Combining Response Times and Choice Data Using A Neuroeconomic Model of the Decision Process Improves Out-of-Sample Predictions*

John A. Clithero†
Antonio Rangel

November 2013

Abstract

A basic problem in empirical economics involves using data from one domain to make out-of-sample predictions for a different, but related environment. When the choice data are binary, a canonical method for making these types of predictions is the logistic choice model. We investigate if it is possible to improve out-of-sample predictions by changing two aspects of the canonical approach: 1) Using response times in addition to the choice data, and 2) Combining them using a model from the neuroeconomics literature, called the Drift-Diffusion Model (DDM). We compare the out-of-sample prediction accuracies of both methods using both simulations and real experimental data. In both cases we find that the DDM method outperforms the logistic prediction method. Furthermore, although the improvement in prediction accuracy is small for the case in which items have very similar or very different values, the DDM method improves prediction accuracy substantially for intermediate cases.

JEL Classification: C9, D03, D87

Keywords: drift diffusion, neuroeconomics, stochastic choice, prediction

*The research was financed with financial support from NSF (AR3.SELFCNTRL-1-NSF.ARR1), the Lipper Foundation, and the Betty and Gordon Moore Foundation. Very useful comments from Doug Bernheim and Mike Woodford are gratefully acknowledged.

†JAC and AR are both at the California Institute of Technology, Pasadena, California. Email: clithero@hss.caltech.edu (JAC).
1 Introduction

Suppose that you are the owner of a restaurant that every day offers a very simple menu: patrons can choose between one of two items at a fixed and unchanging price. Furthermore, suppose that your kitchen can only prepare a fixed and relatively small number of dishes, so that there is also a relatively small number of possible menus. In order to maximize profits, you would like to be able to forecast the relative demand among your patrons for different menu combinations, which among other things, would allow you to avoid wasting food. You are fairly certain that the choices that they made in the past can be used to forecast their choices among new menu combinations. However, you are not sure about how to do this. Overwhelmed by the problem, you decide to consult your two siblings: the pride of the family, who is a professor of empirical economics, and the black sheep, who is a neuroeconomist.

Our hypothetical restaurateur is facing a basic empirical problem, in which observed choice data are used to fit a model of a patron’s decision problem in order to make out-of-sample predictions for new menu combinations. In fact, a simple but powerful way of approaching this problem is known to any doctoral student in economics, and is likely to be the answer provided by the economist sibling [78]. Let \( n = 1, \ldots, N \) denote the possible meals that can be prepared by the restaurant’s kitchen. For every day \( t \) in which choice data are available, let \( L(t) \) denote the item listed on the left side of the menu, and \( R(t) \) denote the one listed on the right side. For every frequent patron, estimate the following basic logistic regression,

\[
Pr(choice = Left) = \text{Logit} \left( \beta_0 + \sum_n \beta_n (I(L(t) = n) - I(R(t) = n)) \right)
\]

where \( I(L(t) = n) \) and \( I(R(t) = n) \) are indicator functions specifying whether item \( n \) was presented on the left or the right side of the menu.\(^1\) In this simple model, \( \beta_0 \) captures any biases for the left item, and \( \beta_n \) provides a measure of the average utility of consuming item \( n \). Once a parameter \( \hat{\beta}_n \) has been estimated for each possible meal, the estimated probability of choosing any item \( n \), from a menu consisting of dishes \( n \) and \( m \), is given by

\[
Pr(choice = n | menu = n, m) = \frac{e^{\hat{\beta}_n}}{e^{\hat{\beta}_n} + e^{\hat{\beta}_m}}.
\]

Of course, more sophisticated models are possible (e.g., involving interactions between spa-

\(^1\)The \text{Logit} function is the canonical \( \text{Logit}(p) = \log \left( \frac{p}{1-p} \right) \), for \( p \in (0,1) \).
tial location in the menu and underlying meal preferences, or choice set effects). However, in many applications, this simple model performs quite well in predicting out-of-sample choices, which is why it is a workhorse of empirical economics. In fact, as we will see, it does quite well in our dataset.

The goal of this paper is to explore the answer that the neuroeconomist sibling might provide to our restaurateur. In particular, we ask if it is possible to improve out-of-sample predictions by changing two aspects of the canonical prediction exercise. The first change entails using response times in addition to the choice data. The second change entails combining them using a model from the psychology and neuroscience literature, called the Drift-Diffusion Model (DDM), which provides an algorithmic description of how the value of items maps to choice probabilities and response times.

There are two motivations for asking this question. First, there has been an increased interest in exploring the usefulness of response time data for economic analysis. Several studies have used response times as a proxy for decision-making style or effort, and have found that choices consistent with less deliberation seem to be made faster than those consistent with more deliberation [3, 69, 70, 85]. In addition, it is well known in psychology that response times are longer for more difficult tasks. For example, it takes longer to decide if 9 is larger or smaller than 10, than to carry out a similar comparison between 1 and 10 [57, 25]. Several economic studies have built on this observation to show that response times are correlated with preference parameters and choices in a variety of settings. This suggests that response time data might contain useful information in making out-of-sample predictions.

Second, the DDM has been shown to fit the choice and response time data from simple decision tasks quite well [5, 56], especially when eye fixations are taken into account [43]. Furthermore, functional magnetic resonance imaging (fMRI) data have shown that the computations described by the DDM seem to be implemented by the brain at the time of choice in a ‘comparator’ network of brain regions that transforms value signals into the motor commands necessary to implement the choices [5, 35]. In other words, there is increasing evidence that the DDM is a valid approximation to the algorithm used by the brain to make simple choices.

There is a growing RT literature with a variety of applications in economics. A few examples are Agranov et al. [1], Chabris et al. [20], Rand et al. [60], Rubinstein [69], and Schotter and Trevino [72].
Here we study the ability of the DDM to improve out-of-sample choice predictions by carrying out an experiment that resembles the restaurateur’s problem. Subjects are asked to make real choices among snack foods in two different situations. First, they make choices in a Yes-No Task (YN-Task) in which they are shown one item at a time, and have to choose whether they prefer to eat it at the end of the experiment (chosen by selecting Yes), or if they instead prefer to have a reference item (chosen by selecting No) that is kept constant across trials. Second, they make choices among all possible pairs of non-reference items. The YN-Task generates a dataset of choices and response times that can be used to predict the choices in the second task either using the canonical logistic regression approach, or by calibrating the parameters of a DDM.

