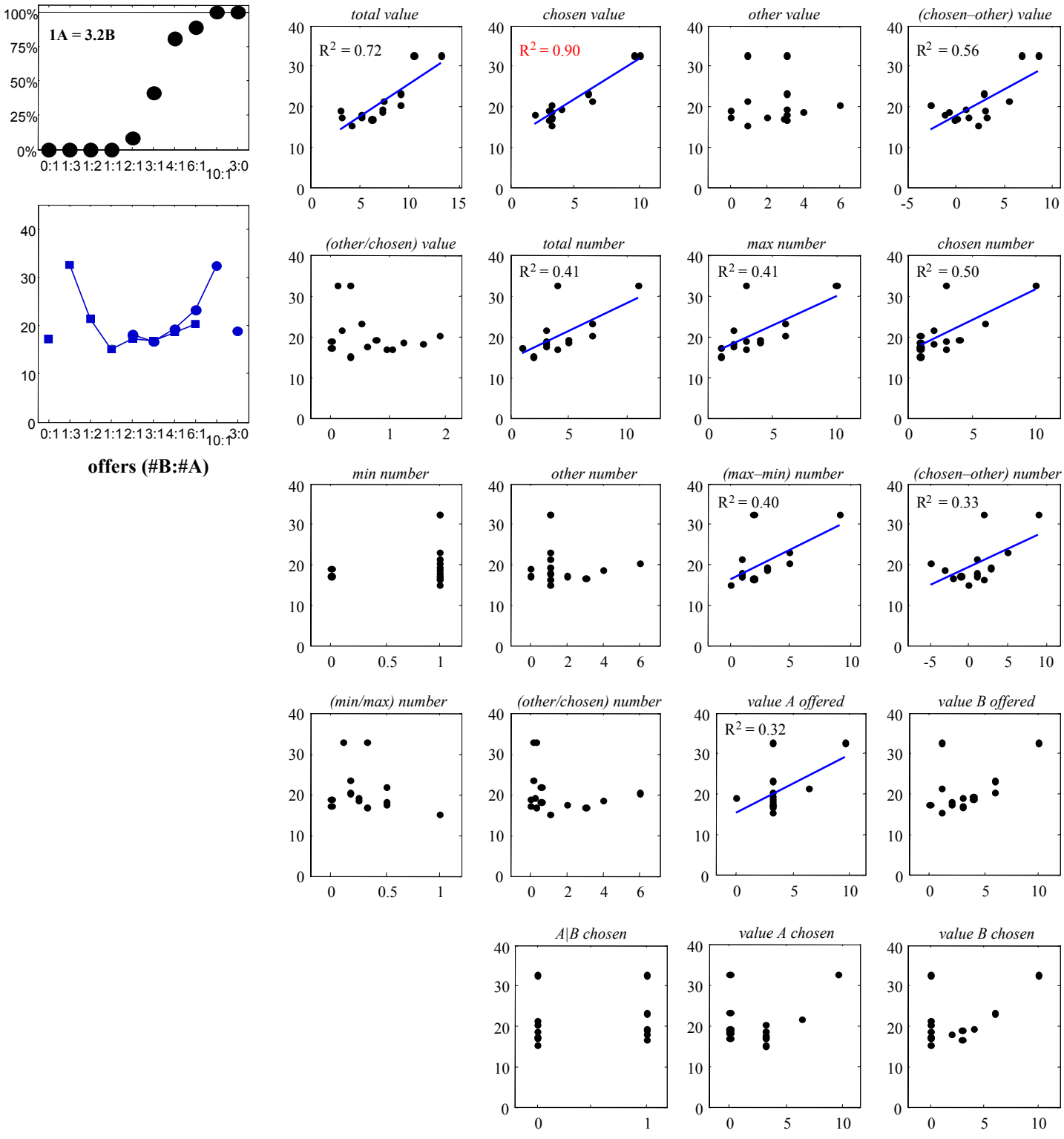


Figure s4. Linear regressions. The two top left panels show the behavioral choice pattern and one response plotted in the same coordinates. The response is the same shown in figure 3a (main text), except that here we grouped trials by trial type (squares for "A" choices; circles for "B" choices). In the other panels, the neuronal response (y-axes) is plotted against each of the 19 variables (x-axes), and each dot represents one trial type. Values are expressed in units of V(B). Blue regression lines indicate regression slopes significantly different from zero. For variables that explain the response, the R^2 is indicated on the top left corner in the panel. The highest R^2 (*chosen value*) is marked in red.

