Neural Activity Reveals Preferences Without Choices

by Alec Smith¹, B. Douglas Bernheim², Colin Camerer¹,³, and Antonio Rangel¹,³

¹ HSS, Caltech
² Department of Economics, Stanford, and NBER
³ CNS, Caltech

April 2011
PRELIMINARY AND INCOMPLETE
DO NOT QUOTE OR DISTRIBUTE WITHOUT PERMISSION

Abstract.

We investigate the feasibility of inferring the choices people would make (if given the opportunity) based on their neural responses to the pertinent prospects when they are not engaged in actual decision making. We develop procedures that involve prediction models relating choices to these “non-choice” neural responses. We estimate these models with data for one set of individuals and options, and use the estimated models to predict choices for new items (but the same set of individuals), for new individuals (but the same set of items), and for new items and individuals simultaneously. We find that individual-specific and group-specific models predict well out of sample (to choices involving new items). Our prediction models for groups are somewhat portable simultaneously across groups and items, but our prediction models for individuals do not work well for other individuals (even for the same items). These methods are of potential value when choice data are unavailable or otherwise limited in ways that render the estimation of choice mappings problematic. Equipped with predictive relationships between choices and non-choice responses (including neural responses) that are stable across a reasonably broad domain of alternatives, one could in principle synthesize choice data for domains within which actual choice data are absent, deficient, or contaminated by selection effects and/or endogeneity.
I. Introduction

Empirical analyses of economic decision making generally involve the estimation of choice mappings (such as demand curves). By far the dominant tradition in economics is to build up choice mappings from observations of actual choices within the same domain—that is, to predict choices from closely related choices, adjusting as necessary for differences in the objective characteristics of the available alternatives. Unfortunately, this traditional approach proves problematic in a variety of circumstances due to the various practical limitations of choice data. First, in some settings, there is simply no opportunity to observe actual choices within the same domain.\(^1\) Second, observed choices that pertain in some fashion to the domain of interest sometimes bear only a distant or indirect relation to the choices of interest. To forecast the latter choices from the former, economists must then make restrictive assumptions.\(^2\) Third, even when pertinent choice data are available, economists must frequently make strong assumptions to overcome the data’s practical limitations and/or accommodate the complexity of the setting.\(^3\) Fourth, naturally occurring choice data rarely reflect clean experiments, and generally implicate considerations that potentially confound reliable extrapolation. Concerns about uncontrolled factors and the endogeneity of opportunity sets are endemic.\(^4\) To some extent, one can address these limitations by supplementing naturally occurring choices with experimental choices. Experimental evidence has proven highly valuable, but is necessarily limited by considerations of practicality (e.g., cost and scale), generalizability, and representativeness.

---

\(^1\) For example, although environmental economists are interested in measuring the typical individual’s willingness to pay for a given reduction in the probability of environmental damage caused by an oil spill, no naturally occurring choice reveals that preference. These considerations motivate the use of “stated preference techniques;” see, for example, Bateman et al. (2002).

\(^2\) As an example, a variety of studies have invoked strong assumptions concerning the process of wage determination to infer the average willingness to accept an incremental likelihood of death from the variation in wages across occupations with different risk exposures; see, for example, Viscusi (1978).

\(^3\) For example, industrial organization economists frequently estimate demand systems for markets with highly heterogeneous goods (such as automobiles), but often only have snapshots of the distribution of sales across products at a given point in time. In such settings, heavy reliance on structural models is the norm; see, for example, Berry, Levinsohn, and Pakes (1995).

\(^4\) Thus, for example, the absence of exogenous variation in eligibility for tax-favored retirement accounts ultimately precludes available choice data from shedding much light on the question of whether (and to what extent) such accounts increase saving; see Bernheim (2002).
An alternative approach – one that this paper adopts – is to posit the existence of stable relationship between people’s responses to potential prospects when they are not engaged in actual decision making (henceforth, “non-choice responses” for brevity), and the decisions they make when they do confront actual choices. We hypothesize that the relation between choices and non-choice responses is more likely to be stable across a broad range of choice domains (such as food and clothing) than the relation between choices and the objective characteristics of the pertinent options (such as size or color). If the former relation proves sufficiently stable, then it may be possible to forecast choices reliably by assessing and extrapolating from non-choice responses, without observing actual choices within the same domain. Such forecasts would “reveal preferences” in the classic sense of identifying what an individual would choose, but without relying directly on actual choices.5

Equipped with predictive relationships between choices and non-choice responses that are stable across a reasonably broad domain of alternatives, one could in effect synthesize choice data for domains within which choice data are absent or deficient.

Some non-choice responses to potential prospects involve subjective reports. For example, the answer to a hypothetical choice question is a subjective report, and indeed such answers are often used to project real choices (see the discussion in Section II). In this paper, we explore the predictive power of non-choice physiological responses – in particular, neural responses – to potential prospects. We hypothesize that some of the neural valuation processes governing decisions operate even when an individual encounters prospects that are not immediately available (see also Tusche et al., 2010, Lebreton et al., 2009, and Kang et al., 2011). Conceivably, automatic neural valuation could be a beneficial evolutionary adaptation: an individual who constantly assesses the values of potential prospects and proceeds to a choice when one or more apparently high-value prospect is potentially available is more likely to achieve a beneficial outcome. Ultimately, we envision using various types of non-choice responses, including subjective reports and physiological responses, in combination to develop more accurate predictive models. This paper, however, is concerned with the threshold question: do non-choice neural responses contain information that is useful in predicting actual choices?

5 To be clear, this strategy does rely indirectly on actual choices, in that the stability of a prediction model can be evaluated only by comparing predictions with outcomes in some domains. However, if such a model proves stable across many and varied domains, it can be applied in other domains where pertinent choice data are unavailable, severely limited, or otherwise problematic.
The canonical task motivating our investigation is to determine how people will behave when confronted with some new or difficult-to-observe choice situations. Imagine assembling a group of individuals, measuring their non-choice neural responses to prospects that we can actually implement, as well as to new choice situations, and then presenting them with unanticipated choices among the implementable prospects. We can then estimate the relationship between their choices and non-choice responses for the implementable prospects, and use that relationship along with non-choice neural responses for the new situations to predict behavior in the new situations. The question of interest is whether the non-choice neural responses contain enough information either to make reasonably accurate predictions concerning choices in the new situations, or at least to improve predictions based on other data.

Because the purpose of this study is to provide proof of concept, we confine attention to a narrow choice domain, consisting entirely of snack foods. Procedurally, subjects view pictures of 100 snacks while undergoing an fMRI brain scan. After the scan is complete, they are asked (unexpectedly) to make choices among 50 pairs of snacks (one of which is implemented), with each snack appearing in one and only one pair. While they are informed in advance that they will eat one of the snacks, they are not told that they will have choices until the scanning phase is concluded; thus, neural activity during the scan is not in any way influenced by choice tasks. After completing the choices, they are asked to rate the extent to which they liked each item. Section III describes our experimental procedures in greater detail.

To determine whether non-choice neural responses contain information that is useful in predicting actual choices, we ask first whether it is possible, for any particular individual, to calibrate a prediction model relating choices to non-choice neural responses for one set of objects, and use that model successfully to predict choices involving objects not contained in the original set, based on the non-choice neural responses they induce (see Section IV.A). Leaving out two pairs of items at a time, we estimate a prediction model based on the other 48 pairs, and use it to predict the individual’s choice for the excluded pairs. The alternative identified as the most likely choice according to the prediction model is in fact chosen 61% of the time, on average across subjects. If decisions were unrelated to non-choice neural activity, we would observe a 50% success rate; the difference between this baseline and the observed success rate is both substantively and statistically significant.
We show that subjects are divided into two groups of roughly equal sizes. For one group, success rates are even higher (68%), and the realized frequencies of predicted events mirror the predicted frequencies across groups of items. For the other group, success rates are considerably lower (54%), and the realized frequencies of predicted events bear essentially no relation to the predicted frequencies. We conclude that our within-subject procedure for predicting choices involving new items performs with considerable success for roughly half of our subjects, and not at all for the other half.

Predictive accuracy on the level of a single individual is a demanding objective, one that goes beyond the requirements of most economic analyses, which are more typically concerned with aspects of group behavior – averages, aggregates, and possibly distributions. Group averages may be easier to predict than individual choices for a number of reasons, including the fact that various types of noise average out over multiple individuals. Accordingly, we next ask whether it is possible, for any particular group of individuals, to calibrate a prediction model relating a measure of average subjective value to average non-choice neural responses for one set of items, and use that model successfully to predict the average subjective values of items not contained in the original set, based on the non-choice neural responses they induce (see Section IV.B). We achieve even greater success predicting group averages: we correctly predict whether an item's average value lies above or below the median for 75% of items, and once again predicted probabilities match up reasonably well with realized frequencies.

If non-choice neural activity exhibits a sufficiently similar relation to choice across subjects, then it should be possible to construct a single prediction model and use it without recalibration to predict choices based on neural measurements taken from new individuals or groups. Such a model would have considerable practical value in that, once estimated, it would vastly simplify the steps required to formulate additional predictions. To predict behavior in new situations, one could simply collect data on non-choice neural responses to the relevant prospects for a new group of individuals, and apply an existing predictive model; it would not be necessary in addition to gather the requisite data to estimate new predictive models for those subjects. Accordingly, we also investigate whether predictive models are portable across individuals and/or groups. Here, our results are considerably weaker. When we use a model estimated for one group of individuals to predict the choices of another individual, our analysis exhibits only a hint of predictive power. However, we do
achieve a moderate degree of success (roughly 60% accuracy) when predicting a group’s average valuation for new objects from a relationship estimated with data pertaining to other objects, gathered from another group.

Taken together, our results demonstrate that non-choice neural reactions to images of potentially desirable objects contain a great deal of information that can be used to predict decisions made by a particular individual, or average decisions made by a group of individuals, in new choice situations. Improvements in methods and advancements in technology are likely to enhance the success of this approach. Thus, our study provides unambiguous proof-of-concept for the feasibility of the larger agenda.

II. Related literature

Our work is related to several existing lines of research. Within economics, there is a large and important literature concerning techniques for predicting choices from answers to hypothetical questions. It is well-established that hypothetical choices and/or valuations are highly correlated with real choices and/or valuations, but sizable biases have been documented. For example, subjects exaggerate their willingness-to-pay for new products, and express preferences that are more virtuous than their actual choices. See, for example, Cummings et al. (1995), Johannesson et al. (1998), List and Gallet (2001), Little and Berrens (2004), Murphy et al. (2005), Blumenschein et al. (2007). Much of that literature examines the effectiveness of protocols for reducing bias. Bernheim et al. (2011) pursue an alternative strategy – one that is more closely related methodologically to the current paper – by treating the problem as one of optimal statistical prediction, where (potentially biased) hypothetical choices are used in combination with other subjective non-choice responses as potential predictors.

There is also a substantial literature in neuroscience concerning the neural correlates of choice. However, with very few exceptions (discussed below), that literature is concerned with neural activity that codes reliably for value signals during the act of choice; see, for example, Knutson et al. (2007), Plassmann et al. (2007, 2010), and Hare et al. (2008). Consequently, the issues those studies address differ fundamentally from the ones that motivate our inquiry. Certainly, as Knutson et al. (2007) emphasize, it is possible to predict choices from neural activity measured during the act of decision making. However,
for economists, there is little value in predicting choices in a setting where choices are themselves observable. If one's objective is to extrapolate choices based on neural activity in settings where choices are not observed, correlations between choice and choice-related neural reactions are simply unhelpful.

Two recent studies suggest, however, that the brain's valuation circuitry may be active even when people are not actively engaged in choice. Kang et al. (2011) have shown that the same area of the vmPFC and vStr correlates with the value of the stimuli during both real and hypothetical choices, which suggests that neural responses to real and hypothetical choices may share many common features. Lebreton et al. (2009) show that activity in vmPFC while subjects were asked to judge the age of paintings correlates with their liking ratings for the same paintings (elicited in a separate task). Thus, there is reason to hope that one can also reliably predict choices based on non-choice neural responses.