We compare the out-of-sample prediction accuracies of both methods using simulations and real experimental data. In both cases we find that the DDM method that incorporates responses times outperforms the logistic prediction model. Furthermore, although the improvement in prediction accuracy is small for the case in which items have similar, or very different values, the DDM improves prediction accuracy substantially (sometimes by as much as 50%) for intermediate cases involving choices among items with values that are somewhat, but not extremely different.

The rest of the paper is organized as follows. Section 2 provides a brief description of related papers in the literature. Section 3 provides a primer on the DDM. Section 4 describes the details of the prediction exercise. Section 5 compares the properties of both prediction methods using simulations. Section 6 compares their properties using real choice data from an experiment. Section 7 carries out a robustness check designed to show that the predictive advantage of the DDM does not come solely from the use of the response times. Section 8 concludes.

2 Related Literature

Several recent papers have shown that it is possible to predict choices using a variety of non-standard, non-choice measures. Wang, Spezio and Camerer [82] show that fixation patterns and pupil dilation can be used to predict behavior and deception in sender-receiver games. Similarly, knowledge of eye fixation patterns helps to predict aspects of out-of-sample binary and trinary choice [43, 44, 45]. Finally, several recent neuroeconomic studies

3
have shown that it is possible to predict choices out-of-sample using measures of neural responses to potential choice items taken in a non-choice context with fMRI [47, 74, 79].

Building on the pioneering work of Luce [50] and McFadden [53], a recent series of papers have investigated the theoretical foundations of random choice rules [2, 28, 33], and the role of salience, attention, and imperfect information in these processes [10, 18, 52]. However, the algorithms that have been proposed are static and thus cannot make use of response times. A notable exception is Natenzon [58], which re-discovers the basic ideas behind the DDM, but does not investigate the role of response times in making choice predictions. In contrast to the DDM, the choice ‘algorithms’ proposed in this literature have not been tested empirically using behavioral or neural data.

More generally, the value of introducing neuroeconomic models and concepts into economic theory has been discussed both approvingly [16, 17, 24, 27, 29, 46, 71, 86], and skeptically [6, 34]. In particular, Bernheim [6] offers the following challenge to neuroeconomists (italics added): “Mainstream economists will relinquish their skepticism of neuroeconomics only when confronted with examples of superior out-of-sample prediction in contexts involving the types of environmental conditions and behaviors that economists ordinarily study.” The results in this paper provide a first step in responding to this important challenge.

3 A Primer on the Drift-Diffusion Model

The brain constantly needs to select among competing responses in domains as varied as perception, memory and choice. The DDM proposes an algorithmic description of how these conflicts are resolved in all of these domains. The model was originally developed by Ratcliff and collaborators in an influential series of papers [62, 63, 65], which built on the seminal work of Wald [81] and Stone [75].

The DDM has been extremely influential in psychology, neuroscience, and neuroeconomics for two reasons. First, it has been able to provide highly accurate quantitative descriptions of choices and response times in a wide range of tasks spanning domains from linguistics to perceptual discrimination [32, 38, 59, 63]. Second, a growing body of evidence from the domains of perceptual discrimination and neuroeconomics suggests that

---

3See also Townsend and Ashby [77], Luce [51], and Link [49]. A related discussion for economics also takes place in Rieskamp et al. [68] and Brocas [13].
the computations described by the model resemble those implemented by the brain. In the perceptual discrimination domain, fMRI and neurophysiology studies have found that the lateral intraparietal area implements computations consistent with the DDM \cite{12, 32, 38}. In the realm of economic decision making, fMRI studies have shown evidence for computations described by the parameters in the DDM. In addition to the robust evidence that the brain can compute values for available options \cite{29, 61}, there is also evidence at the time of choice for a ‘comparator’ network that includes dorsomedial prefrontal cortex, dorsolateral prefrontal cortex, and inferior parietal sulcus \cite{5, 30, 35, 84}.

3.1 DDM in Perceptual Discrimination

A useful way of introducing the DDM to economists is to explore it first in the context of perceptual discrimination, where the logic and motivation for the model are perhaps clearest. Figure 1 illustrates a typical task used in this literature. Subjects are shown an image that is built by taking a linear combination of a picture of a face, with weight $\lambda \in [0, 1]$, and a car, with weight $1 - \lambda$.\footnote{All of the pictures are the same size. Different cars and faces are used in different trials. Other common tasks include pictures of faces and houses \cite{37} or dot-motion discrimination \cite{32}.} The subject is asked to guess whether the dominant item in the image (in the sense of having the larger weight) is either a face or a car. Correct responses are rewarded with a gain of $1$. Incorrect responses are penalized with a loss of $1$. When $\lambda$ is close to 0.5, as illustrated in Figure 1, the task is quite hard and requires filtering out a lot of noise to find the dominant representation. In contrast, the task is quite easy when $\lambda$ is close to 0 or 1. For obvious reasons, the parameter $\lambda$ is called the stimulus coherence.

Figure 1 provides a qualitative description of the typical choice data generated by this task. It shows that a logistic curve approximates choice probabilities, according to

$$P(\text{choice} = \text{Face}) = \frac{e^{A\lambda}}{e^{A\lambda} + e^{A(1-\lambda)}},$$

(1)

for some constant $A > 0$, with the choice curve crossing 50% when $\lambda = 0.5$. In addition, average responses times, when conditioned on $\lambda$, show an inverted U-shape: response times are largest when the stimulus has minimal coherence, at $\lambda = 0.5$, and then decrease as the coherence of the stimulus increases, either towards 0 or 1. The intuition is straightforward: response time increases with task difficulty, so that they are smallest as $\lambda$ approaches 0 or 1, and largest when it is close to 0.5.
Figure 1: Overview of the DDM in perceptual decisions. A) Subjects are shown blurred images of faces and cars, created through a linear combination that gives a weight $\lambda$ to the car image and a weight $1 - \lambda$ to the face image. B) Sample trial from a typical perceptual decision-making experiment. Individuals are shown a blurred image, with unknown $\lambda$, and have to decide if the Face ($F$ response) or the Car ($C$ response) image is dominant. C) Typical psychometric choice curve. D) Typical mean response time (RT) curve. E) Graphical summary of the DDM, and sample paths for a case with $\mu > 0$, in blue, and for the case $\mu < 0$, in orange. See text for more details.