Figure s5

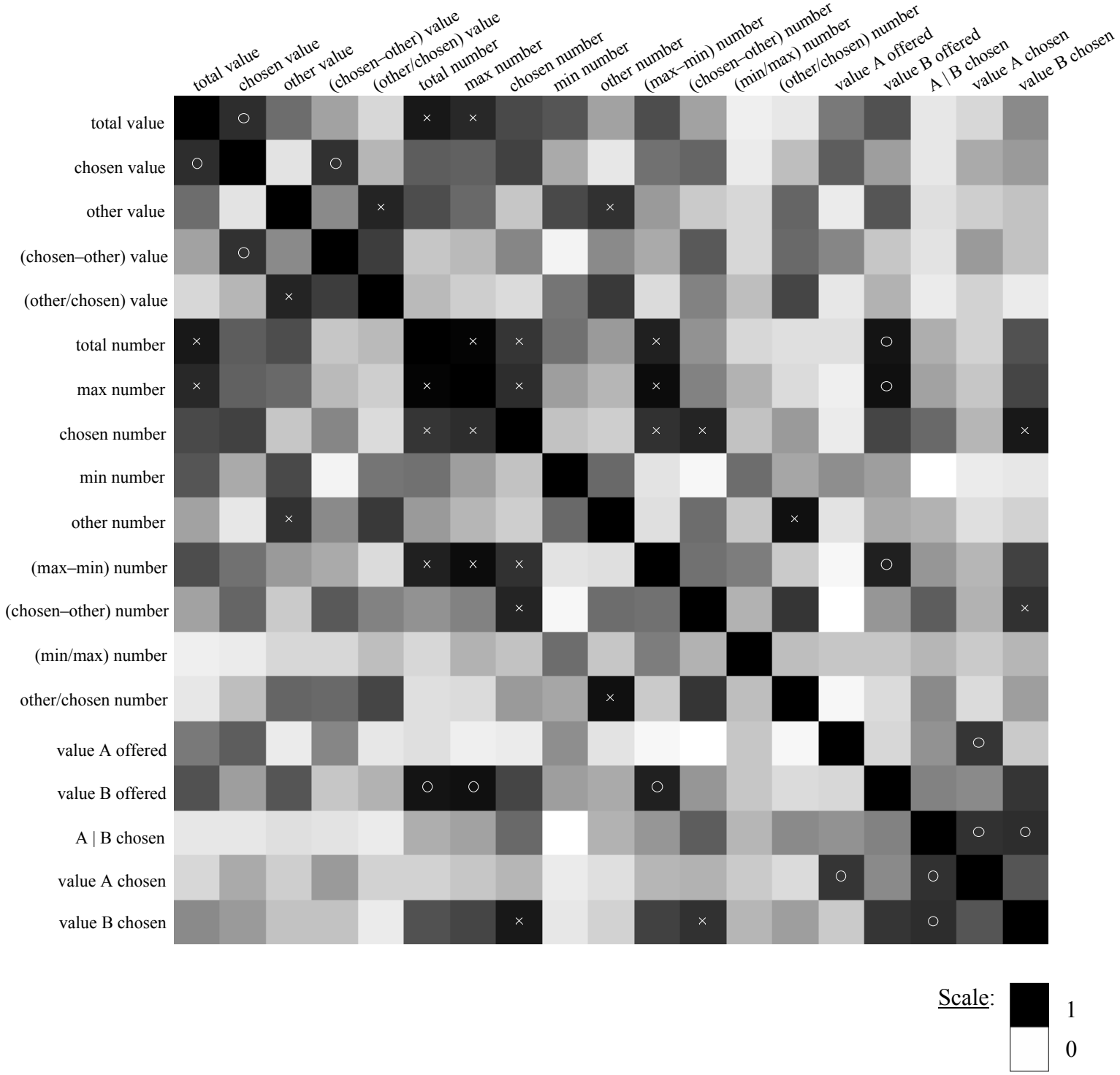


Figure s5. Correlation matrix. Elements of the correlation matrix vary between 0 and 1 (see Supplementary Methods). Here the correlation matrix ρ is rendered in gray scale so that ρ=1 is represented in black (diagonal elements) and ρ=0 is represented in white. Small circles and crosses indicate matrix elements for which ρ>0.8 (excluding the diagonal). We use circles for correlations that include one of the selected variables and crosses otherwise.