The current study is most closely related to recent papers in the neuroscience literature by Tusche et al. (2010), Levy et al. (2011), Luu and Chau (2009), all of which have elements of predicting choice (or tasks related to choice) from non-choice neural responses. Because the primary objectives of these papers pertain to neuroscience rather than economics, they are not designed to answer the specific questions that motivate our study. To understand the key differences between their work and ours, it is helpful to summarize several features of our analysis that are critical for the economic applications we envision. First, we are concerned with predicting real choices from neural responses during non-choice activity. Second, our interest is in out-of-sample prediction, rather than within-sample fit. We are concerned with predicting choices over one set of alternatives using a relationship estimated with data for a disjoint set of alternatives. There is both an economic reason and a technical reason for this requirement. The economic reason is that we are ultimately concerned about predicting choices over alternatives that are difficult or impossible to implement in practice. The technical reason, which we explain in detail at the end of Section IV.A.1, is that statistical procedures might otherwise predict choices correctly by exploiting neural indicators of the alternatives' identities, rather than of their perceived values. Third, our objective is not merely to predict the more likely choice, but in addition to derive reliable probability statements concerning the alternatives. We seek a procedure that will indicate whether a particular alternative will be chosen with, say, 60% probability rather than 90% probability, where outcomes bear out those probabilistic statements.
Finally, as discussed in Section I, we are concerned with several distinct types of prediction exercises: within subject, within group, across subjects, and across groups. Predicting average behavior within and across groups likely has the greatest potential value for economics (though an ability to predict behavior for subjects plainly implies an ability to predict average behavior for groups).

These four features distinguish our paper from the three studies listed above. None of them attempts to predict choices among one set of alternatives from a relationship estimated with a disjoint set of alternatives, nor do they attempt to derive and validate probability statements concerning alternatives. All three focus exclusively on within-subject classification or prediction; none attempts to predict average behavior for groups, or choices across subjects. Luu and Chau (2009) focus on within-sample classification rather than out-of-sample prediction. Both Tusche et al. (2010) and Luu and Chau (2009) study the neural correlates of hypothetical choices rather than real choices. In contrast, Levy et al. (2011) attempt to predict real choices, but their subjects were aware that they would be making some real decisions concerning the same objects during the scanning phase, and hence their procedure does not truly involve non-choice neural responses in the sense defined here.

III. Experimental Design

A. Subjects

Thirty-five right-handed subjects participated in the experiment (age range: 19 to 36 years old, 11 female). Subjects were pre-screened to ensure that they regularly ate the types of foods used in the experiment, and that they met the standard medical criteria required for the safe and reliable acquisition of fMRI data. Subjects were paid $100 for participating, and were offered a $10 bonus for limiting their head motion during the fMRI task (which, if excessive, invalidates the procedure). Despite these incentives, in-scanner head motion for eight subjects exceeded a pre-specified limit of 2mm in any direction during a scanner run. After excluding those eight subjects, 27 usable subjects remained.

Subjects were recruited at Caltech, and the experimental procedures were approved by the Caltech Institutional Review Board.
B. Procedures

Subjects were instructed to refrain from eating or drinking anything other than water for four hours prior to the experiment. At the outset of a session, they were advised that the experiment would consist of three stages. Importantly, they were told that they would receive the instructions for each stage immediately before it began, and not until they had completed the previous stage. Thus, when viewing images of snack foods in stage 1 (as described below), subjects were not aware that they would subsequently face choices among those items.

**Stage 1. Passive viewing of foods during fMRI scan.** In the first stage, subjects viewed images of 100 different snack foods while we measured their neural responses; see Figure 1 for sample images, and the Appendix for a list of all foods used in the experiment. Foods were shown in randomized order with each item appearing three times. Each image was visible for 2.75 seconds. Between images, a small white fixation cross centered on a black screen was shown for 8.25 seconds. For technical reasons related to the acquisition of the neural data, each session was divided into 6 identical blocks each consisting of 50 image presentations, separated by breaks of roughly one minute.

Presumably, the informational content (and hence predictive power) of non-choice neural responses measured in Stage 1 depends at least in part on the degree to which subjects attend to and think about each item. To enhance the psychological salience of the images, we told subjects that they would be required to eat at least three spoonfuls of one of the food items at the end of the session; with 50% probability, the item would be selected at random, and with 50% probability it would be determined in a subsequent stage of the experiment. However, subjects were not told that that they would be asked to make choices among the alternatives, or that such choices would affect which item they received.

Given the tedious nature of the task, there was a risk that subjects would fall sleep or tune out during their sessions. We therefore inserted five additional “wake-up” trials at irregular intervals within every block. During each such trial, the subject was instructed to press a button indicating whether the displayed item was sweet or salty. Subjects were given a maximum of 2.75 seconds to respond. Wake-up trials were also followed by a

---

7 A copy of the instructions is included in the Appendix.
fixation cross screen for 8.25 seconds. The foods shown in the wake-up trials were different from those used in the passive viewing trials, and we did not use the neural responses from the wake-up trials in the prediction exercises described below. In 93.1% of the wake-up trials, subjects responded within the time allowed, which suggests that they generally attended to the images.⁸

We collected measures of neural activity using BOLD-fMRI (which stands for blood-oxygenated level dependent functional magnetic resonance imaging).⁹ Because we did not wish to make assumptions as to which brain regions that were likely to generate predictive non-choice responses, we measured activity throughout the entire brain.

BOLD-fMRI operates by measuring changes in local magnetic fields resulting from local inflows of oxygenated hemoglobin and outflows of de-oxygenated hemoglobin that occur when neurons fire. One complication is that the hemoglobin responses are slower than the associated neuronal responses. Specifically, although the bulk of the neuronal response takes place in 4 to 6 seconds, subsequent BOLD measurements are affected for as much as 24 seconds. Despite this consideration, the BOLD signal is widely thought to be highly correlated with aggregate neural activity within relatively small “neighborhoods” (tiny cubes, known as voxels). Furthermore, as long as trials are spaced sufficiently far apart, one can attribute most of the BOLD signal to trial-specific neural responses. In our experiment, each trial spanned a total of 11 seconds (2.75 seconds for an image, and 8.25 seconds for a fixation cross on a black screen), and BOLD measurements were obtained in 3-mm³ voxels every 2.75 seconds. With this design, the BOLD signal provides a good measure of neural responses to each image. This is an approximation, but it suffices for our purposes. Presumably, the predictive power of a sharper measure of neural activity would yield even greater predictive power than that of the somewhat noisy measure used here.

---

⁸ For one subject, we did not observe any responses to wake-up trials during the last two blocks, which suggests that he might have fallen asleep. This subject is included in our analyses, but excluding him does not affect our results substantially.

⁹ The fMRI data were collected at the Caltech Brain Imaging Center using a Siemens 3T Trio scanner. We acquired gradient-echo T2* weighted echo planar (EPI) images with BOLD contrast. We used an oblique acquisition angle of 30 degrees relative to the anterior commissure-posterior commissure line (Deichmann et al., 2003) and an 8-channel phased array head coil to maximize functional contrast-to-noise in areas of the ventro-medial prefrontal cortex which, as described in Section II, have been shown to play a critical role in valuation. Each volume consisted of 44 axial slices covering the entire brain. The imaging parameters were: echo time, 30ms; field of view, 192mm; in-plane resolution and slice thickness, 3mm; repetition time (TR), 2.75s.
Stage 2: Pairwise choices. The second stage of the experiment was conducted outside the scanner. Subjects were shown pairs of food items, and were asked to choose their preferred item from each pair. They were told that, with 50% probability, one of the pairs would be selected at random, and they would receive their choice from that pair at the end of the experiment.

The first ten subjects were shown 200 pairs of items constructed by drawing with replacement from the 100 foods viewed in stage 1. The remaining 17 subjects were shown 50 randomly selected pairs, with each item appearing in one and only one pair. As discussed below, the first procedure is not appropriate for some portions of our analysis (a fact which, unfortunately, we did not notice until we scrutinized some preliminary results based on the first ten subjects that were too good to be true). Accordingly, some of the results reported below are based on all 27 subjects, while others are based on the last 17.

Foods were randomly assigned to left and right positions on the screen. As is common in such tasks, there was a small but significant spatial bias: subjects chose the left item 53% of the time ($p<.05$, binomial test). When estimating our forecasting models, it is important to ensure that our predictions do not benefit artificially from this bias (as they would, for example, if we used models describing the probability of choosing the object displayed on the left). Accordingly, for every subject, the choice pairs were randomly divided into two equal sets; in one set, the chosen item was designated as the “target,” while in the other the item not chosen was so designated. The choice for any trial was then recorded as a 1 if the target item was chosen, and as a 0 otherwise. With this assignment, the unconditional probability that our discrete choice variable equals 1 in any given trial is exactly 50 percent, and the predictive success of more elaborate models must be judged against this neutral benchmark (rather than 53 percent). Because the target item is designated at random, spatial bias effectively introduces random variation into the discrete choice variable that is inherently not predictable from stage 1 measurements. Therefore, the existence of spatial bias necessarily reduces the predictive accuracy of the models we consider.

Stage 3: Preference ratings. In the final stage of the experiment, subjects were asked to indicate the extent to which they liked each food item, using a discrete scale from $-3$ (strongly dislike) to 3 (strongly like). They viewed pictures of all 100 items sequentially and entered liking ratings through button presses, proceeding at their own pace. They were
told that their ratings would not affect the item they received at the end of stage 3, but they were also strongly encouraged to provide ratings that reflected their true preferences.

After each subject finished entering liking ratings, we tossed a coin to determine whether he or she would receive an item chosen at random, or the item chosen in a randomly selected choice trail from stage 2 (where the item or choice trial was selected by drawing a number from an envelope). Subjects were then required to eat at least three spoonfuls of the selected item, and were allowed to eat more if desired. Subjects were instructed to remain in the lab for 30 minutes, during which time they were not permitted to eat anything else.

C. Initial data processing

Before analyzing the predictive power of non-choice BOLD responses, the raw data must be converted into a usable form. First, we corrected for head motion to ensure that the time series of BOLD measurements recorded at a specific spatial location is always associated with the same location throughout the experiment.10 Second, we removed low-frequency signals that are unlikely to be associated with neuronal responses to individual trials.11 Third, we realigned the BOLD responses for each individual into a common neuroanatomical frame (the standard Montreal Neurological Institute EPI template). This step, called spatial normalization, is necessary because brains come on different shapes and sizes, and as a result a given spatial location maps to different brain regions in different subjects. Spatial normalization involves a non-linear re-shaping of the brain to maximize the match with a target template. Although the transformed data are not perfectly aligned across subjects due to remaining neuroanatomical heterogeneity, the process suffices for the purposes of this study. Any imperfections in the re-alignment process simply introduce noise that reduces our ability to predict choices based on the neural responses.

For the analyses described in Sections V (which involve comparisons across subjects), we also spatially smoothed the BOLD data for each subject, by making BOLD

10 BOLD measurements were corrected for head motion by aligning them to the first full brain scan and normalizing to the Montreal Neurological Institute's EPI template. This entails estimating a six-parameter model of the head motion (3 parameters for center movement, and 3 parameters for rotation) for each volume, and then removing the motion using these parameters. For details, see Friston et al. (1996).
11 Specifically, we applied a high-pass temporal filter to the BOLD data with a cut-off of 128 seconds.
responses for each voxel a weighted sum of the responses in neighboring voxels, with the weights decreasing with distance.\textsuperscript{12} The purpose of this transformation is to address residual problems arising from neuroanatomical heterogeneity across subjects. In effect, smoothing assumes that any particular voxel in one subject’s brain can play the same predictive role as neighboring voxels in another subject’s brain; without smoothing, we would be assuming that only the same voxel can play the same predictive role.

The final step was to construct, for each subject and each voxel, the average non-choice neural response to each food item. We began by removing predicted neural responses that were related to the task (e.g., seeing the image of a food item) but common to all items.\textsuperscript{13} The object of this step is to increase the preponderance of the BOLD responses that are specific to the individual food items, and therefore likely to be helpful in predicting choices. Second, we averaged the residual response over the three presentations of each food item, collected 2.75 and 5.5 seconds after the onset of stimulus. In constructing this average, we omitted measurements from full brain scans (known as volumes) that exhibited excessive within-volume variation across voxels.\textsuperscript{14} This exclusion criterion reduces noise (and thereby improves predictive accuracy) by eliminating BOLD responses that are outliers with respect to the typical range of responses for food items.