Figure 1E illustrates the DDM algorithm. Let $t$ denote the time elapsed from the beginning of the trial. At any instant $t$ during the decision process, the brain is assumed to compute a relative value signal, $RVS(t)$, which measures the current estimate of the value of Face minus the value of Car. The variable can have negative values, which denote instances in which the Car is perceived as more likely to be the dominant image. The $RVS$ signal evolves until it crosses a pre-established barrier either at $+B > 0$, in which case Face is chosen, or at $-B < 0$, in which case Car is chosen. More precisely, the signal evolves in
small time increments, $\Delta t$, according to the following stochastic difference equation

$$RVS(t + \Delta t) = RVS(t) + \mu(\lambda) + \epsilon_t,$$

(2)

where $\mu(\lambda)$ is a drift-rate that depends linearly on the stimulus coherence, and $\epsilon_t$ is i.i.d. Gaussian noise with zero mean and $\sigma^2$ variance. The process begins at $RVS(0) = 0$, so there is no initial bias towards Face or Car. The model also assumes that the observed response times include a period of non-decision time ($NDT$) during which no response values are computed or compared. This period is normally thought to involve basic perceptual processing, such as recognizing that a trial has been initiated and determining the location of the stimuli. Its length is assumed to vary across subjects, but not with the properties of the stimuli shown on each trial. Sample paths for $RVS$, one for each possible choice, are also shown in Figure 1E.

Several properties of the DDM are worth highlighting. First, there are well-known relationships between the parameters of the DDM and the parameters of the logistic choice and response time curves. In particular, suppose that drift rates are linearly related to stimulus coherence so that $\mu_t = k(0.5 - \lambda_t)$, where $k > 0$ is a constant. In this case we have that the choice probabilities in trial $t$ are given by

$$Pr(\text{choice} = \text{Face}|\lambda_t) = \frac{1}{1 + e^{-2B\mu_t/\sigma^2}}.$$  

(3)

This is a logistic response curve, as illustrated in Figure 1C. The probability of choosing Face is largest when $\lambda = 1$, smallest when $\lambda = 0$, and equal to 50% when $\lambda = 0.5$ (as this implies $\mu_t = 0$). The conditional mean response times, exclusive of $NDT$, are given by

$$E[rt|\lambda_t] = B \frac{\mu_t}{\mu^2} \tanh(B/\sigma^2\mu_t).$$

(4)

Second, the core parameters of the model are only identified up to ratios of the free parameters. In particular, multiplying $B$, $\sigma^2$, and $\mu$ by the same positive constant has no effect on the choices and response times predicted by the DDM [63, 73]). For this reason, it is customary to fix one of these three parameters, and to only estimate the other two. Following a common convention in the literature, we set $\sigma^2 = 1$ in all subsequent analyses. \footnote{The hyperbolic tangent function used here is $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. In the case of $\mu_t = 0$, the limiting result for equation 4 is $\lim_{\mu_t \to 0} E[rt|\lambda_t] = \frac{\mu^2}{\sigma^2}$.}

\footnote{This implies that $B$ and $\mu$ can be thought of as being measured in units of standard deviation of the Gaussian noise.}
Third, the DDM algorithm induces decision mistakes, since the non-dominant item is chosen with positive probability, regardless of the coherence parameter $\lambda$. The mistakes are unavoidable since, in essence, the brain faces a statistical decision problem with noise. Every instant $\Delta t$, the comparator process receives a noisy but informative signal, and has to aggregate those signals to select the most likely option. In fact, it can be shown that under appropriate assumptions about the sampling process, the DDM implements an optimal sequential likelihood ratio test using the sequence of noisy signals.\footnote{There is an overview of this process in Gold and Shadlen \cite{31}. For more details see Bogacz et al. \cite{9}.}

### 3.2 DDM in Economic Choice

The same model can be used to describe the types of simple economic choices studied here. In particular, consider the binary choice task depicted in Figure 2, in which a subject is shown a pair of foods and has to decide which one to eat, by pressing either a left or a right button. A growing body of research in neuroeconomics suggests that the brain makes these choices by assigning a value to each of the stimuli, and then comparing those values to make a choice \cite{61}.\footnote{We have also completed a meta-analysis of the fMRI literature of value-based choice \cite{23}.} A minor modification of the DDM described above provides an algorithmic description of how choices can be made in this case. In particular, suppose that the drift rate $\mu_t$ in any given trial is proportional to the value difference between the two items; i.e., $\mu_t = s(V_{Left} - V_{Right})$, where $s$ is a positive constant. This relationship is depicted in Figure 2. In this case, the $RVS$ computed by the DDM denotes the estimate of the relative value of the left and right items. As illustrated in Figure 2, the left item is chosen when the $RVS$ exceeds the upper barrier at $+B > 0$, and the right item is chosen when it becomes lower than the lower barrier at $-B < 0$.