Figure s6

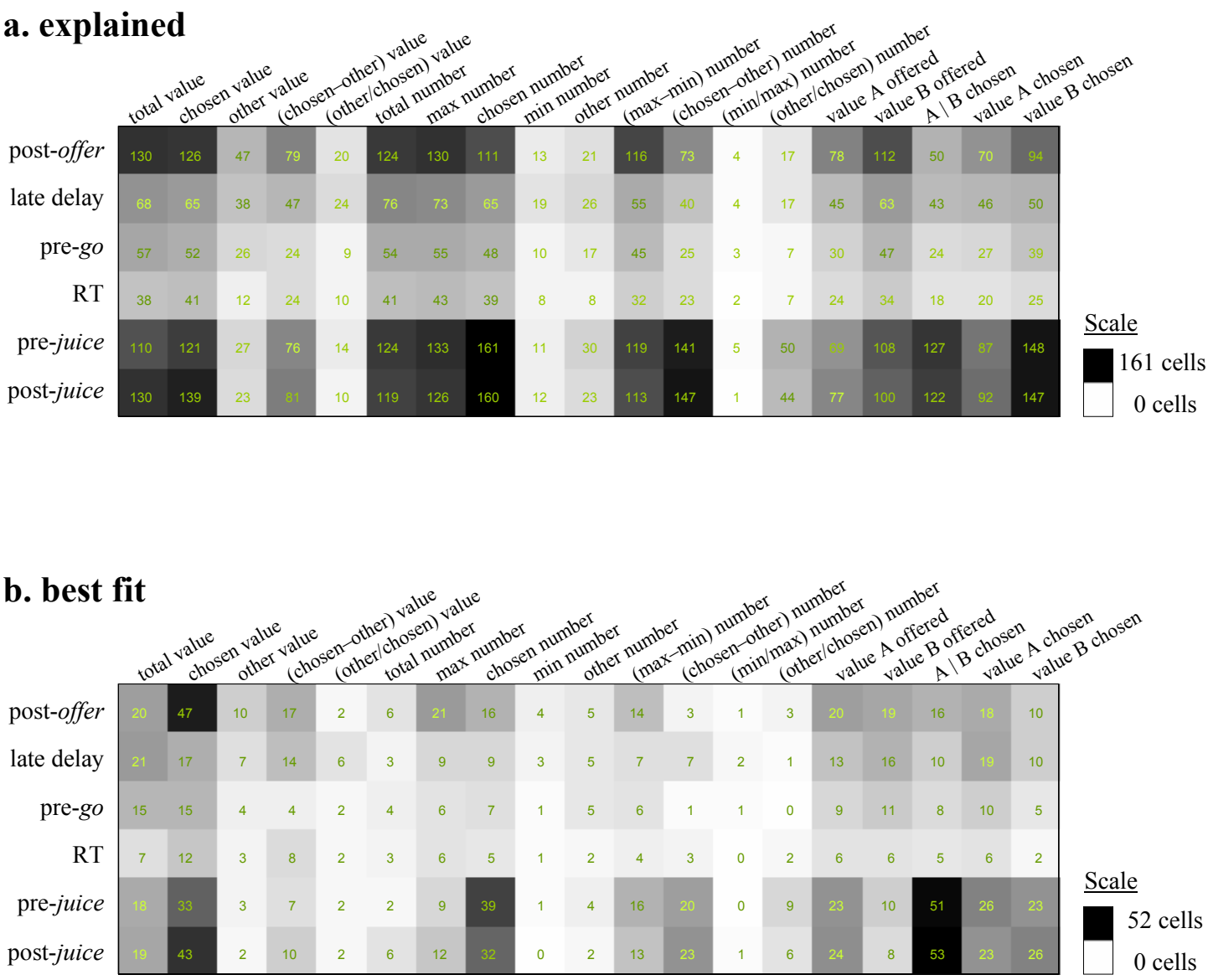
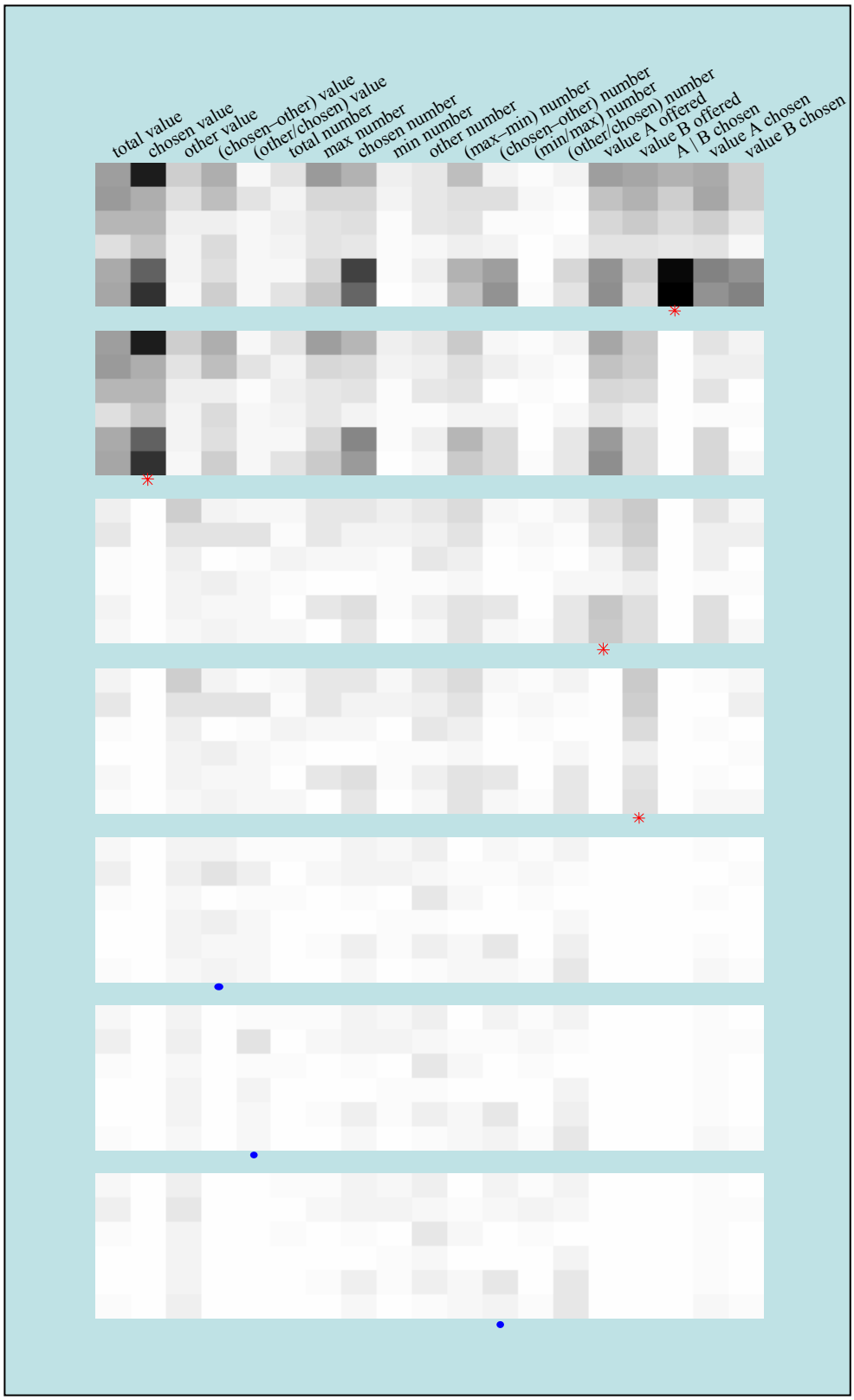