\section*{IV. Predicting choices involving new items, within subjects and groups}

\textsuperscript{12} Smoothing was performed using an 8 mm full-width half-maximum Gaussian kernel.
\textsuperscript{13} We carried out this step by estimating a general linear model (GLM) of BOLD responses with an AR(1) structure. The model included the following regressors: 1) An indicator function for the moment at which the image of an item appears on the screen, convolved with a canonical hemodynamic responses function (Friston et al., 1998) that captures the manner in which neural responses are mapped to delayed changes in the BOLD signal, six block dummies, and the time series of head motion parameters estimated in the pre-processing step described above. The outputs of interest from this regression are the residuals, which capture the BOLD responses from each trial that are item-specific, and not a reflection of common responses. For reasons of practicality, we performed this calculation only for gray-matter voxels (of which there are approximately 45,000 per-subject). We identified gray matter in each subject using the Automated Anatomical Labeling (AAL) Tool and the MNI gray-matter mask (Tzourio-Mayer et al 2002).

\textsuperscript{14} For each volume we computed the variance across voxels (known as global signal variation), as well as the mean and standard deviation of this variance across volumes. We excluded data on volumes for which the within-area variance exceeded the mean plus five standard deviations.
As explained in the introduction, the canonical task motivating our investigation is to determine how people will behave when confronted with some new or difficult-to-observe choice situations. Imagine assembling a group of individuals, measuring their non-choice neural responses to prospects that we can actually implement, as well as to the new choice situations, and then presenting them with unanticipated choices among the implementable prospects. We can then estimate the relationship between their choices and non-choice responses for the implementable prospects, and use that relationship along with non-choice neural responses for the new situations to predict behavior in the new situations. Do the non-choice neural responses contain enough information to make reasonably accurate predictions?

In this section, we implement the procedure outlined in the previous paragraph and use it to make and evaluate predictions both within subjects and within groups. Subsection A investigates the accuracy of within-subject predictions. Accuracy on the level of a single individual is a demanding objective, one that goes beyond the requirements of most economic analyses, which are more typically concerned with aspects of group behavior – averages, aggregates, and possibly distributions. For example, a per capita demand curve is ordinarily of greater interest than the demand curve for a specific person. It is therefore worth emphasizing that averages and distributions of choices among groups may well be easier to predict than the choices of individuals, for at least two reasons. First, individual-level predictions suffer due to the presence of substantial noise in neural measurements, which should presumably average out over multiple subjects. Second, independent random variation in choice should also average out over subjects. For example, if a model tells us only that an individual will choose the target item from a given pair with 50% probability, we would view it as providing no value over an uninformed benchmark. But if the realizations are independent across individuals, the same model predicts with high confidence that, in a large group of subjects, the fraction picking the target item will be very close to one-half. Accordingly, in Subsection B, we shift our attention to the potentially easier (and arguably more important) objective of predicting the average behavior of a group of individuals from their average neural responses.

A. Within-subject predictions

In this subsection, we focus on the accuracy of within-subject predictions. For reasons detailed below, we performed this analysis using subjects eleven through twenty-
seven, each of whom made decisions for 50 pairs of food items, with no item appearing twice.

1. Statistical methods

We adopt a logit probability model for choices. To describe the model formally, we introduce the following notation. For every subject $i$ and choice pair $t$, let $y_{it} = 1$ if the target food was chosen, and $y_{it} = 0$ otherwise. For every brain voxel $v$ and choice pair $t$, let $d_{itv}$ denote the difference between the measured neural responses in voxel $v$ to the target and non-target food items offered in choice pair $t$ (i.e., the response for the target food minus the response for the non-target food). Also let $D_t$ denote the vector of differential neural responses for all voxels. Our model is described by the following probability statement:

$$\Pr(y_{it} = 1|D_t) = \frac{\exp(y_0 + \gamma D_t)}{1 + \exp(y_0 + \gamma D_t)}$$

Because our object is accurate out-of-sample prediction, we employ standard methods for estimating and evaluating forecasting models. A central tenet of the forecasting literature is that within-sample fit may be a poor gauge of out-of-sample fit. Accordingly, one generally proceeds by dividing the sample into a “training sample” which is used for estimation, and a “hold-out” sample which is used to evaluate predictive performance. By removing two choice pairs at a time from the set of 50, we create 25 training samples (each consisting of 48 observations) and 25 associated hold-out samples (each consisting of 2 observations). For each training sample, we then estimate the model and use it to predict the choice for the associated hold-out observations (which are not used in the estimation). We then assess the model’s out-of-sample predictive performance over all 25 hold-out samples (50 predictions in all).

To ensure the representativeness of both the training and hold-out samples, we randomly partitioned the 50 choices into 25 pairs, with each pair containing one choice from which the target item was chosen, and another from which it was not chosen. Each such pair served as a hold-out sample, and the complement served as a training sample. This procedure yields training and hold-out samples in which the target item is chosen exactly 50 percent of the time, just as in the full sample (by construction).

---

15 As described in the previous section, one item in every pair was randomly designated as the target.
As the literature recognizes, evaluating the predictive performance of a categorical probability model involves some inherent ambiguities. Alternative standards for defining a "predicted outcome" have been proposed. In the context of binary models, Cramer (1999) proposes identifying an alternative as the predicted outcome if its predicted probability exceeds its baseline frequency in the population.\textsuperscript{16} By construction, in our experiment, the baseline frequency for selecting the target item is exactly 50%. Consequently, we classify the target item as the predicted choice if its predicted probability exceeds 50%; otherwise, the non-target item is the predicted choice. We classify a particular prediction as a "success" if the predicted item was in fact chosen.

Notice that our task involves prediction from small samples (48 observations). It therefore raises two important and closely related issues: variable selection and overfitting.

The variable selection problem is obvious: because we estimate each model with only 48 observations, we cannot use all 45,000 potential predictive variables (voxel-specific BOLD signals). Rather, we must eventually focus on a small handful of predictors, in effect leaving out large numbers of presumably relevant variables. If we intended to interpret estimated parameters as reflecting causal effects, the left-out variable problem would be fatal. Accordingly, it is essential to emphasize that our objective here is purely prediction. When predicting from a small sample, it is worthwhile to include a variable only if the incremental predictive information it carries is sufficient to justify sacrificing a scarce degree of freedom. Thus, for example, when two important causal factors are highly correlated, it is often appropriate to include only one (when predicting from a small sample), because each reflects most of the predictive information contained in the other. Naturally, with either factor omitted, the coefficient of the included factor will not measure its causal effect; on the contrary, that coefficient will be "biased" (relative to the factor's causal effect) due to contamination from the causal effect of the omitted factor. Even so, the omission of a causal factor does not impart a bias to predictions (conditional on the included predictors), and may well reduce variance. A variety of criteria have been developed to assist with the task of variable selection for predictive models; they include the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), cross-validated predictive performance, LASSO (which we describe and implement below), and others.

\textsuperscript{16} Even that alternative is recognized as somewhat arbitrary; see Green (2003), p. 685.
The overfitting problem arises because, as sample size shrinks, it becomes more likely that some predictor will be highly correlated (within sample) with the outcome variable purely by chance. Standard regression techniques will assign a large coefficient to that predictor. Unfortunately, because the largest within-small-sample correlations tend to shrink out of sample, estimated models (though potentially unbiased) will tend to overpredict out-of-sample responses to the variables that seem most important within sample. The most obvious case occurs when the number of predictors equals the number of observations. In the context of our analysis, any combination of 48 independent predictors will yield a perfect fit within sample, but the resulting model will generally perform very poorly out of sample.

Various techniques have also been developed to address the overfitting problem. For example, ridge regression compensates for the fact that the largest within-small-sample correlations tend to shrink out of sample by shrinking the estimated coefficients (we describe how this is accomplished below). Such shrinkage estimators generally impart some bias to predictions, but reduce variance, and thus can improve out-of-sample predictive performance according to measures such as mean-squared prediction error.

We address the variable selection and overfitting issues simultaneously using LASSO (the Least Absolute Shrinkage and Selection Operator; see Tibshirani, 1996). As the name implies, LASSO, like ridge regression, is a shrinkage procedure. For both procedures, one optimizes a standard criterion for within-sample fit (for example, minimizing the sum of squared residuals in the case of a regression, or maximizing likelihood) subject to a penalty that increases monotonically in the size of the coefficient vector. For ridge regression, one measures the size of the coefficient vector using the L2-norm (i.e., the square root of the sum of squared coefficients). For LASSO, one uses the L1-norm (i.e., the sum of the absolute coefficients).

While both methods of penalization lead to shrinkage, only LASSO also accomplishes variable selection. Relative to an L2-penalty, an L1-penalty favors coefficient vectors wherein some elements equal zero. Notice, for example, that in a model with two coefficients, $\gamma_1$ and $\gamma_2$, as we move linearly from $(\gamma_1, \gamma_2) = (\alpha, 0)$ to $(\gamma_1, \gamma_2) = (\alpha/2, \alpha/2)$,

---

17 In a linear regression context, one can also interpret LASSO as a Bayesian regression with double exponential priors; see Tibshirani (1996). In the Bayesian context, shrinkage results from the priors.
the L1-penalty remains constant while the L2-penalty declines monotonically. More importantly, because iso-penalty curves are smooth when one uses the L2-norm, solutions involve coefficients of zero only by coincidence. In contrast, because iso-penalty curves are kinked at the axes when one uses the L1-norm, solutions typically involve setting many coefficients equal to zero.

In our context, the LASSO procedure involves maximizing the following penalized log-likelihood function over the parameters $\gamma_0$ and $\gamma$:

$$
\frac{1}{T} \sum_{t=1}^{T} \left[ y_{it} \log[\Pr(y_{it} = 1|D_t)] + (1 - y_{it}) \log[1 - \Pr(y_{it} = 1|D_t)] \right] - \lambda \sum_v |\gamma(v)|
$$

where $T$ denotes the number of trials in the training set.

In the LASSO objective function, the L1 penalty receives a weight of $\lambda$. Larger values of $\lambda$ lead to greater shrinkage and more aggressive variable selection. The best value of $\lambda$ is generally determined through cross-validation, which is effectively a procedure for simulating out-of-sample predictive performance entirely within a training sample. For our analysis, we randomly divided each training sample into five sets of approximately equal size, indexed $k = 1, \ldots, 5$ (called folds in the statistical prediction and machine learning literatures). For each $k$, we estimated the penalized regression model for each possible value of $\lambda$ in a pre-specified grid using only the data from the $k - 1$ other folds. We then used the estimated models to predict choices using the neural data for the left-out fold, and computed the accuracy of the predictions by comparing them to the actual choices. The value of $\lambda$ with the best prediction rate across all of the folds, $\lambda^*$, was then used to estimate the model with all of the observations in the training sample. Importantly, note that the selection of $\lambda^*$ is completely blind with respect to outcomes in the actual hold-out sample; thus, accuracy within the hold-out samples remains a valid gauge of the procedure’s out-of-sample performance.

With LASSO, selecting $\lambda^*$ through cross-validation not only accomplishes optimal shrinkage, but also in effect ensures that a variable remains in the model with a non-zero coefficient only if its incremental predictive value is sufficient to justify the sacrifice of a degree of freedom. Thus, in our setting the procedure selects the brain voxels with the
neural responses that provide the most valuable predictive informative concerning subsequent choices.

Prior to implementing the LASSO procedure, we reduced the vast set of candidate voxels by excluding those that failed to meet a simple screening criterion. Ryali et al. (2010) have shown that this initial screening step can improve predictive accuracy in studies employing fMRI data, even when the subsequent estimation procedure selects voxels automatically (as is the case with the LASSO procedure used here). Our initial selection step proceeded as follows. For every voxel, we computed a simple two-sided t-test for the hypothesis of no difference between neural responses (within the training sample) to foods that were chosen and those that were not. We then ranked voxels by their t-statistics, and retained only those exceeding some threshold percentile.18 Except where otherwise specified, we present detailed results based on analyses for which the top 1% of voxels (roughly 450) were retained, but we also examine robustness with respect to alternative screening criteria. Intuitively, the purpose of this initial screening step is to focus attention on voxels that are likely to contain highly predictive information. Note that the voxel selection procedure, like the selection of $\lambda^*$, is completely blind with respect to outcomes in the actual hold-out sample; thus, accuracy within the hold-out samples remains a valid gauge of the procedure’s out-of-sample performance.