We highlight several aspects the application of the DDM to the realm of economic choice. First, it makes quantitative predictions for how choice probabilities and response times depend on the underlying value differences, as described in equations (3) and (4). Second, since the DDM is defined only up to positive affine transformations, the predictions of the model remain unchanged if a positive affine transformation is applied to the core DDM parameters. Third, given that the DDM predicts a logistic choice function, it provides a foundation for the type of Random Utility Models (RUMs) widely used in economics [54, 78]. However, the DDM leads to a different interpretation of the RUM: the choice randomness is driven by computational noise, and not by random shocks to the utility.
Figure 2: Applying the DDM to economic choice. A) Sample trial from a typical binary choice task. The subject is shown a pair of foods and has to decide which one to eat by pressing either the left or right buttons. B) Heat map describing how the DDM slope parameter changes as a function of $V_{Left}$ and $V_{Right}$, where $s$ is a positive constant. C) Graphical summary, and sample paths of the DDM, for the case of binary economic choice. The blue denotes a sample path with $V_{Left} > V_{Right}$. The orange path denotes a sample path with $V_{Left} < V_{Right}$.

function; as a result, it leads to decision mistakes.\footnote{See Webb et al. \cite{83} for a complementary approach to providing neuroeconomic foundations to RUMs.} Fourth, the reason the brain uses a DDM in the case of economic choice is similar to the reason it uses it in perceptual discrimination: the decision-making circuitry only has access to noisy samples of stimulus properties, and the DDM filters out this noise to compute a sequential estimate of which item is best. In fact, as in perceptual discrimination, under appropriate assumptions on the sampling process, it can be shown that the $RVS$ measures the log-likelihood ratio of the hypothesis that $V_{Left} > V_{Right}$.\footnote{For details see Bogacz et al. \cite{9}.}
4 The Prediction Exercise

In this section we describe the structure of the prediction exercise that we employ to investigate if using the DDM can improve our ability to make out-of-sample predictions. This structure is used in later sections to investigate this question using simulations, and to validate the methodology using real choice data.

4.1 Datasets

The prediction exercise is based on two datasets, collected in the laboratory, involving choices made by the same group of individuals.

The first dataset consists of a Yes-No Choice Task (YN-Task). Every trial, subjects are shown a picture of one of \( N \) different snack foods on a computer monitor, and have to decide whether they prefer to eat the displayed food, or a constant reference item that does not change across trials, at the end of the experiment. They are free to take as long as they need to make their decision. A preference for the non-reference item is indicated by pressing a Yes-button and a preference for the reference item is indicated by pressing a No-button. Figure 3A depicts the set of 17 items used in the actual experiment, as well as the reference item, which is highlighted by a blue border.\(^{11}\) Each item is shown \( T \) times, and the order of presentations is randomized within individuals subjects. For every individual, this leads to a dataset with \( N \times T \) decisions, and for each of them we record the choice \( (c_t) \) and the response time \( (r_t) \).

The second dataset describes the results of a Two-Alternative Forced-Choice Task (2AFC-Task). Every trial subjects are shown a picture of two snack foods, one placed on the left side of the computer screen, and one on the right, as illustrated in Figure 3B. Subjects have to decide which of the two items they would prefer to eat at the end of the experiment, by pressing a left- or a right-button on the keyboard. They are free to take as long as they need to make their decision. Subjects are asked to make a choice between all possible pairs of non-reference stimuli shown in the YN-Task. The order in which pairs are presented, as well as their location on the screen, are fully randomized within subjects. For every individual, this leads to a dataset with \( N(N - 1)/2 \) decisions. For each of the trials we record the choice and the response time.

\(^{11}\)Subjects are shown a picture of the reference item on the monitor before the YN-Task begins.
Figure 3: Experimental design and prediction procedure. A) Foods used in the experiment. The image highlighted with the blue box was the constant reference item in the YN-Task. B) There were two main choice tasks in the experiment. First, subjects completed a YN-Task (top), where they had to choose between eating the shown item at the end of the experiment (response = Y) or eating the constant reference item (not shown, response = N). Second, subjects completed a 2AFC-Task where two items were shown in the screen, and the subjects pressed a left or right button to indicate their choice. The basic logic of the prediction exercise is shown in the dashed boxes.

Subjects were incentivized to care about every choice in two ways. First, they were asked to abstain from eating for three hours prior to the participation of the experiment. Second, since the subjects did not know which trial from each task would count, they had an incentive to treat every choice as if it were the only one that mattered. An additional benefit of this rule is that it makes the choice trials independent.

For all subjects, the YN-Task was always completed before the 2AFC-Task. Furthermore, subjects were not read instructions about the second task until they finished the first one. The data collection was structured this way because we wanted a situation that resembles the type of prediction problems commonly faced by empirical economists, in which choice data by a given population in a particular domain is used to predict how the same population would make choices in a different but related domain.
4.2 Prediction Using Logit

The prediction exercise involves constructing a model from the YN-data to predict the choices in the 2AFC-Task. In this section we describe how to carry out this prediction using a logistic regression model. We view this model as an important benchmark because it represents the canonical way in which most economists would attack the problem.

The logistic prediction exercise involves two steps. First, the choice data from the YN-Task are used to estimate the following logistic choice probability:

$$P(c_t = \text{Yes}) = \frac{1}{1 + e^{\beta_0 - \beta_i}}.$$  \hfill (5)

Given the well-known link between the logistic regression model and the RUM [53], the estimated coefficients have a nice interpretation: $\hat{\beta}_i$ denotes the fitted value (or utility) of item $i$, and $\hat{\beta}_0$ denotes the fitted ‘average’ value of the reference option.

Under the assumption that the logistic model also describes the choice process for the 2AFC-Task, the estimated coefficients can then be used to predict choices. In particular, we have that

$$P(c_t = \text{Left}) = \frac{1}{1 + e^{\hat{\beta}_R(t) - \hat{\beta}_L(t)}},$$  \hfill (6)

where $\hat{\beta}_R(t)$ refers to the value for the right item estimated from the YN data and $\hat{\beta}_L(t)$ is defined analogously for the left item.\footnote{Note that $\hat{\beta}_0$ is not part of this prediction because the expected value difference between the two items is $\left(\hat{\beta}_R(t) - \hat{\beta}_0\right) - \left(\hat{\beta}_L(t) - \hat{\beta}_0\right) = \hat{\beta}_R(t) - \hat{\beta}_L(t)$.}

The Logit prediction exercise described here ignores the response time data, in sharp contrast to the DDM prediction exercise described next. We set up the prediction exercise this way because, traditionally, economists have not used response times in empirical applications. However, in Section 7 we explore what happens when response times are added to the logistic prediction model.
4.3 Prediction Using DDM

Now we describe the prediction exercise using the DDM. As before, the analysis consists of two steps: first we estimate the model parameters using the data from the YN-Task, and then we use the estimated parameters to predict choices in the 2AFC-Task.