Figure s6. Qualitative analysis. **Top.** Numbers in the top panel represent for each variable (x-axis) and for each time window (y-axis) the number of responses for which the linear regression provided a non-zero slope (i.e., the number of responses explained by the variable). The color table underneath reports the same numbers in gray scale. **Bottom**. Numbers in the bottom panel represent for each variable and for each time window the number of responses for which the corresponding variable provided the best fit (highest R^2). Again, the color table represents the same numbers in gray scale.

Figure s7



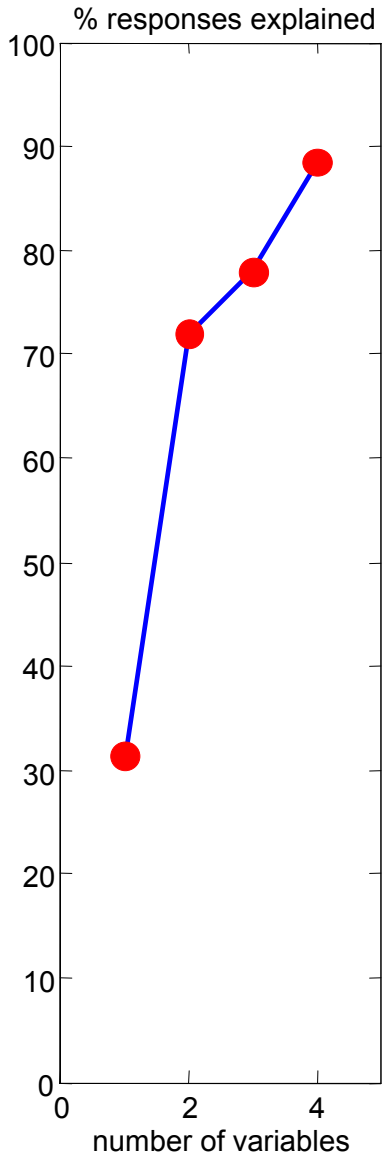


Figure s7. Stepwise selection method. See Supplementary Results.