As we mentioned at the outset of this section, data gathered from our first 10 subjects were not used for within-subject predictions. Recall that those subjects made choices from 200 pairs of items, drawn randomly with replacement from our set of 100 items. Accordingly, when the full sample is divided into a training sample and a hold-out sample, the items that belong to pairs in the hold-out sample also typically belong to pairs in the training sample. The resulting overlap between the sets of items represented in the training and hold-out samples can lead to spurious predictive power. To understand why, suppose for the purpose of illustration that the subject’s choices are pair-wise transitive. From the choices in the training sample, one can then predict many choices perfectly out of sample. For example, if the individual chooses $a$ over $b$ as well as $b$ over $c$ in the training sample.

18 The problems with using t-statistics for variable selection are well-known; see, for example, Greene (2003), p. 151. While the screening step is not ideal from the perspective of statistical theory, Ryali et al. (2010) demonstrate that it tends to improve predictive performance in these types of settings. We note that any superior procedure for variable selection would presumably yield predictions that are even more accurate than the ones presented in subsequent sections.
sample, we can confidently predict that he will pick $a$ over $c$ out of sample; no neural information is required. This observation is problematic for our investigation because, with 45,000 voxels, there is a substantial likelihood that for each item there will be some voxel within which neural activity was high when the item in question, and only that item, was presented. That voxel becomes a spurious neural identifier for the item. LASSO will tend to latch on to those voxels and assign them coefficients that reflect each item’s place in the subject’s preference ordering. In the preceding example, it might assign a coefficient of 1 to the voxel identifying item $a$, 0 to the voxel identifying item $b$, and -1 to the voxel identifying item $c$. Accordingly, the resulting model will predict that $a$ will be chosen over $c$ out of sample, but only because the neural activity spuriously identifies the item, and not because it is correlated with some provisional assessment of subjective value. We avoided this potential problem by selecting the choice pairs for subjects 11 through 27 so that each item appeared in one and only one pair.\(^{19}\)

2. Results

Figure 3 plots the mean success rates, defined as the fraction of hold-out observations for which the predicted item was chosen, as a function of the fraction of voxels retained after initial screening, with the retained fractions ranging from 0.0001 (0.01% of voxels), to 1 (all voxels). Our procedure maximizes predictive accuracy when 1% of voxels are retained. The mean success rate is then 61.3%, which represents an economically and statistically significant improvement over the uninformed 50% benchmark ($p<.0001$, one-sided t-test). Performance falls sharply when fewer than 1% of voxels are retained in the initial screening step, but declines only slightly when fewer are eliminated. Indeed, when we abandon the initial screening step (i.e., retain 100% of the voxels), our overall success rate, 59.3%, remains significantly better than the uninformed benchmark ($p<0.001$), and is not significantly different from the rate obtained when retaining 1% of voxels ($p=0.23$, paired t-test). Thus we find, in contrast to Ryali et al. (2010), that the initial voxel screening step yields only a small and statistically insignificant improvement in predictive accuracy. For the remainder of this section, we will focus on the results obtained using the 1%
screening criterion (which maximizes the success rate); our conclusions are not substantially affected by applying less restrictive screens.

The first data column in Table 1 provides results on predictive accuracy for each subject (numbered 11 through 27 because this analysis excludes the first ten subjects). As the table shows, there was considerable cross-subject variation in success rates, which ranged from a low of 44% to a high of 76%, with all but one exceeding 50% and four exceeding 70%. Predictive accuracy exceeded the uninformed benchmark by a statistically significant margin for 9 out of 17 subjects at the 5% level (amongst whom the overall success rate was 68%), and for 8 out of 17 subjects at the 1% level. On the basis of these results alone, we can unambiguously conclude that non-choice neural responses contain a substantial amount of predictive information for a large fraction of subjects. For subsequent reference, we have shaded all of the rows in the table associated with high-success-rate subjects (i.e., those whose success rates exceeded the uninformed benchmark by statistically significant margins), so that their results are easily distinguished from those of low-success-rate subjects (i.e., the complementary set).

In interpreting success rates, it is helpful to distinguish between what we will call the strength and accuracy of a probabilistic prediction. By strength, we mean the degree of certainty. The statement that an individual will choose the target item with either 1% or 99% probability is extremely strong, while the statement that he will choose that item with either 49% or 51% probability is extremely weak. By accuracy, we mean the degree to which the probabilistic prediction is born out in practice. To illustrate, suppose that for some group of pairwise choices, a model predicts that the more likely items will be chosen with an average probability of 75%. Then, according to the model, the expected success rate for those items is 75%; if the model is correct and the set of choices is reasonably large, the realized success rate should be close to 0.75. Thus, we say that the model's predictions are accurate if the success rate for any reasonably large group of observations is close to the average predicted probability of the more likely item. If it is not close, we say that the model's predictions are inaccurate. In the preceding example, if the realized success rate is 55% rather than 75%, the model's probabilistic predictions are plainly inaccurate. The same is true if the success rate is 95%; in that case, while the model might appear to predict choices quite well, its probabilistic predictions would nevertheless be substantially off.
According to these definitions, we would classify the predictions of the uninformed benchmark (50-50) as extremely weak, but nevertheless highly accurate (because the overall success rate, 50%, exactly matches the predicted probability of the most likely item in every pair). In contrast, we would classify the typical deterministic model as very strong, but in all likelihood highly inaccurate (because it is rarely possible to forecast outcomes with certainty).

Knowing only that the average success rate for our procedure is 61.3%, one cannot say whether the underlying predictions are strong or weak, accurate or inaccurate. Yet such distinctions are plainly crucial. If our procedure typically yielded predicted probabilities on the order of 90% but achieved an overall success rate near 60%, its success would be only directional, and one would not be able to take its probabilistic predictions seriously. On the other hand, if on average our procedure yielded predicted probabilities near 60% (i.e., in line with the observed success rate), then although one could complain that its predictions were somewhat weak, at least they would be highly accurate.20

With respect to the aforementioned potential complaint concerning low predicted probabilities, it bears emphasizing that the value of an accurate predictive model should not be discounted merely because its predictions are not as strong as one might like. On the individual level, certain determinants of choice may be fundamentally unpredictable (see, e.g., Krajbich, Armel, and Rangel, 2010), in which case the strength of any accurate probabilistic prediction is necessarily limited. Fortunately, such idiosyncratic randomness likely averages out over multiple decisions, so it should still be possible to predict the average behavior of groups with a high degree of accuracy (see Sections VI and VII).

The second data column in Table 1 sheds light on the strength of our procedure’s predictions. Focusing for the moment on the second-to-last row, we see that the mean predicted probability of the more likely item is 72.7%. Thus, our predictions are reasonably strong, but not remarkably so. More significantly, there is a sizable and highly statistically significant gap (or bias) of 11.4 percentage points between the mean predicted probability and the overall success rate of 61.3% ($p < 0.001$). At this level of aggregation, the models’ predictions certainly cannot be classified as highly accurate.

20 As explained below, further investigation would be required before reaching that conclusion.
A careful examination of the results for individual subjects tells a more interesting and nuanced story. Based on our initial analysis of success rates for individual subjects, it is entirely possible that our procedure works well for some subjects, and poorly (or not at all) for others. For example, some subjects simply may not meaningfully attend to the images of food items during stage 1; indeed, this is a common problem in methodologically related studies.

It is natural to conjecture that subjects with high success rates are those for whom the procedure has worked well, and that subjects with low success rates are those for whom it has not. A preliminary (albeit somewhat weak) test of that hypothesis involves comparing success rates with mean predicted probabilities separately for high-success-rate and low-success-rate subjects. The fourth column of Table 1 contains the p-values for those subject-specific tests. Comparing the shaded and unshaded lines, we see a striking pattern. We cannot reject equality of the success rate and the mean predicted probability with 95% confidence for any of the high-success-rate subjects, and we reject equality with 90% confidence for only two of these nine subjects (and would have expected roughly one rejection by chance). In contrast, we reject equality at the 90% confidence level for seven of the eight low-success rate subjects (and with 88% confidence for the eighth), at the 95% confidence level for five of the eight, and at the 99% confidence level for three of the eight. Visually, asterisks (indicating levels of statistical significance) tend to appear in the first data column when no asterisks appear in the fourth, and vice versa.

Overall, for high-success-rate subjects, the mean success rate is 68.2%, while the mean predicted probability is 72.9%; the difference (or bias), 4.7 percentage points, is modest but statistically significant \((p = 0.042)\). Though the predictions are not right on the mark, they are in our view impressively close given the nature of our out-of-sample prediction exercise. Interestingly, our predictions are equally strong for the low-success-rate subjects: the mean predicted probability is 72.6%. However, the mean success rate is only 53.5%, and the difference (or bias), 19.1 percentage points, is large and highly statistically significant \((p < 0.001)\).

One might be tempted to discount the results of the preceding test as a possible coincidence. After all, if the overall success rate is below the overall mean predicted probability, and if the latter does not vary between low- and high-success-rate subjects, then it is not too surprising that the success rate for high-success-rate subjects is closer to
that group's mean predicted probability. Accordingly, we view this first test as providing only a relatively weak preliminary indication concerning the model's performance among high-success-rate subjects.

Fortunately, a much more demanding test is available. So far, we have made no use of variation in the strength of predictions across hold-out observations (e.g., whether the predicted probability of choosing the target item is 51% or 98%). According to Table 1, the mean within-subject standard deviation of the predicted probability is 0.140, which indicates considerable variation. Moreover, as shown in Figure 3, the predicted probability of the more likely item is distributed fairly evenly between 50% and 100%. Using this variation allows us to determine whether our predictive procedure is functioning properly. Imagine, for example, that the predicted probability averages 60% within one large group of hold-out observations, and averages 80% within a second group. If the model is generating valid out-of-sample probabilities, the frequency with which the target item is chosen should be approximately 60% in the first group, and approximately 80% in the second. Even if the model is just capturing tendencies, that frequency should be noticeably higher in the second group than in the first.

We implement this idea as follows. First, we rank the hold-out observations (pooled across all subjects) according to the predicted probability of the more likely choice (i.e., the probability of choosing the target item if the model indicates that the target is more likely, and the probability of choosing the non-target item if the model indicates that the non-target item is more likely). Second, we divide the observations into deciles based on the aforementioned probability. Third, for each decile, we compute the frequency with which the item identified as more likely was in fact chosen (i.e., the success frequency). Finally, we examine the relationship between the average predicted probability of choosing the more likely item and the actual frequency with which that item was actually chosen across deciles.21

Figure 4A plots the results, pooled over all subjects. The horizontal axis shows the predicted probability of choosing the more likely item, while the vertical axis shows the frequency with which that item was actually chosen. For an ideal predictive model, the data

21 This procedure is motivated by and closely related to a goodness-of-fit test for binary choice models described Lemeshow and Hosmer’s (1982).
points would line up along the 45 degree line (i.e., the predicted probabilities and the success frequencies would always coincide). Though our procedure does not achieve this ideal, there is nevertheless an obvious and reasonably strong positive relationship between the predicted probabilities and success frequencies. Between the first and eighth deciles, the actual success rate increases roughly half a percentage point for every one percentage point increase in the predicted probability; beyond the eighth decile, it declines modestly. Overall, the predictive performance of the model is encouraging, at least directionally.

Figure 4B performs the same analysis separately for low-success-rate and high-success-rate subjects. The results are striking. For the eight low-success-rate subjects, there is no apparent relationship between success frequencies and predicted probabilities: the red line moves up and down a bit, but overall appears reasonably flat. With these problematic subjects removed, the procedure’s performance is much improved. For the nine high-success-rate subjects, the relationship between success frequencies and predicted probabilities (shown as a blue line) increases more sharply than the one in Figure 5A, and is much closer to the ideal (i.e., the 45 degree line). For the lowest two deciles, within which the average predicted probability is 53.8%, the overall success frequency is 56.7%, while for the highest two deciles, within which the average predicted probability is 92.7%, the overall success frequency is 84.4%.