The DDM for the YN-Task has the following free parameters: the size of the barriers (B), a drift rate for each choice item i (denoted by \( \mu_i \), where i is the non-reference item on the screen), and the amount of non-decision-time (NDT). Recall that, without loss of generality, we can set \( \sigma^2 = 1 \) [63]. We use maximum likelihood, as outlined below, to fit these parameters for every subject separately. In particular, we first compute a likelihood function for the outcome of each trial using the grid method. In particular, for each combination of the parameters \((B, \mu_1, ..., \mu_N, NDT)\) we simulate the DDM 10,000 times, and compute the choice probabilities \((c_t = \text{Yes}, \text{No})\) and the distribution of response times (using 20 millisecond bins). The grid had the following characteristics. Possible values of the drift rate \( \mu_i \) for each item ranged from \(-0.15\) to \(0.15\) in intervals of \(0.01\). Possible barrier sizes ranged from 10 to 50, in intervals of 5. Finally, possible non-decision times (NDT) ranged from 120 to 600 ms, in intervals of 40 ms. Let \( L(c_t, brt_t|B, \mu, NDT) \) denote the likelihood of the observation in trial \( t \), where \( brt_t \) denotes the binned response time.

We can then define the following log-likelihood function for a subject’s entire data set:

\[
LL(data|B, \{\mu_i\}, NDT) = \sum_{t=1}^{NT} \log \left( L(c_t, brt_t|B, \mu_{I(t)}, NDT) \right),
\]

where \( I(t) \) denotes the non-reference item shown in trial \( t \). The estimated parameters for each subject, denoted by \((\hat{B}, \{\hat{\mu}_i\}, \hat{NDT})\), are those that maximize this log-likelihood function.

Evidence from neuroeconomics suggests that the drift rate \( \mu_i \) for each item \( i \) is proportional to the value of the non-reference item minus the value of the reference item [56]. Under the maintained assumption that the DDM also describes the choice process for the 2AFC-Task, the prediction step is then straightforward. In particular, given that the DDM predicts logistic choice curves, we have that

\[
P(c_t = \text{Left}) = \frac{1}{1 + e^{\hat{B}(\hat{\mu}_{R(t)} - \hat{\mu}_{L(t)})}},
\]

where \( \hat{\mu}_{R(t)} \) refers to the value for the right item estimated from the YN-Task, \( \hat{\mu}_{L(t)} \) is
defined analogously for left, and \( \hat{B} \) is the estimated barrier from the YN-Task.

We conclude this section by emphasizing several features of the prediction exercises. First, in both cases the prediction exercise entails using the data from the YN-Task to estimate the value of each item, and then using these estimated values to predict the choices on the 2AFC-Task using a logistic choice function. The key difference between the two exercises has to do with how the values used in the predictions are estimated. In the Logit exercise they are estimated through a logistic regression on the choice data, and the response times are not used. In contrast, the DDM exercise uses both choices and response times to estimate the free model parameters.

Second, the prediction steps for the Logit and DDM exercises are almost identical, except for the fact that the estimated slope differences are multiplied by the estimated barrier \( B \) in the DDM. Intuitively, this is necessary to deal with the fact that in the DDM the non-decision time parameters are identified only up to a positive affine transformation.

Third, in the DDM we use the response times to estimate the values, but we do not use them to make the choice predictions. The rationale for this is that we are trying to simulate a standard economic prediction exercise in which out-of-sample response times are not available, and are not a variable of interest [6].

5 Simulations

In this section, we carry out simulations of the comparative prediction exercise under the maintained hypothesis that the DDM provides a good description of the choice process. As described in Section 3, this assumption is justified by previous findings in neuroeconomics, which have shown that this model provides accurate descriptions of choices and response times in simple choice tasks. The goals of the section are to develop intuition about the relative properties of the Logit and DDM prediction methods, and to investigate if the DDM is able to consistently generate better predictions than the Logit.

5.1 Example

In order to build some intuition, consider a simulated dataset with 30 non-reference items, and assume that the true values of the items are equally separated within the range \(-3\) to
The choice for every item is made on 20 different occasions, for a total of 600 observations. Suppose also that in the YN-Task the reference item has a value of 0. The true value of the items is unknown in real applications, but it is a useful variable to track in these simulation exercises.

We begin by simulating the YN-dataset for this case, under the assumption that the choice process is described by a DDM with the following parameters: \( B = 30, NDT = 350 \) ms, \( \sigma^2 = 1 \) and \( \mu_i = V_i / B \). The resulting data (Figure 4A, top) can be summarized by plotting the frequency of Yes responses as a function of the non-reference item’s true value. As predicted by the theory, the choice probabilities are well described by a logistic choice function that crosses 0.5 for items with a zero value. The individual observations fluctuate around the best-fitting logistic curve due to sampling noise built into the simulation.

![Figure 4](image-url)

Figure 4: Psychometric properties of simulated YN-Task with 30 items. A) Mean choice frequency for each item, and best-fitting logistic curve. True values for each item are plotted below. B) Observed mean response times (RT) and standard error band, binned by underlying true value. Note that the simulations assume that \( V_{Ref} = 0 \).

Figure 4 summarizes the response time (RT) data, by plotting the mean RT as a function of the item’s true value. This RT-curve has an inverted U-shape, with a peak at a true value of zero. Intuitively, since the value of the reference item is zero, choices are hardest when the non-reference item also has a value of zero, and become easier as its value becomes larger or smaller, in a symmetric way. We chose the simulation parameters so that the RTs
of the simulation (mean = 917 ms, SD = 454 ms, min = 403 ms, max = 4142 ms) are in the same range as those previously observed in this type of task [4, 5]. Across trials, 71% of the simulated choices were made under 1000 ms, consistent with previous findings indicating that subjects are able to make quality simple choices at sub-second speeds [36, 40, 55].