Figure s8

position juice A	move direction	offer type		*offer value*	*chosen value*	*taste*	other class	unclas-sified
0	0	1	pre-*offer*	0	0	0	1	0
18	14	284	post-*offer*	113	87	25	27	32
4	6	207	late delay	72	46	26	33	30
4	8	133	pre-*go*	46	37	16	15	19
5	5	93	RT	29	30	11	13	10
3	14	329	pre-*juice*	95	79	93	29	33
5	11	332	post-*juice*	90	93	97	24	28
39	58	1379	total	445	372	268	142	152
33	46	505	at least 1					

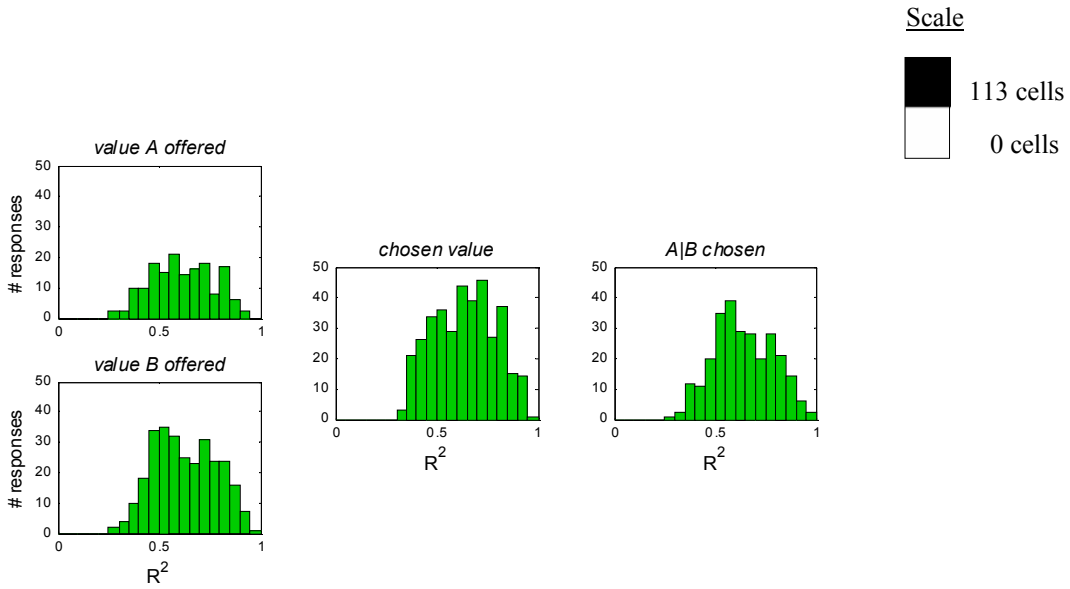
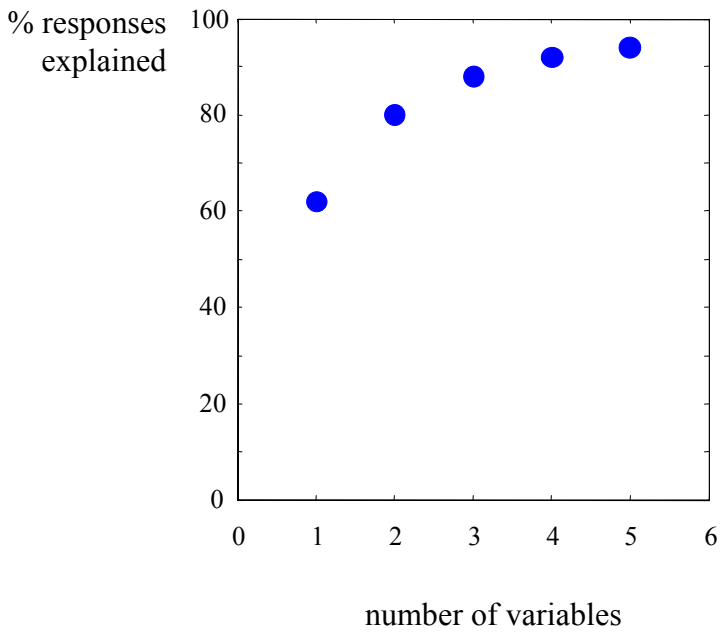