To sharpen these impressions, we conduct additional statistical analyses. For each subject $i$ and choice trial $t$, we define a binary success indicator, $S_{it}$, which equals unity when the subject chooses the item predicted as more likely (with this trial treated as a hold-out observation), and zero otherwise. Let $P_{it}$ denote the predicted probability that the subject $i$ will choose the item identified as more likely in choice trial $t$ (again, when this choice trial is treated as a hold-out observation). Assuming that $P_{it}$ is in fact a correct probability, it follows trivially that $E(S_{it} | P_{it}) = P_{it}$. Thus, $S_{it} = P_{it} + \epsilon_{it}$, where $E(\epsilon_{it} | P_{it}) = 0$ (in particular, $\epsilon_{it}$ equals $1 - P_{it}$ with probability $P_{it}$, and $- P_{it}$ with probability $1 - P_{it}$). Accordingly, our strategy is to estimate simple linear probability models (LPMs) of the following form:

$$S_{it} = \alpha + \beta P_{it} + \epsilon_{it}$$

If the predicted probability statements are in fact correct, by the preceding reasoning we should obtain $\alpha = 0$ and $\beta = 1$.  

25
We began by estimating two versions of the preceding linear probability models, one for the nine high-success-rate subjects, and one for the eight low-success-rate subjects. In these regressions, each observation consists of a single hold-out choice pair; thus, the regression for high-success-rate subjects uses $50 \times 9 = 450$ observations, while the regression for low-success-rate subjects uses $50 \times 8 = 400$ observations.

When we estimate the linear probability model for the eight low-success-rate subjects, we obtain an intercept of 0.551 (s.e. = 0.129) and a slope of $-0.023$ (s.e. = 0.174). The combination of low success rates and the absence of any detectable relationship between the two variables indicates that our forecasting procedure fails completely for those subjects. In contrast, when we estimate the same model for the nine high-success-rate subjects, we obtain an intercept to 0.129 (s.e. = 0.113) and a slope 0.759 (s.e. = 0.152). Here, the relationship between the two variables is strong, positive, highly statistically significant, and within the general vicinity of the ideal. However, we reject the hypothesis that the intercept is in fact zero and the slope unity ($p = 0.027$). With that qualification, our prediction model performs demonstrably well out of sample for the nine high-success-rate subjects.

In principle, the strong results obtained for the LPM estimated with high-success-rate subjects could be attributable to compositional effects. For example, success rates might be unrelated to predicted probabilities within subject, but subjects with higher success rates might also have higher predicted probabilities. In practice however, Table 1 provides little reason to anticipate significant compositional effects. Notice that the means and standard deviations of the predicted probabilities (the second and third data columns) are quite similar across subjects. Moreover, the cross-subject standard-deviations of these statistics are only 0.028 in the case of the within-subject mean, and 0.010 in the case of the within-subject standard deviation.

To completely eliminate the possibility that our LPM results for high-success-rate subjects are driven by compositional effects, we estimate an additional LPM with subject-fixed effects. Our estimate of $\beta$ increases to 0.798 (s.e. = 0.157). We also estimate an LPM separately for every subject. The slope coefficients and associated standard errors are reported in the last two data columns of Table 1. Each regression employs only 50 observations, so as one would expect, the standard errors are large. Still, the overall pattern is striking. For the high-success-rate subjects, the slopes are all positive and range from a
low of 0.144 to a high of 1.632. The mean slope for this group is 0.784, and the median is 0.840, with three of the nine slopes exceeding unity. In sharp contrast, for the low-success-rate subjects, five of the eight slopes are negative. They range from a low of −0.416 to a high of 0.649, with a mean of −0.010 and a median of −0.190.

We conclude that our within-subject procedure for predicting choices involving new items performs with considerable success for roughly half (nine of seventeen) of our subjects. The overall success rate is 68% for that group, and subject-specific success rates are close to subject-specific mean predicted probabilities. Moreover, success frequencies mirror predicted probabilities across hold-out observations, both overall and within subjects. The predicted probabilities are not always spot-on for this group, but they are close. We of course acknowledge that the procedure works dismally for the rest of our subjects: the overall success rate is only 54%, subject-specific success rates differ considerably from subject-specific mean predicted probabilities, and success frequencies bear no discernable relation to predicted probabilities across hold-out observations.

B. Within-group predictions

Our investigation in this subsection parallels that of Section 4.A, except that we concern ourselves here with average behavior among groups, rather than the choices of specific individuals. Thus, our objective is determine whether the average non-choice neural responses among a group of individuals contain enough information to make reasonably accurate predictions concerning the group’s average behavior in new situations, using a model estimated with data concerning the same group.

We will attempt to predict measures of subjective valuation, averaged across group members. A natural alternative would have been to predict the fraction of subjects choosing the target item from a given pair. That alternative is, however, inconsistent with our experimental design, which employed different random pairings of the items for different subjects.

As explained in Section III, stage 3 of our experiment elicits preference ratings (on a scale of −3 to +3) for each item from every subject. We acknowledge that these ratings do not provide cardinally meaningful measures of willingness-to-pay (WTP), and that our elicitation protocol is not incentive-compatible, but we nevertheless offer the following observations in defense of our short-cut for measuring subjective valuation. First,
preference ratings were elicited after the subjects made incentivized choices, from which it follows that (i) subjects had already thought about their preferences for each item in an incentive-compatible context, and (ii) subjects were likely to provide ratings that rationalized their choices. Second, ratings were in fact highly correlated with choices: subjects choose the item with the highest rating 85.1% of the time ($p<10^{-12}$, one-sided t-test vs. chance) in the 50-choice condition and 90.1% ($p<10^{-8}$, one-sided t-test vs. chance) of the time in the 200-choice condition. Third, to the extent preference ratings are noisy measures of subjective valuation, our results likely understate the true predictive power of non-choice neural responses.

1. Statistical methods

Before aggregating subjective ratings across our 27 subjects, we normalized each subject’s ratings using a z-score transformation. We then computed the mean normalized ratings for the group, denoted $Z_j$ for item $j$, as well as the group’s mean non-choice neural responses, denoted $M_j$ for item $j$, where $M_j$ is a vector containing the average neural response for each voxel $v$, denoted $M_{vj}$. The distribution of mean normalized ratings across food items is shown in Figure 5.

As a first step, we simply ask whether the average non-choice neural responses to an item predict whether its average subjective rating is above or below the median rating (denoted $Z_{med}$). This is a natural and interesting comparison because it stands in for a binary choice between the item in question and the median-rated alternative. We assume that the probability of an above-median rating for any item $j$ is given by the logistic function

$$
\Pr(Z_j > Z_{med} | M_j) = \frac{\exp(y_0 + \gamma M_j)}{1 + \exp(y_0 + \gamma M_j)}
$$

Plainly, realizations of this process cannot be independent across items (because half of the items must be above the median). However, with a sufficient number of items, correlations

---

22 The probability of any item falling above the median clearly depends on the entire vector of neural responses to all items. However, in our analysis, that vector is identical for all items (because all items all part of the same group); consequently, we suppress it in the notation.
across observations are presumably small, so we ignore them and treat the model as a simple approximation of the true process.

By removing two items at a time from the set of 100, we create 50 training samples (each consisting of 98 observations) and 50 associated hold-out samples (each consisting of two observations). For each training sample, we then estimate the model and use it to predict whether the average valuations for the hold-out observations (which are not used in the estimation) will fall above or below the median valuation of items within the training sample. We then assess the model’s out-of-sample predictive performance over all 100 predictions. We classify a prediction as a success if the item’s average subjective rating falls into the half of the training sample rating distribution that the model identifies as more likely.

As in the previous section, we applied a screening criterion to reduce the number of candidate voxels prior to estimating the model for any given training sample. Using only the training data, for each voxel \( v \) we regressed \( M_{vj} \) on a binary variable indicating whether \( Z_j \) was above \( Z_{med} \). We then ranked the voxels according to the t-statistics of the slope coefficients and, for our baseline calculations, retained the highest 1% (as in the previous section; see below for a discussion of alternatives). Then we estimated the probability model using the LASSO procedure, selecting the penalty parameter through 5-fold cross-validation, where the folds were assigned at random.

The second step in our analysis of group behavior was to predict the actual value of \( Z_j \), an item’s average subjective rating across all subjects, rather than a binary indicator of its position relative to the median. For this analysis, we employed a LASSO-penalized linear regression of \( Z_j \) on \( M_j \). In the initial screening step, for each voxel \( v \) we regressed \( M_{vj} \) on \( Z_j \), then ranked all voxels by the t-statistics of the slope coefficients, and retained the highest 1%. All other procedures were identical, except that the LASSO penalty parameter, \( \lambda^* \), was chosen to maximize cross-validated mean-squared-error (which is appropriate here given that the objective is to predict a continuous variable).

As mentioned previously, the data gathered from our first 10 subjects are suitable for this analysis. Only the stage 2 choice data for those subjects have the feature that a single item plays a role in more than one observation (which produces violations of the
assumed separation between training and hold-out samples), and we do not use those data here. Thus, throughout this section we present results based on all 27 subjects.

2. Results

We begin with an analysis of predictions concerning the probability that the average subjective rating for a given hold-out item will fall above the median rating for items in the training sample. Figure 2 plots the overall success rate as a function of the fraction of voxels retained after initial screening, with the retained fractions ranging from 0.0001 (0.01% of voxels), to 1 (all voxels). Our procedure maximizes predictive accuracy when 0.5% of voxels are retained. The overall success rate is then 77%, which represents an economically and statistically significant improvement over the uninformed 50% benchmark ($p < 0.001$, one-sided $t$-test). Performance falls sharply when fewer than 0.5% of voxels are retained in the initial screening step, but is fairly robust when fewer are eliminated, with success rates generally exceeding 70%. Recalling that classifications of ratings relative to the median stand in for binary choices between any given item and an alternative of median value, we note that our procedure achieves a significantly higher overall success rate for within-group predictions than for the within-subject predictions discussed in Sections IV.A (compare the two lines in Figure 2). To avoid cherry-picking results section-by-section, we will adopt the same screening criterion here as in the previous section (1%), which yields a success rate of 75%, rather than the success-rate-maximizing 0.5% criterion. Our conclusions are not substantially affected by applying less restrictive screens.

Figure 5A employs a scatterplot to illustrate the relationship between the predicted probability of an above-median rating and an item’s average rating. Each data point corresponds to a food item; circles represent correctly classified items, while crosses represent incorrectly classified items. An unmistakable and strong positive relationship between the variables is easily discerned. Thus, our model plainly tends to predict higher probabilities of above-median ratings for more highly rated items.

As in Section IV.A, we perform an initial test of the validity of the model’s predictive probability statements by comparing the strength of the typical prediction with the overall success rate. On average, the model predicts that items will fall into the more likely half of the rating distribution with 79% probability. This figure is close to the actual success rate (75%), and the gap is statistically insignificant ($p = 0.388$, two-sided $t$-test).
For a more discerning assessment of the model’s predictive validity, we grouped items into quintiles (20 items in each) based on the predicted probability that the item’s average rating exceeded the median, and then, for each quintile, computed the frequency with which the group’s ratings of those items actually fell above the median. Results appear in Figure 6B. A strong positive relationship between predicted probabilities and realized frequencies is readily apparent. Thus, for example, when the model places the likelihood of an above-median rating in excess of 90 percent, the odds are indeed roughly nine in ten that the item’s rating was in fact above the median. While the five data points do not line up along the 45 degree line, the empirical relation bears a noticeable resemblance to that ideal.

To sharpen this impression, we estimated a linear probability model relating a binary variable indicating whether an item’s average rating was above the median to the predicted probability of that event. The estimated intercept is 0.099 (s.e. = 0.088), and the slope is 0.721 (s.e. = 0.140). We reject the joint hypothesis that the intercept is zero and the slope is unity with 90% confidence ($p = 0.076$). Although the point estimates may not support a literal interpretation of the model’s predictive probability statements, on the whole its quantitative out-of-sample performance is rather promising.

Next we turn to predictions of the average rating itself, rather than its relation to the median. Figure 7 is a scatterplot of average normalized ratings against predicted ratings. Although the predictions are by no means exact, once again a strong positive relationship is immediately evident. To summarize that relation, we regress the actual rating on the predicted rating, and plot the regression line in the figure. If predicted ratings were unbiased, then the actual ratings would equal the predictions plus random noise. In that case, our regression would yield an intercept of zero and a slope of unity. In practice, we obtain an intercept of -0.012 (s.e. = 0.060) and a slope of 0.712 (s.e. = 0.144). Here we fail to reject the joint hypothesis of interest with 90% confidence ($p = 0.136$). Overall, the predicted ratings account for 20% of the variation in the actual ratings.