We use the data from the YN-Task to estimate the value of the different options using the Logit and DDM methods. Let \( \hat{\beta}_i \) denote the value of item \( i \) estimated by the Logit method, and \( \hat{\mu}_i \) and \( \hat{B} \) denote the drift-rate and barriers estimated by the DDM. As discussed above, we can think of \( \hat{B}\hat{\mu}_i \) as the ‘normalized’ estimate of value for item \( i \) in the DDM. These values allow us to predict choices in the 2AFC-Task, by plugging them into the logistic choice functions described in equations 6 and 8.

Figure 5A compares the performance of the Logit and DDM methods, plotting the true value difference for every trial in the 2AFC-Task versus the estimated value difference (DDM = blue diamonds; Logit = red circles). If a method predicted the values perfectly, all of the observations would lie along the 45° line. For the DDM, the estimated value differences are symmetrically distributed along a narrow band around the 45° line (estimated ordinary least squares (OLS) regression coefficient: \( \hat{\gamma} = 0.998, 95\% \text{ CI} = 0.969 - 1.027, \) blue line). For the Logit, the estimated value differences exhibit more noise, and are not symmetrically distributed along the 45° line (estimated OLS regression coefficient: \( \hat{\gamma} = 1.163, 95\% \text{ CI} = 1.131 - 1.196, \) red line). This implies that a hypothesis for \( \gamma = 1 \) cannot be rejected for the DDM, but it can be rejected for the Logit. In other words, the Logit overestimates the true value differences. Note also that the relative performance of the Logit is particularly poor at intermediate value differences, a pattern that we will revisit below.

Panel B in Figure 5 compares the true probability of choosing Left (which can only be computed if the true values are known, as in this simulation exercise) versus the predicted probability of choosing Left. Each point denotes a trial, with blue diamonds indicating DDM predictions, and red circles indicating Logit predictions. Each of these probabilities was estimated by feeding the estimated value differences plotted in Figure 5A to the predictive logistic choice functions in equations 6 and 8. A regression analysis shows that the predicted choice probabilities are an unbiased estimate of the true choice probabilities in both cases (DDM \( \hat{\gamma} = 1.008, 95\% \text{ CI} = 0.986 - 1.030 \) Logit \( \hat{\gamma} = 0.998, 95\% \text{ CI} = 0.969 - 1.027 \)). However, the estimated choice probabilities are noisier for the Logit (DDM: regression \( R^2 = 0.949 \); Logit: regression \( R^2 = 0.915 \)).
A better statistic for prediction accuracy is the absolute error (AE), which is given by the absolute value of the difference between the true choice probability and the estimated choice probability: $AE = |P(c_t = Left) - \hat{P}(c_t = Left)|$. Figure 5C plots the distribution of the difference in AE across the two methods, $AE_{DDM} - AE_{Logit}$. In this example, the average difference in AE is $-0.027$ (paired $t(434) = 7.85$). This means that, on average, the DDM predictions of choice probabilities are more accurate, but only by about 2.7 points. However, Figure 5D shows that the advantage of the DDM depends on the true value difference between the items. We see that these differences can be as large as 75% (e.g., a true choice probability of 0.4 can be associated with a difference in AE of 0.3), especially when the true choice probabilities are away from 0 or 1.
5.2 General Properties of the Prediction Methods

The previous section shows, by example, that the DDM method can outperform the Logit method in some circumstances. In this section, we investigate the robustness of the example using more general simulations.

To address these questions we repeat the previous simulation exercise 1000 times, using randomly sampled parameters and values every time (with replacement). We always assume that there are 30 non-reference items, with uniformly distributed true values, and that each choice is repeated 20 times in the YN-Task. However, the set of possible item values was selected from one out of five ranges: −3 to 3, −2.4 to 2.4, −1.8 to 1.8, −1.2 to 1.2, or −0.6 to 0.6. Similarly, the barrier \( B \) had five possible values: 20, 25, 30, 35, 40. Finally, \( NDT \) was drawn from the set 200, 225, 250, 275, 300, 325, and 350 ms.

In each simulation, we draw random parameters from these sets, which are then used to simulate a YN-Task dataset with 600 trials. As before, the data are used to estimate values with the Logit (i.e., \( \hat{\beta}_i \) for each item), and to estimate the choice-relevant parameters of the DDM (i.e., \( \hat{B} \) and \( \hat{\mu}_i \) for each item). We then used these values to predict the choices in the 2AFC-Task, by plugging the estimated value differences into equations 6 and 8.

Figure 6 summarizes the results across the 1000 simulations. First, consider Panel A, which plots the mean AE in the Logit and DDM cases. Each point represents the mean AE (MAE) in a different simulation. Data along the 45° line represent simulations in which the DDM and Logit perform equally well, according to the MAE criterion. We found that in all 1000 cases, the MAE was lower for the DDM than for the Logit. Furthermore, on average the MAE was nearly twice as large in the Logit case (DDM average MAE = 0.058; Logit average MAE = 0.111; \( t(999) = 77.90 \), two-tailed paired \( t \)-test). Panel B in Figure 6 compares the noise in the estimates across the two conditions. In particular, each point denotes the standard deviation of the AE within a simulation, for both the DDM and Logit methods. Again, we find that in the vast majority of the simulations, the Logit predictions were noisier (DDM average SD AE = 0.050; Logit average SD AE = 0.073; \( t(999) = 69.28 \), two-tailed paired \( t \)-test).

Does the relative accuracy of the two methods depend on the underlying true choice probabilities? Although the true values and choice probabilities are not observable in real data, here we can condition on them to investigate how the relative accuracy depends on
Figure 6: General simulation results, based on 1000 simulations. A) Mean absolute error (MAE) of the estimated choice probabilities for the DDM (horizontal axis) and Logit (vertical axis) methods. Each dot represents one simulated dataset. B) Standard deviation (SD) of AE for choice probabilities for the DDM and Logit methods. C) Difference in MAE binned by true choice probability. D) MAE as a function of the range of true values in the choice set (standard errors are shown as the band above and below the mean).