Figure s8. Population summary. **Top.** The figure illustrates the final result of the classification. The three columns on the left repeat teh results of the s-way ANOVA (figure s2). The factor [offer type] is significant for a total of 1379 responses. The five columns on the right summarize the results of the variable selection analysis. Responses that cannot be explained by any of the 19 variables appear in the rightmost column (unclassified). Variables *value A|B offered*, *chosen value* and *A|B chosen* collectively explain 1085 responses (79% of the total, 88% of responses explained by the 19 variables). Responses explained by one of the 19 variables but not explained by any of the selected variables appear on the second column on the right (other class). The remaining three columns indicate the number of responses classified as *value A|B offered*, *chosen value,* and *A|B chosen*, respectively. The color table represents the same number in gray scale. The prevalence of different response classes varies over time windows: *value A|B offered* is most prevalent shortly after the *offer*, *chosen value* is prevalent throughout the trial, and *A|B chosen* is most prevalent late in the trial, before and after *juice* delivery. This time course can be also observed at much higher resolution (figure 4, main text). **Bottom.** How well are responses accounted for in the final classification? The histograms show for the four response classes the number of responses (y-axis) with the corresponding R^2 (x-axis). The responses included in the four histograms are 159, 286, 372 and 268, respectively. The mean (avg) and median (med) of the four distributions are (avg=0.61, med=0.60), (avg=0.63, med=0.61), (avg=0.64, med=0.64) and (avg=0.64, med=0.63), respectively.

Figure s9



selected variables	$d = 1$	$d = 2$	$d = 3$	$d = 4$	$d = 5$
	value A|B offered	*total value* *value A|B chosen*	*chosen value* *value A|B offered* *A|B chosen*	*chosen value* *value A|B offered* *A|B chosen* *chosen number*	*chosen value* *value A|B offered* *A|B chosen* *other value* *(chosen–other) number*

Figure s9. Best-subset method. See Supplementary Results.

Figure s10

variable X	variable Y	nX	nY	p
chosen value	*total value*	58	21	$<10^{-5}$
chosen value	*(chosen–other) value*	91	34	$<10^{-7}$
chosen value	*chosen number*	53	20	$<10^{-5}$
value A|B offered	*total number*	99	14	$<10^{-10}$
value A|B offered	*max number*	93	19	$<10^{-10}$
value A|B offered	*(max–min) number*	102	15	$<10^{-10}$
value A|B offered	*value A|B chosen*	100	19	$<10^{-10}$
A|B chosen	*value A|B chosen*	26	19	0.12

Figure s10. Post hoc analysis. The best subset of three variables includes *chosen value*, *value A|B offered*, and *A|B chosen*. We tested these three variables separately against all other variables highly correlated with them (ρ>0.8). The relevant pairs of variables are marked by small white circles in the correlation matrix in figure s5. The collapsed variable *value A|B offered* is tested here against variables highly correlated either with *value A offered* or with *value B offered*. In addition to the comparisons dictated by the ρ>0.8 criterion, we tested the variable *chosen value* against the alternative variable *chosen number*. In the table, the two left columns indicate the tested variable and the alternative variable, the next two columns indicate the marginal explanatory power of the two variables, and the right column indicates the result of a binomial test. In essence, all tests indicate that selected variables have significantly higher explanatory power than the challenging alternatives, except that the variable *A|B chosen* does not reach significance level against the alternative *value A|B chosen* (bottom row).

a. # responses

	value A offered	*value B offered*	*chosen value*	*A|B chosen*	*value A chosen*	*value B chosen*	*chosen number*	*value A offered*^{2}	*value B offered*^{2}	*chosen value*^{2}	*none*
value A offered	-	2	1	2	7	1	0	19	-	-	127
value B offered	7	-	12	5	4	12	24	-	19	-	203
chosen value	12	15	-	13	4	3	4	-	-	16	305
A|B chosen	12	2	22	-	10	8	12	-	-	-	202

b. % responses

	value A offered	*value B offered*	*chosen value*	*A|B chosen*	*value A chosen*	*value B chosen*	*chosen number*	*value A offered*^{2}	*value B offered*^{2}	*chosen value*^{2}	*none*
value A offered	-	1	1	1	4	1	0	12	-	-	81
value B offered	2	-	4	2	1	4	8	-	7	-	72
chosen value	3	4	-	4	1	1	1	-	-	4	83
A|B chosen	5	1	8	-	4	3	5	-	-	-	77

Figure s11. Second order encoding. See Supplementary Results.