We conclude that our within-group procedure for predicting the average ratings of new items performs with considerable success. For the binary prediction task, the overall success rate is 75%, considerably higher than for within-subject predictions, and predicted probabilities match up reasonably well with realized frequencies. Predicted ratings also track average ratings and plainly contain usefully predictive information.
Conceivably, one might achieve greater predictive accuracy by conditioning on higher moments of the distribution of predicted ratings. Likewise, it may be possible to predict additional parameters of the distribution of actual ratings, such as variance. These are important questions, but we leave them for future research.

C. Discussion

The results in this section demonstrate that non-choice neural reactions to images of potentially desirable objects contain a great deal of information that can be used to predict decisions made by a particular individual, or average decisions made by a group of individuals, in completely novel choice situations. Thus, we have provided unambiguous proof-of-concept for the feasibility of the agenda outlined at the outset of this paper.

That said, it is also important to acknowledge the limitations of our analysis, and to emphasize that it represents only a first step. Our procedure is entirely unsuccessful for nearly half of our subjects. Moreover, even for subjects to whom it is applied successfully, in many instances it yields relatively weak predictions (e.g., predicted probabilities near 50 percent rather than 100%), and consequently achieves only a moderate overall success rate (68.2%).

We would note, however, that our procedure also yields strong predictions in many instances, and that those predictions are associated with very high success rates (in the case of the subjects for whom the procedure is successful, as well as for groups). In addition, there is every reason to believe that refinements of the procedure will ultimately yield substantial improvements in predictive accuracy. Better methods can be developed to enhance attentiveness in the scanner and to weed out inattentive subjects. Advancements in knowledge of the brain and improved statistical methods may provide better guides to voxel selection. And technological advances will undoubtedly enhance our ability to detect and measure stimulus-specific neural responses.

V. Predicting choices across subjects and groups

The method of prediction developed and implemented in the previous section requires the use of separate forecasting models calibrated to each individual or group. If non-choice neural activity exhibits a sufficiently similar relation to choice across subjects,
then it should be possible to construct a single prediction model and use it without recalibration to predict choices based on neural measurements taken from new individuals or groups. Such a model would have considerable practical value in that, once estimated, it would vastly simplify the steps required to formulate additional predictions. In particular, to predict behavior in new situations, one could simply collect data on non-choice neural responses to the relevant prospects for a new group of individuals, and apply the existing model. It would not be necessary to collect new measurements from the same set of individuals used to estimate the original model, or to re-estimate the model with additional data elicited from the new group. Indeed, with sufficient research, it might be possible to converge upon a single, stable formula for predicting new choices based on non-choice neural responses.

Accordingly, in this section we explore the feasibility of developing a single model for predicting choices from non-choice neural responses that is portable from one individual (or group) to another, rather than applicable to a single individual or group. Subsection V.A concerns the accuracy of cross-individual predictions. Because subject-level prediction is especially challenging, we investigate whether a model estimated with data on one set of individuals accurately predicts the choices of another individual over the same group of items. Ideally, we would like to use such models to predict the choices of other individuals in new situations (e.g., over different items), but as we will see, even the more modest objective pushes the limits of our current methods. Subsection V.B concerns the accuracy of cross-group predictions. In that context, we examine the portability of the predictive model both across groups and across items. In other words, we investigate whether a model estimated with data on one group’s choices over a given set of items can, with reasonable accuracy, predict another group’s choices over new items.

A. Cross-subject predictions

In this subsection, we examine the accuracy of the choices predicted for a particular subject using a model estimated with data pertaining to the same items but different subjects. For the same reason as in Section IV.A, only subjects 11 through 27 were included in this analysis.

1. Statistical methods
The methods used here are generally the same as those described in Section IV.A, with the following important exceptions.

Here, we begin with a data set consisting of all pair-wise choices for all subjects. By removing one subject at a time, we create 17 training samples (each consisting of $16 \times 50 = 800$ observations) and 17 associated hold-out samples (each consisting of 50 observations). For each training sample, we then estimate the model and use it to predict the choices for the associated hold-out observations (which are not used in the estimation). We then assess the model's out-of-sample predictive performance over all 17 hold-out samples (850 predictions in all). To set the LASSO penalty parameter $\lambda^*$, we use 16-fold cross-validation where each fold consists of all observations associated with a single subject, rather than 5-fold cross-validation as in the previous section. This alternative method of cross-validation is appropriate here because our objective is to maximize predictive accuracy when the model is estimated with one group of subjects and used to predict the choices of another subject.

Given the large number of available observations, we were also able to estimate the logistic choice model without a LASSO penalty for those initial voxel selection criteria that left us with adequate degrees of freedom. The results do not differ qualitatively from the LASSO-penalized results, which we report here for consistency with Section IV.A.

2. Results and discussion

Figure 8 plots mean success rates as a function of the fraction of voxels retained after initial screening, with the retained fractions ranging from 0.0001 (0.01% of voxels), to 1 (all voxels). Our procedure maximizes predictive accuracy when 0.05% of voxels are retained, and is only slightly lower when 10% of voxels are retained. In both cases the mean success rate exceeds 57%, which represents an economically and statistically significant improvement over the uninformed 50% benchmark ($p < .0005$, one-sided t-tests). A number of other screening criteria (including all retention rates of at least 5%) also yield statistically significant improvements over the benchmark, but in some cases those improvements are considerably smaller, and for some intermediate screening criteria the procedure performs no better than chance. Thus, the level of performance for between-subject predictions is much less robust with respect to the initial screening criterion than that of within-subject predictions (Section IV.A).
Further analysis reveals that our prediction methods have limited validity in the context of cross-subject exercises. To underscore that point, we will focus on the relatively favorable results obtained using the 10% screening criterion, acknowledging explicitly that this case seems to be a bit of an outlier, and possibly the result of chance. (We avoid using the 0.05% screening criterion, even though it achieves a slightly higher success rate, because in that case virtually all variable selection results from initial screening based on t-tests, rather than the more principled LASSO penalty criterion). We return to the role of the voxel selection criterion at the end of this section.

The first data column in Table 2 provides results on predictive accuracy for each subject (numbered 11 through 27 because this analysis excludes the first ten subjects). Once again, there was considerable cross-subject variation in success rates, which ranged from a low of 48% to a high of 72%, with all but one exceeding 50%. In this case, predictive accuracy exceeded the uninformed benchmark by a statistically significant margin for 7 out of 17 subjects at the 10% level (amongst whom the overall success rate was 63%), for 5 subjects at the 5% level, and for one subject at the 1% level. For subsequent reference, we have once again shaded all of the rows in the table associated with high-success-rate subjects (i.e., those whose success rates exceeded the uninformed benchmark by statistically significant margins), so that their results are easily distinguished from those of low-success-rate subjects (i.e., the complementary set).

One notable feature of the results in Tables 1 and 2 is that six of the nine high-success-rate subjects in the within-subject analysis (Table 1) were also high-success-rate subjects in the between-subject analysis (Table 2), and seven of the eight low-success-rate subjects in the within-subject analysis were also low-success-rate subjects in the between-subject analysis. This pattern is striking because the models used to predict choices for each subject are completely different in the two instances. The finding is certainly consistent with the hypothesis that our procedure for measuring pertinent neural responses works for some subjects but not for others.

As in Section IV, we pursue that hypothesis further by comparing success rates with mean predicted probabilities separately for high-success-rate and low-success-rate subjects. The fourth column of Table 2 contains the p-values for those subject-specific tests. There are qualitative similarities between the patterns of p-values in Tables 1 and 2, but critical quantitative differences. Qualitatively, the hypothesis of equality between the
success rate and mean predicted probability is rejected more emphatically for the low-
success-rate subjects (visually, more asterisks appear in the fourth column of the unshaded
rows than in the shaded rows). However, that hypothesis is rejected with 90% confidence
for five of the seven high-success-rate subjects (and nearly so for a sixth), and with 95%
confidence for two. These discrepancies provide an indication that the between-subject
models perform rather less well out of sample than the within-subject models.

Overall, for high-success-rate subjects, the mean success rate is 63.1%, while the
mean predicted probability is 75.8%; the difference (or bias), 12.7 percentage points, is
highly significant statistically (p <0.001) and large economically. Notably, this discrepancy
is a considerably larger than the analogous gap in Section IV. Once again, our predictions are
equally strong for the low-success-rate subjects: the mean predicted probability is 77.9%.
However, the mean success rate is only 53.2%, and the difference (or bias), 24.7 percentage
points, is large and highly statistically significant (p < 0.001).

As in Section IV, we implemented a stronger test of predictive validity based on
variation in the strength of predictions across observations. According to Table 2, the mean
within-subject standard deviation of the predicted probability is 0.125, which indicates
considerable variation. Graphs of the relationship between success rates and predicted
probabilities, analogous to Figure 4B (but omitted for the sake of brevity) show only a weak
relationship for high-success-rate subjects (primarily at high values of predicted
probabilities), and once again no relationship at all for low-success-rate subjects. The same
point is summarized more succinctly by estimates of linear probability models relating a
successful prediction indicator to the predicted probability of the more likely choice. The
resulting slope coefficient is 0.255 (s.e. = 0.176) with no fixed effects, and 0.304 (s.e. =
0.182) when subject-fixed effects are included. While these estimates provide some hint of
a relationship, particularly in comparison to the results for the low-success-rate subjects
(slope = 0.031, s.e. = 0.147), they are far weaker than those obtained for within-subject
predictions.

With a voxel screening criterion of 0.05%, we achieve greater success in some
dimensions: the differences between success rates and predicted probabilities for high-
success-rate subjects are much lower, and the slope coefficients in the LPMs are a bit larger
and statistically significant (e.g., slope = 0.442 and s.e. = 0.208 for a model with subject-fixed
Arguably, selecting the voxel screening condition based on high overall success rates does not automatically create a systematic bias toward the latter pattern. However, given the obvious sensitivity of our results to the screening criterion, our analysis does little more than hint at the potential feasibility of reliable cross-subject predictions.

If one can conclude from the within-subject analysis that our measurement procedure works poorly some of our subjects, then it may be reasonable to exclude those subjects prior to attempting cross-subject predictions. Unfortunately, a cross-subject prediction analysis that is confined to the nine high-success-rate subjects from Section IV.A does not perform appreciable better than the analysis described above.

The fact that it is harder to predict choices across subjects than within subjects is not surprising, for several reasons. First, differences in brain size and shape make it difficult to align brain regions spatially across individuals. As a result, the same voxel may perform a different function in different subjects, which can introduce substantial cross-subject noise. Second, our method does not account for the fact that the range of functional neural responses varies across subjects. In other words, the difference in neural responses between the most and least liked foods may be much larger for some subjects than for others. This consideration introduces additional noise, in that our procedure treats the BOLD signals of all subjects identically. Third, different subjects may engage in different cognitive activities during the passive viewing stage, and these activities may be related to underlying preferences in different ways. In that case, the best predictive model for each individual might be highly idiosyncratic.

That said, one should view our findings as a first step toward the development of methods for cross-subject choice prediction. Prediction rate may improve as technology evolves. For example, more sophisticated procedures for aligning brains may reduce the noise introduced when moving from one brain to another (see, e.g., Thirion et al., 2006). Other techniques, such as normalizing neural responses voxel-by-voxel for each subject may yield gains in predictive power.

B. Cross-group predictions

---

23 We also achieve less success in another dimension: the sets of high-success-rate subjects for the within- and between-subject analyses overlap to a much smaller degree.
In this subsection, we examine the accuracy of the average ratings predicted for a
group of subjects using a model estimated with data pertaining to the different items and a
different group of subjects. As in Section IV.B, all twenty-seven subjects were included in
this analysis.

1. Statistical methods

The methods used here are identical to those of Section VI.B, with some exceptions
involving the nature of the training and hold-out samples. In the current section, we
randomly divide the subjects into a training group of 14 subjects and a hold-out group of 13
subjects. By removing two items at a time from the set of 100, we create 50 training sets
(each consisting of 98 items) and 50 associated hold-out sets (each consisting of two items).

For each set of training items, we then estimate the same models as in Section IV.B
using data on the training subjects – one model to predict whether an item’s average rating
is above the median, and another to predict the value of its average rating. We use the first
model to predict whether the average ratings of the hold-out subjects for the hold-out items
will fall above or below the average rating of the median item for the hold-out subjects, and
the second to predict the average ratings themselves.