The underlying true parameters. Figure 6C shows that the two methods perform equally well when the items are sufficiently different, so that the true choice probabilities are near 0 or 1, or when the two items have sufficiently similar values, so that the true choice probabilities are close to 0.5. However, the DDM method performs substantially better for the intermediate cases, when the item values are somewhat different, and the true choice probabilities are in the neighborhoods of 0.25 or 0.75.

How does the range of true values among the choice set items affect prediction accuracy? Panel D in Figure 6 plots the relationship between the true value range (given by the value of the best item minus the value of the worse item) and the MAE (across all
simulations) for the Logit and DDM cases, while holding other DDM parameters constant. We find a striking difference between the two methods: whereas the MAE was relatively stable for the DDM prediction method, the MAE from the Logit procedure was about 78% larger for items that are close together in true value than for items that are far apart.\textsuperscript{13}

Figure 7 provides an alternative way of comparing the two prediction methods. For every simulation we run a simple OLS of the estimated value differences for each 2AFC trial \( t, \hat{VD}_t \), on the true value differences \( VD_t \),

\[
\hat{VD}_t = \beta_{VD} VD_t + \beta_K + \epsilon_t. \tag{9}
\]

The estimated coefficients of this regression provide a measure of the average quality of the value estimates. In particular, if the constant coefficient \( \hat{\beta}_K \) is approximately 0, and the slope coefficient \( \hat{\beta}_{VD} \) is close to 1, then the value estimates are unbiased and responsive to changes in the underlying true values. Panel A plots the distribution of slope coefficients across all simulations. The estimated coefficients for the DDM predictions have significant overlap with 1 (mean \( \hat{\beta}_{VD} = 1.10, \) SD = 0.103), whereas those for the Logit do not (mean \( \hat{\beta}_{VD} = 1.71, \) SD = 0.416). Each point in Panel A (bottom left) represents the regression results for a different simulated dataset. A paired \( t \)-test revealed that the Logit regression coefficients are significantly larger on average (\( t(999) = 47.42 \)), indicating a significantly greater bias for the Logit value estimates.

Panel B in Figure 7 provides a similar analysis for the constant coefficients. We found that the DDM constant coefficients were not significantly different from 0 (mean \( \hat{\beta}_K = 0.002, \) SD = 0.070; \( t(999) = 0.829, \) \( p > 0.40 \)), whereas the Logit coefficients were (mean \( \hat{\beta}_K = 0.017, \) SD = 0.097; \( t(999) = 5.557, \) \( p < 0.0001 \)). A paired \( t \)-test revealed that the Logit constants were also significantly larger (\( t(999) = 7.156, \) \( p < 0.0001 \)). Together, these results indicate significantly less bias in the DDM value estimates for the YN-data. This is important because bias in the fitted values is transmitted to biases in the predictions.

Two important properties of prediction models are calibration and resolution. A prediction model is well-calibrated if predicted choice probabilities closely match actual choice frequencies.\textsuperscript{14} For example, in a well-calibrated model, events that are predicted to occur

\textsuperscript{13}There is no zero true value difference on the \( x \)-axis because that would imply all the items in the set had the same value.

\textsuperscript{14}See Smith et al. [74] and Bernheim et al. [7] for further discussion of calibration.
with 40% probability actually occur 40% of the time. Importantly, a high degree of calibration does not necessarily imply that a model is a good predictor of choices. To see why, consider an example model for our data that uses no information and simply assumes that the probability of choosing Left is always 50%. This model would be well-calibrated, but would have poor resolution, in the sense of not assigning large or small probabilities to choosing Left. In particular, a prediction model has high resolution when it can identify high- or low-probability events, giving predicted choice probabilities close to 0% (e.g., 5%) or 100% (e.g., 95%). An ideal prediction model is one that is perfectly calibrated and has perfect resolution, which means that it predicts every outcome correctly. This ideal is not achievable in our context since, as long as the DDM is a valid approximation of the choice process, decisions contain a non-trivial amount of stochasticity.

Figure 7C addresses the issues of calibration and resolution by comparing estimated versus actual choice probabilities for the two methods. The black line denotes the 45° locus. The red and blue lines denote the relationship between estimated and predicted choice probabilities, respectively, for the Logit and the DDM methods, averaged across all 1000 simulations. On average, both models have high resolution in the sense that the set of predicted probabilities spans the entire range of probabilities, from 0 to 1. Although, as seen in the figure, the DDM has higher resolution. In addition, a Hosmer-Lemeshow test, which is commonly employed to assess model fit and calibration [39], shows that the observed probabilities follow the estimated probabilities more closely in the DDM than in the Logit (DDM: $\chi^2(8) = 3.24$, Logit: $\chi^2(8) = 61.81$). In other words, we find that the DDM is also better calibrated.\(^{15}\)

This simulation exercise assumes $T = 20$ observations for every item in the YN-Task. In order to investigate the robustness of this assumption, we repeated the exercise with $T = 5, 10, 15$. As shown in Table 4, the qualitative results remain unchanged, as the DDM significantly outperforms the Logit in each case.

\(^{15}\)This result should not be surprising, given the results from the estimation performed using equation 9, which performed an analogous test for the values, which map directly to choice probabilities.
Figure 7: Properties of the DDM (blue) and Logit (red) prediction methods, based on 1000 simulations. A) Slope coefficients for a linear regression of underlying estimated value differences on true value differences, estimated on each simulation separately (see equation 9). Note that the unit of observation in each regression is a trial of the 2AFC-Task. B) Constant coefficients for the same regression. C) Estimated versus observed choice probabilities in the 2AFC-Task. The black line denotes the 45° line, for a perfectly calibrated model. The color bands depict the mean and standard error for each prediction method.
Together, the simulation results in this section lead to the following conclusions. First, the DDM method consistently leads to more accurate out-of-sample choice predictions than the Logit method. Second, the average improvement in prediction accuracy is about 95%, with a MAE of 0.058 in the DDM and of 0.111 in the Logit. Third, although both prediction methods are well-resolved and well-calibrated, the DDM method is better resolved and better calibrated than the Logit. Fourth, the relative advantage of the DDM is especially large at intermediate choice probabilities, when the items are somewhat different in their true values, but not excessively so.