Figure s12

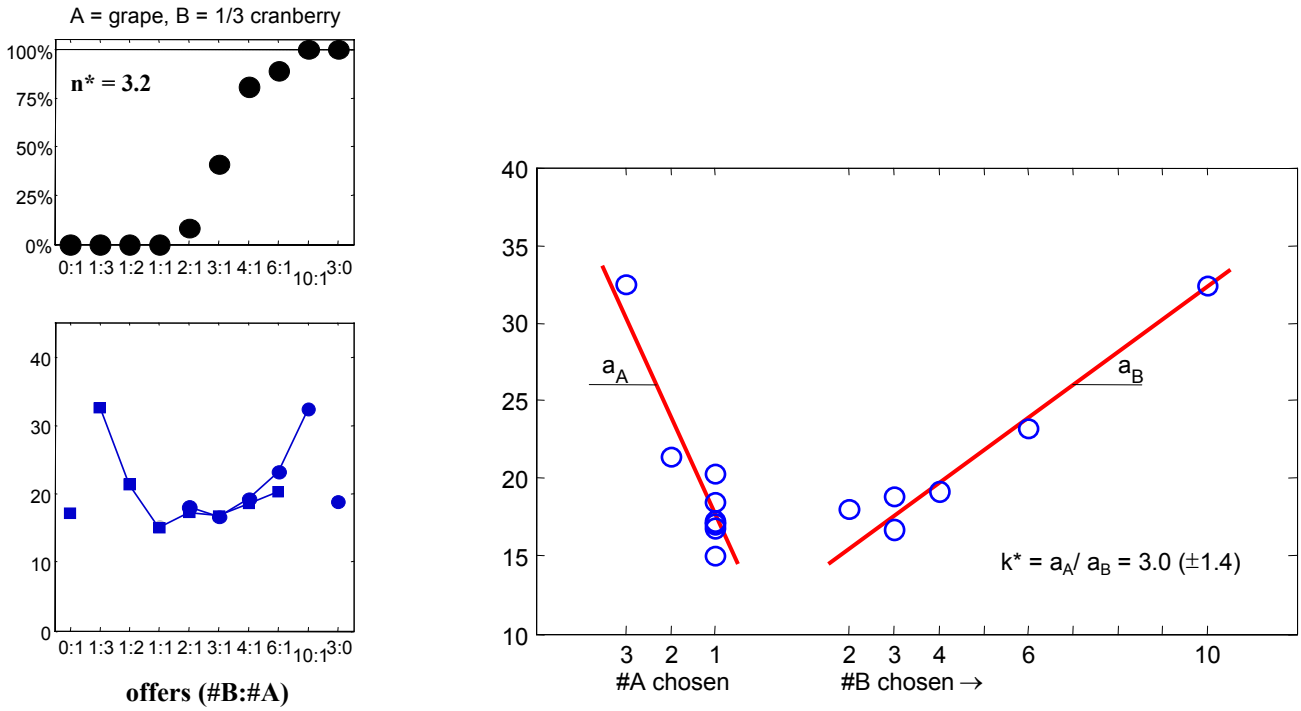


Figure s12. One example of U-shaped response (same shown in figures 3a and s4). In the left panels, the behavioral choice pattern and the neuronal response are plotted in the usual "ordinal" x-axis coordinates. From the behavioral choice pattern, we obtain the relative value n*=3.2. In the right panel, the neuronal response is plotted in "cardinal" x-axis coordinates, separately for trials in which the monkey chose juice A and juice B. The two regression slopes are plotted in red. From the bi-linear regression, we obtain a slope ratio k* = 3.0 (±1.4) (±95% confidence interval).