To ensure that our results cannot be attributed to a potentially idiosyncratic
division of the subjects, we repeat this exercise 200 times, selecting the training and hold-
out groups randomly each time. In this way, we generate a total of 20,000 predictions (100
for each population draw).

2. Results and discussion

We begin with an analysis of predictions concerning the probability that the hold-
out group’s average subjective rating for a given hold-out item will fall above the median
rating for items in the training data. Figure 8 plots the overall success rate (averaged over
the 200 population draws) as a function of the fraction of voxels retained after initial
screening, with the retained fractions ranging from 0.0001 (0.01% of voxels), to 1 (all
voxels). Our procedure maximizes predictive accuracy when 50% of voxels are retained.
The average overall success rate is then 61.3%, which represents an economically and

24 This figure represents the overall success rate averaged over the 200 population draws.
statistically significant improvement over the uninformed 50% benchmark ($p < 0.001$, one-sided t-test). Here, the initial voxel selection criterion has a fairly small effect on the success rate. To avoid cherry-picking results section-by-section, we will adopt the same screening criterion here as in Section IV (1%), which yields an average overall success rate of 60.3%, rather than the success-rate-maximizing 50% criterion. Our conclusions are not substantially affected by applying less restrictive screens.

As in Section IV, we perform an initial check on the validity of the model’s predictive probability statements by comparing the strength of the typical prediction with the average overall success rate. On average, the procedure predicts that items will fall into the more likely half of the rating distribution with 79.7% probability. That figure is not close to the average overall success rate of 60.3%, and the gap is statistically significant ($p < 0.001$, two-sided t-test). Consequently, the procedure does not generate quantitatively accurate probability statements for the hold-out data.

For a more revealing assessment of the model’s predictive validity, we grouped individual predictions into deciles (2000 predictions in each) based on the predicted probability that the hold-out item’s average rating among the hold-out group would exceed the median, and then, for each decile, computed the frequency with which the hold-out group’s average ratings of those items actually fell above the median. Results appear in Figure 9. A strong positive relationship between predicted probabilities and realized frequencies is readily apparent. The relationship does not, however, lie close to the 45 degree line.

To sharpen these impressions, we estimated linear probability models relating a binary variable indicating whether the hold-out group’s average rating of a hold-out item was above the median, to the predicted probability of that event. Pooling all 20,000 predictions, the estimated intercept is 0.317 (s.e. = 0.006), and the slope is 0.365 (s.e. = 0.010). Adding fixed effects for each of the 200 population draws, the slope increases slightly to 0.366 (s.e. = 0.010). We also estimated a separate LPM for each population draw. Figure 10 shows the distribution of the resulting slope coefficients. The mean slope is 0.358 (s.e. = 0.158), and the median coincides with the mean. Although these estimates do not support a literal interpretation of the model’s predictive probability statements, they are directionally accurate. Thus, there is clear evidence that the predicted probabilities contain a good deal of information that is useful for forecasting ratings.
VI. Concluding remarks

The preceding analysis points to the feasibility of inferring the choices people would make (if given the opportunity) at least in part based on their neural responses to the pertinent prospects when they are not engaged in actual decision making. It represents a first step toward developing methods for estimating choice mappings that could be used in settings where pertinent choice data are non-existent, limited, or contaminated by spurious factors, so that more conventional methods of estimation are inapplicable or problematic. Possible examples include inferring willingness-to-pay for the avoidance of environmental damage, and the estimation of the behavioral impact of interventions where naturally occurring interventions are insufficiently clean to permit reliable inferences (e.g., concerning the impact of tax-favored retirement account on saving).

The accuracy of the predictive methods developed here will no doubt improve with further research. Better methods can be developed to enhance attentiveness in the scanner and to weed out inattentive subjects. Advancements in knowledge of the brain and improved statistical methods may provide better guides to voxel selection. Technological advances will undoubtedly enhance our ability to detect and measure stimulus-specific neural responses.

Perhaps the greatest potential for improving predictive accuracy lies in exploring combinations of diverse non-choice responses to potential prospects. One promising avenue is to supplement fMRI information with subjective non-choice responses, such as hypothetical choices, as well as other neurometric data, such as pupil dilation, facial temperature, SCRs, and the like. The latter types of measurements are easier and less costly to obtain than fMRI data, and may ultimately turn out to be equally predictive. Physiological responses may prove particularly valuable in detecting discrepancies between answers to hypothetical questions and true tendencies.
Table 1. Predictive accuracy for choices involving new items, within subject.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Success rate</th>
<th>Mean</th>
<th>Std dev</th>
<th>p-value for bias</th>
<th>Slope</th>
<th>Std dev of slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>0.66***</td>
<td>0.663</td>
<td>0.138</td>
<td>0.966</td>
<td>0.840</td>
<td>0.485</td>
</tr>
<tr>
<td>12</td>
<td>0.52</td>
<td>0.702</td>
<td>0.156</td>
<td>0.013**</td>
<td>0.563</td>
<td>0.459</td>
</tr>
<tr>
<td>13</td>
<td>0.62**</td>
<td>0.728</td>
<td>0.143</td>
<td>0.090*</td>
<td>1.632</td>
<td>0.434</td>
</tr>
<tr>
<td>14</td>
<td>0.66***</td>
<td>0.768</td>
<td>0.133</td>
<td>0.126</td>
<td>0.144</td>
<td>0.517</td>
</tr>
<tr>
<td>15</td>
<td>0.58</td>
<td>0.711</td>
<td>0.128</td>
<td>0.074*</td>
<td>0.222</td>
<td>0.559</td>
</tr>
<tr>
<td>16</td>
<td>0.52</td>
<td>0.742</td>
<td>0.147</td>
<td>0.005***</td>
<td>-0.075</td>
<td>0.494</td>
</tr>
<tr>
<td>17</td>
<td>0.58</td>
<td>0.730</td>
<td>0.136</td>
<td>0.050*</td>
<td>-0.364</td>
<td>0.527</td>
</tr>
<tr>
<td>18</td>
<td>0.76***</td>
<td>0.738</td>
<td>0.141</td>
<td>0.726</td>
<td>0.394</td>
<td>0.437</td>
</tr>
<tr>
<td>19</td>
<td>0.54</td>
<td>0.733</td>
<td>0.142</td>
<td>0.014**</td>
<td>-0.345</td>
<td>0.508</td>
</tr>
<tr>
<td>20</td>
<td>0.58</td>
<td>0.692</td>
<td>0.148</td>
<td>0.114</td>
<td>0.649</td>
<td>0.476</td>
</tr>
<tr>
<td>21</td>
<td>0.62**</td>
<td>0.748</td>
<td>0.140</td>
<td>0.070*</td>
<td>0.500</td>
<td>0.499</td>
</tr>
<tr>
<td>22</td>
<td>0.72***</td>
<td>0.697</td>
<td>0.130</td>
<td>0.727</td>
<td>0.345</td>
<td>0.501</td>
</tr>
<tr>
<td>23</td>
<td>0.68***</td>
<td>0.717</td>
<td>0.144</td>
<td>0.567</td>
<td>0.936</td>
<td>0.454</td>
</tr>
<tr>
<td>24</td>
<td>0.70***</td>
<td>0.763</td>
<td>0.118</td>
<td>0.320</td>
<td>1.159</td>
<td>0.543</td>
</tr>
<tr>
<td>25</td>
<td>0.44</td>
<td>0.760</td>
<td>0.145</td>
<td>0.000***</td>
<td>-0.416</td>
<td>0.494</td>
</tr>
<tr>
<td>26</td>
<td>0.52</td>
<td>0.733</td>
<td>0.141</td>
<td>0.007***</td>
<td>-0.312</td>
<td>0.515</td>
</tr>
<tr>
<td>27</td>
<td>0.72***</td>
<td>0.738</td>
<td>0.155</td>
<td>0.761</td>
<td>1.110</td>
<td>0.392</td>
</tr>
</tbody>
</table>

Group Mean | 0.613*** | 0.727 | 0.140 | < 0.001*** | 0.411 | 0.488 |
Group Std Dev | 0.089 | 0.028 | 0.010 |

NOTES: Based on an initial voxel selection threshold of 0.01. Asterisks are used to denote statistical significance only in the columns for “success rate” (difference from uninformed benchmark, binomial test for individual rates, 1-sided t-test for group mean rate) and “p-value for bias” (difference success rate and mean predicted probability, two-sided t-test), as follows: * denotes p<0.1; ** denotes p<0.05; *** denotes p<0.01. “Success rate” is the frequency with which the item with highest predicted choice probability in each pair was actually chosen; “p-value for bias” refers to the test statistic for the hypothesis that the success rate equals the mean predicted probability, and “LPM” refers to a simple linear probability model relating a success indicator to the predicted probability. “Group Mean” is the mean of within-subject means, and “Std Dev” is the standard deviation of within-subject means.
Table 2. Predictive accuracy for choices involving no new items, across subjects.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Success rate</th>
<th>Mean of predicted probability</th>
<th>Std dev of predicted probability</th>
<th>p-value for bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>0.60*</td>
<td>0.824</td>
<td>0.170</td>
<td>0.003***</td>
</tr>
<tr>
<td>12</td>
<td>0.62**</td>
<td>0.762</td>
<td>0.136</td>
<td>0.051*</td>
</tr>
<tr>
<td>13</td>
<td>0.62**</td>
<td>0.764</td>
<td>0.147</td>
<td>0.049**</td>
</tr>
<tr>
<td>14</td>
<td>0.64**</td>
<td>0.779</td>
<td>0.149</td>
<td>0.056*</td>
</tr>
<tr>
<td>15</td>
<td>0.56</td>
<td>0.740</td>
<td>0.146</td>
<td>&lt;0.001**</td>
</tr>
<tr>
<td>16</td>
<td>0.54</td>
<td>0.858</td>
<td>0.128</td>
<td>0.005***</td>
</tr>
<tr>
<td>17</td>
<td>0.54</td>
<td>0.755</td>
<td>0.145</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>18</td>
<td>0.62**</td>
<td>0.752</td>
<td>0.149</td>
<td>0.071*</td>
</tr>
<tr>
<td>19</td>
<td>0.58</td>
<td>0.805</td>
<td>0.162</td>
<td>0.003***</td>
</tr>
<tr>
<td>20</td>
<td>0.54</td>
<td>0.779</td>
<td>0.135</td>
<td>0.002***</td>
</tr>
<tr>
<td>21</td>
<td>0.60*</td>
<td>0.715</td>
<td>0.139</td>
<td>0.118</td>
</tr>
<tr>
<td>22</td>
<td>0.48</td>
<td>0.836</td>
<td>0.143</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>23</td>
<td>0.50</td>
<td>0.774</td>
<td>0.147</td>
<td>&lt;0.001***</td>
</tr>
<tr>
<td>24</td>
<td>0.52</td>
<td>0.778</td>
<td>0.152</td>
<td>0.001***</td>
</tr>
<tr>
<td>25</td>
<td>0.56</td>
<td>0.727</td>
<td>0.165</td>
<td>0.027**</td>
</tr>
<tr>
<td>26</td>
<td>0.50</td>
<td>0.737</td>
<td>0.155</td>
<td>0.002***</td>
</tr>
<tr>
<td>27</td>
<td>0.72***</td>
<td>0.712</td>
<td>0.109</td>
<td>0.903</td>
</tr>
</tbody>
</table>

Group Mean 0.573*** 0.770 0.150 < 0.001***
Std Dev 0.059 0.040 0.014

NOTES: Based on an initial voxel selection threshold of 0.10. Asterisks are used to denote statistical significance only in the columns for “success rate” (difference from uninformed benchmark, binomial test for individual rates, 1-sided t-test for group mean rate) and “p-value for bias” (difference success rate and mean predicted probability, two-sided t-test), as follows: * denotes p<0.1; ** denotes p<0.05; *** denotes p<0.01. “Success rate” is the frequency with which the item with highest predicted choice probability in each pair was actually chosen, and “p-value for bias” refers to the test statistic for the hypothesis that the success rate equals the mean predicted probability. “Group Mean” is the mean of within-subject means, and “Std Dev” is the standard deviation of within-subject means.
Figure 1. Samples of the food pictures used in the experiment.
Figure 2. Overall success rate as a function of the fraction of voxels retained after initial screening when predicting choices for new items either within subject or within group.
Figure 3: Distribution of prediction strength (predicted probability of the more likely alternative), within-subject predictions for new items.
Figure 4. Success rate for within-subject predictions of choices involving new items as a function of predictive choice probability for (A) the entire group, and (B) separately for high-success-rate and low-success-rate subjects.
Figure 5. Distribution of mean normalized ratings across food items.
Fig. 6. Predicting above-and-below-median ratings for new items within groups. (A) Scatter plot of mean ratings versus predicted probability that item is in the upper half of the group's valuation distribution. Circles denote correct predictions. Crosses denote incorrect predictions. (B) Fraction of items with ratings exceeded the median versus average predicted probability of rating exceeding the median, grouped by quintiles of the latter.
Figure 7. Predicting average ratings for new items within groups. Scatter plot of actual vs. predicted mean normalized ratings for each item. Each point represents a different food item. Least-squares regression line included.
Figure 8. Overall success rate as a function of the fraction of voxels retained after initial screening when predicting a new subjects' choices among the same items, and when predicting average ratings for new items among new groups.
Figure 9. Predicting above-and-below-median average ratings for new items and new groups. Fraction of items with ratings exceeded the median versus average predicted probability of rating exceeding the median, grouped by deciles of the latter.
Figure 10. Distribution over 200 population draws of slope coefficients from linear probability models relating an above-median indicator to the predicted probability of an above-median average rating, based on predictions for new items and new groups.
Appendix: Experimental Instructions

This experiment has three parts. We will give you the instructions for each part just before you perform that part.

PART I. During this part, which will be performed inside the fMRI scanner, we will show you pictures of different food items, one at a time. Each item will appear for 2.75 seconds. In total we will show 130 different items, some of which will be repeated several times. It is important that you pay attention to the stimulus while it is on the screen. All of the food items used in the experiment are available for purchase at local grocery stores.

Why should you care about the pictures? Because at the end of the experiment we will ask you to eat one of the snacks shown.

There is a 50% probability that we will randomly select one of the trials from Part 1 and that you will have to eat 3 spoonfuls of the food item shown that trial. If the snack is not chosen from Part 1, it will be chosen from either Part 2 or 3, which we will tell you about later. Either way, you will eat one of the foods shown in the scanner.

Occasionally you will be asked whether a snack is “salty” or “sweet.” Please indicate your response using the buttons as directed by the experimenter.

PART II. During this part you will see 50 screens with pairs of food items that you saw in Part I. In every trial we want you to select the food item that you would prefer to eat. You will do this by a mouse click on the picture that you want to choose.

There is a 50% probability that a trial will be selected from this part of the experiment. If a trial is selected from part II, you will have to eat three spoonfuls of whatever item you select from the pair of items shown in that trial.

PART III. During this part of the experiment we ask you to indicate how much you would like to eat each of the snacks shown, on a scale ranging from -3 (Strong Dislike) to 3 (Strong Like). You will indicate your ratings by clicking on the scale below the image.

Although the snack you will eat has already been chosen, we ask that you take this part of the experiment seriously and indicate your true rating for each snack.
### Appendix: List of foods used in the experiment, with ratings

<table>
<thead>
<tr>
<th>Food</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Food</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>pig feet (pickled)</td>
<td>-1.87</td>
<td>0.56</td>
<td>tomatosoup</td>
<td>0.23</td>
<td>0.60</td>
</tr>
<tr>
<td>lamb babyfood</td>
<td>-1.61</td>
<td>0.77</td>
<td>vanilla pudding</td>
<td>0.23</td>
<td>0.63</td>
</tr>
<tr>
<td>clam juice</td>
<td>-1.56</td>
<td>0.95</td>
<td>grapefruit</td>
<td>0.26</td>
<td>0.78</td>
</tr>
<tr>
<td>babybeans</td>
<td>-1.40</td>
<td>0.78</td>
<td>streusel</td>
<td>0.27</td>
<td>0.78</td>
</tr>
<tr>
<td>carrot babyfood</td>
<td>-1.35</td>
<td>0.65</td>
<td>wheatcrisps</td>
<td>0.27</td>
<td>0.59</td>
</tr>
<tr>
<td>macncheese babyfood</td>
<td>-1.33</td>
<td>0.73</td>
<td>nutrigrain blue</td>
<td>0.29</td>
<td>0.70</td>
</tr>
<tr>
<td>squash babyfood</td>
<td>-1.26</td>
<td>0.63</td>
<td>blueshortbread</td>
<td>0.30</td>
<td>0.67</td>
</tr>
<tr>
<td>babyclams</td>
<td>-1.25</td>
<td>0.76</td>
<td>tomatoes</td>
<td>0.30</td>
<td>0.86</td>
</tr>
<tr>
<td>clams</td>
<td>-1.25</td>
<td>0.72</td>
<td>caramelcrisps</td>
<td>0.31</td>
<td>0.80</td>
</tr>
<tr>
<td>spam spread</td>
<td>-1.14</td>
<td>0.88</td>
<td>chocolate pudding</td>
<td>0.32</td>
<td>0.81</td>
</tr>
<tr>
<td>hamspread</td>
<td>-1.12</td>
<td>0.80</td>
<td>toast peanut butter</td>
<td>0.34</td>
<td>0.48</td>
</tr>
<tr>
<td>sardines</td>
<td>-0.97</td>
<td>0.97</td>
<td>sour cream onion chips</td>
<td>0.35</td>
<td>1.01</td>
</tr>
<tr>
<td>cornedbeef</td>
<td>-0.91</td>
<td>0.95</td>
<td>mixed yogurt</td>
<td>0.37</td>
<td>0.58</td>
</tr>
<tr>
<td>spinach</td>
<td>-0.82</td>
<td>0.73</td>
<td>lays classic chips</td>
<td>0.38</td>
<td>0.66</td>
</tr>
<tr>
<td>plum</td>
<td>-0.82</td>
<td>0.70</td>
<td>frostedcheerios</td>
<td>0.39</td>
<td>0.66</td>
</tr>
<tr>
<td>sweetpeas</td>
<td>-0.73</td>
<td>0.71</td>
<td>oatmeal cookies</td>
<td>0.40</td>
<td>0.50</td>
</tr>
<tr>
<td>garbanzo beans</td>
<td>-0.72</td>
<td>0.81</td>
<td>popcorn buttered</td>
<td>0.40</td>
<td>0.71</td>
</tr>
<tr>
<td>palm heart</td>
<td>-0.71</td>
<td>0.93</td>
<td>starburst</td>
<td>0.41</td>
<td>0.63</td>
</tr>
<tr>
<td>sauerkraut</td>
<td>-0.69</td>
<td>0.94</td>
<td>cashews</td>
<td>0.42</td>
<td>0.90</td>
</tr>
<tr>
<td>artichog</td>
<td>-0.66</td>
<td>1.20</td>
<td>kiwi</td>
<td>0.43</td>
<td>0.93</td>
</tr>
<tr>
<td>beet</td>
<td>-0.58</td>
<td>0.90</td>
<td>cheezits</td>
<td>0.44</td>
<td>0.93</td>
</tr>
<tr>
<td>water chestnuts</td>
<td>-0.58</td>
<td>0.87</td>
<td>doritos ranch</td>
<td>0.45</td>
<td>0.59</td>
</tr>
<tr>
<td>albacore</td>
<td>-0.51</td>
<td>0.99</td>
<td>fruitsnacks</td>
<td>0.45</td>
<td>0.74</td>
</tr>
<tr>
<td>brusselsprouts</td>
<td>-0.48</td>
<td>1.00</td>
<td>ruffles</td>
<td>0.46</td>
<td>0.73</td>
</tr>
<tr>
<td>sprouts</td>
<td>-0.46</td>
<td>0.94</td>
<td>pomegranate smoothie</td>
<td>0.47</td>
<td>0.69</td>
</tr>
<tr>
<td>salmon</td>
<td>-0.46</td>
<td>0.71</td>
<td>cherry ice cream</td>
<td>0.48</td>
<td>0.89</td>
</tr>
<tr>
<td>slim jim</td>
<td>-0.38</td>
<td>0.78</td>
<td>teddy grahams plain</td>
<td>0.48</td>
<td>0.54</td>
</tr>
<tr>
<td>SlimFast</td>
<td>-0.37</td>
<td>0.87</td>
<td>plums</td>
<td>0.49</td>
<td>0.81</td>
</tr>
<tr>
<td>celery</td>
<td>-0.34</td>
<td>0.88</td>
<td>wheat thins</td>
<td>0.52</td>
<td>0.50</td>
</tr>
<tr>
<td>babycorn</td>
<td>-0.29</td>
<td>0.80</td>
<td>pringles</td>
<td>0.52</td>
<td>0.66</td>
</tr>
<tr>
<td>clam chowder</td>
<td>-0.24</td>
<td>0.82</td>
<td>hershey kisses</td>
<td>0.55</td>
<td>0.57</td>
</tr>
<tr>
<td>redvines</td>
<td>-0.14</td>
<td>0.71</td>
<td>blueberry smoothie</td>
<td>0.55</td>
<td>0.62</td>
</tr>
<tr>
<td>beef jerky</td>
<td>-0.13</td>
<td>0.96</td>
<td>grannysmith apple</td>
<td>0.57</td>
<td>0.68</td>
</tr>
<tr>
<td>chow mein</td>
<td>-0.10</td>
<td>0.91</td>
<td>chips ahoy</td>
<td>0.61</td>
<td>0.50</td>
</tr>
<tr>
<td>tuna</td>
<td>-0.10</td>
<td>0.89</td>
<td>coke</td>
<td>0.64</td>
<td>0.85</td>
</tr>
<tr>
<td>broccoli</td>
<td>-0.07</td>
<td>0.85</td>
<td>100grand</td>
<td>0.64</td>
<td>0.52</td>
</tr>
<tr>
<td>lemons</td>
<td>-0.05</td>
<td>0.63</td>
<td>orange</td>
<td>0.66</td>
<td>0.76</td>
</tr>
<tr>
<td>cherry pie</td>
<td>0.03</td>
<td>0.72</td>
<td>banana</td>
<td>0.67</td>
<td>0.75</td>
</tr>
<tr>
<td>macncheese</td>
<td>0.05</td>
<td>0.72</td>
<td>tobleron</td>
<td>0.71</td>
<td>0.74</td>
</tr>
<tr>
<td>goldfish</td>
<td>0.08</td>
<td>0.62</td>
<td>ice creams sandwich</td>
<td>0.72</td>
<td>0.76</td>
</tr>
<tr>
<td>sbc crackers</td>
<td>0.09</td>
<td>0.55</td>
<td>milky way</td>
<td>0.72</td>
<td>0.57</td>
</tr>
<tr>
<td>flaming cheetos</td>
<td>0.09</td>
<td>0.84</td>
<td>oreos</td>
<td>0.75</td>
<td>0.66</td>
</tr>
<tr>
<td>chicken noodle soup</td>
<td>0.11</td>
<td>0.82</td>
<td>lindt</td>
<td>0.75</td>
<td>0.71</td>
</tr>
<tr>
<td>cheese n crackers</td>
<td>0.11</td>
<td>0.68</td>
<td>crunch</td>
<td>0.78</td>
<td>0.55</td>
</tr>
<tr>
<td>lunchables turkey</td>
<td>0.13</td>
<td>0.81</td>
<td>red grapes</td>
<td>0.80</td>
<td>0.55</td>
</tr>
<tr>
<td>trix</td>
<td>0.13</td>
<td>0.66</td>
<td>chipmuffins</td>
<td>0.80</td>
<td>0.60</td>
</tr>
<tr>
<td>cracker jacks</td>
<td>0.14</td>
<td>0.83</td>
<td>white grapes</td>
<td>0.83</td>
<td>0.60</td>
</tr>
<tr>
<td>corn pops</td>
<td>0.15</td>
<td>0.60</td>
<td>ghiradelli milk</td>
<td>0.86</td>
<td>0.52</td>
</tr>
<tr>
<td>sun chips french onion</td>
<td>0.19</td>
<td>0.60</td>
<td>kitkat</td>
<td>0.93</td>
<td>0.57</td>
</tr>
<tr>
<td>fritos</td>
<td>0.20</td>
<td>0.65</td>
<td>ecologist milk</td>
<td>0.96</td>
<td>0.44</td>
</tr>
</tbody>
</table>
References