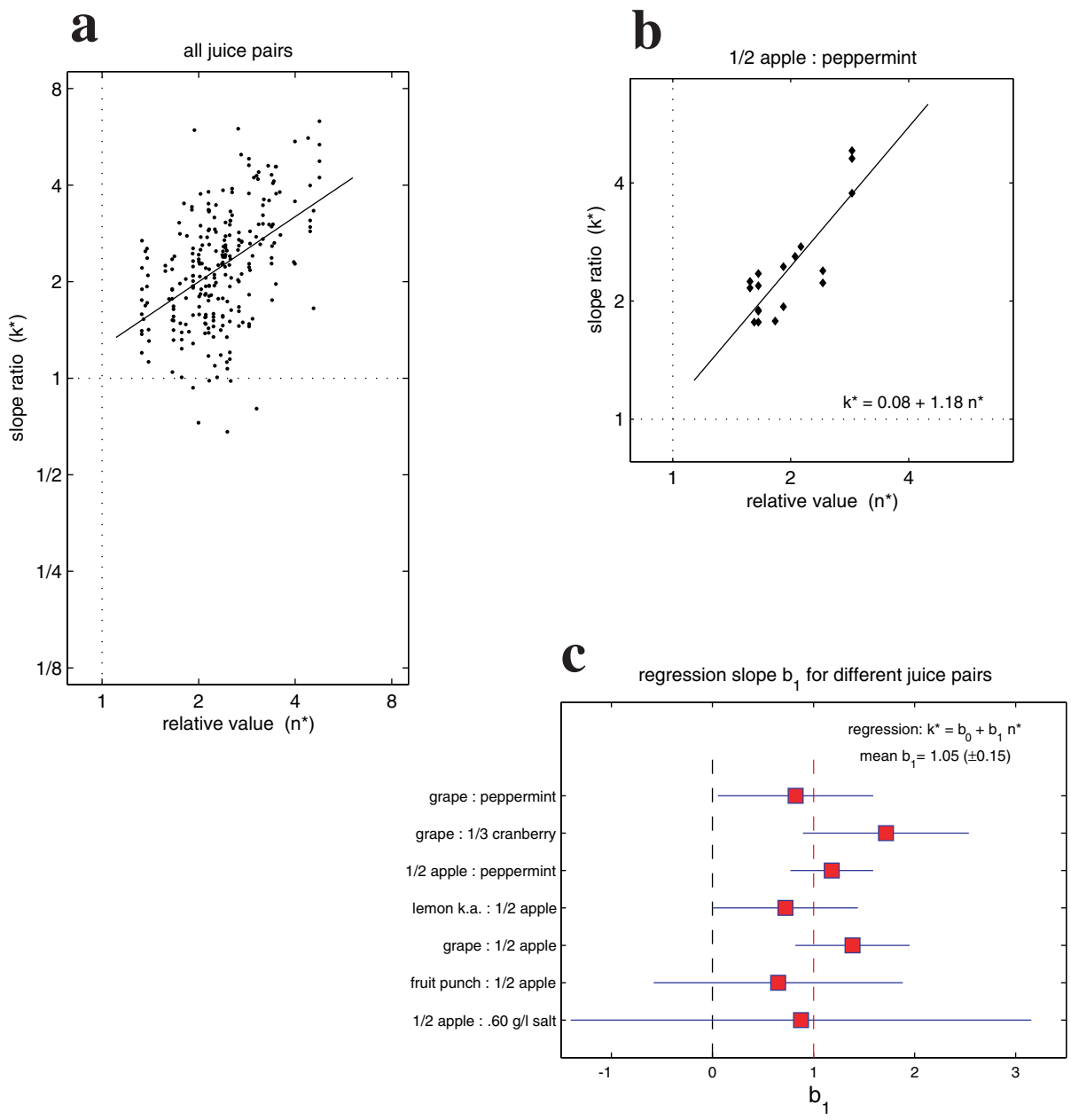


Figure s13. Relationship between slope ratio k* and relative value n*. See Supplementary Results.

Figure s14

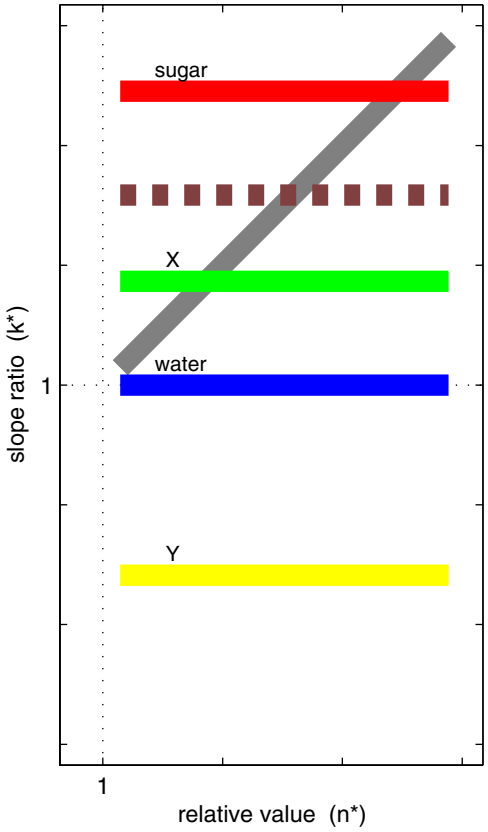


Figure s14. Predicted relationship between k* and n*. If U-shaped responses encode *chosen value*, the values of k* should lie on a diagonal (gray line). If they encode the quantity of some juice ingredient (e.g., water, sugar, or some other ingredient X or Y), values of k* should lie on some horizontal line corresponding to the concentration ratio (blue, red, green, and yellow lines). Similarly, if U-shaped responses encode the linear combination of multiple ingredient, values of k* should lie on some horizontal line (e.g., brown dotted line).