5.3 Intuition

The previous results give rise to the following question: Why does the DDM outperform the Logit? In this section we describe the intuition behind this result.

First, consider the choice curve for item $i$ with value $V_i$ in the YN-Task (Figure 8). For the purpose of the example, suppose that $V_{\text{Ref}} = 0$. The figure highlights seven different non-reference items, labeled $A - G$. Given the shape of the logistic function, the choice probabilities are marginally changed when we move from $A$ to $B$, or from $F$ to $G$, but they change a lot when we move from $C$ to $D$, or from $D$ to $E$. The other panel illustrates the mean RT, as a function of $V_i - V_{\text{Ref}}$, for these same items.

The top plot provides a straightforward intuition for the limitations of the Logit method. In this approach, the value estimates can only depend on the choice probabilities, and thus on the choice curve. However, given the shape of the choice curve, it is relatively easy to estimate the value difference between items $C$ and $D$, or $D$ and $E$, but relatively difficult to estimate the value difference between $A$ and $B$, or $F$ and $G$, even though their true values differ by the same amount. Given this limitation, notice what happens when the estimated values are used to predict two different subsequent binary choices: a choice between $A$ and $C$, and a choice between $B$ and $C$. Without loss of generality, assume that $C$ is placed on the left of the computer screen in both choice sets. Clearly, the true probabilities for choosing Left for the $A$-$C$ pair should be larger than those for the $B$-$C$ pair, since they have a larger underlying true value difference. However, the Logit would not capture this because it cannot reliably estimate the value differences between the items represented by points $A$ and $B$. As a result, it mistakenly assigns similar probabilities of choosing Left to both choice sets. This problem is acute when comparing the predicted choice probabilities
Figure 8: Intuition for why the DDM method outperforms the Logit method. A) Psychometric choice curve for the YN-Task as a function of the relative value of each of seven different items, labeled $A - G$. B) Mean response time in the YN-Task for each of the items. In the example, $V_{Ref} = 0$.

of choice sets $A - C$ or $B - C$, but not when comparing the predictions for choice sets $A - F$ and $B - F$. In this case, the probability of choosing the $F$ item is already approaching 1, thus rendering trivial the failure of the Logit to differentiate between values for the items represented by points $A$ and $B$.

To see the reason why the DDM reliably outperforms the Logit, consider both plots in Figure 8. Since the DDM predicts a logistic choice function, it would face the same limitations as the Logit if it were fitted using only choice data. However, the DDM also provides sharp predictions for how the response times should vary with the true value differences. Importantly, in contrast to the choice curve, the differences in response times
between points $A$ and $B$, and between $F$ and $G$ are not negligible. This is very useful because it means that the model can distinguish the value of the items associated with the points $A$ and $B$, or those in $F$ and $G$, by looking at the difference in response times. In fact, the changes in response times are smaller in the central value regions, from $C$ to $D$ and $D$ to $E$, where the choice curve would be steeper. This means that the response time data are highly complementary to the choice data, providing additional granularity for estimating values in instances when the choice data alone offer a coarser estimate.

6 Experiment

The simulation exercises show that the DDM leads to better predictions than the Logit, as long as the maintained hypothesis that the DDM provides a good description of the choice process is a valid one. Of course, the DDM, or any other simple theory, is unlikely to provide a fully accurate description of the choice process. This raises the important question of whether the advantage of the DDM also holds in practice. In this section we describe the results of a prediction exercise carried out on a real choice dataset, which was designed to address this question.

6.1 Experimental Methods

We collected real choice data sets from 32 subjects (20 male, 12 female) at the Caltech Social Science Experimental Laboratory (SSEL). Subjects were asked to abstain from eating or drinking anything but water for three hours prior to the start of the experiment. Subjects were also pre-screened for not having any allergies for the foods used in the experiment and for liking snack foods, such as potato chips, and candy.

Upon arrival, the subjects performed the YN-Task with 17 non-reference items, each shown in 10 different choice trials. Thus, the task consisted of 170 trials, with the order randomized for each subject. The Yes and No buttons were fixed within subject (i.e., Yes was always on the left or always on the right), but randomized across subjects. Afterwards, subjects completed the 2AFC-Task for every possible distinct combination of the

---

\[^{16}\text{We collected data from two additional subjects that are not reported. The subjects appeared to have ignored the experimental instruction of taking the choices seriously, as they always provided the same speeded response in the YN-Task (one always responded Yes, the other always responded No). These types of artificial responses were selected as an } a \text{ priori } \text{ exclusion criterion since our prediction exercise cannot be carried out when individuals do not exhibit variation in their choices.}\]
non-reference items, for a total of 136 choice trials. Trial order and screen location were randomized within subject. Instructions for the 2AFC-Task were not read until the YN-Task was completed.

Figure 3 provides a complete description of the food stimuli, as well as a picture of the screens that the subjects saw on both tasks. All pictures were of the same size (350 × 150 pixels) and were presented over a gray background. The tasks were presented using the Psychophysics Toolbox [11]. All other details are as described in Section 4.

After completing these tasks, subjects were paid $20 and given the snack that they chose in a randomly-selected trial from each task. The random selection was implemented by asking each subject to pull a trial number from an envelope. The entire experiment lasted approximately 50 minutes.

6.2 Exploring the DDM Fit

The comparison of prediction methods is motivated by the fact that previous neuroeconomic studies have found that the DDM provides a good quantitative description of related choice tasks (see Section 3 for references). In fact, all of our simulation results assume that the choices and response times are generated by a DDM algorithm. In addition, the exercise assumes that DDM parameters, such as the barrier and slopes, remain constant for each subject across the two tasks. In this section, we carry out a qualitative test of the extent to which the experimental data from the YN-Task and the 2AFC-Task are consistent with these assumptions.

The experimental data are summarized in Figure 9. Consider the top panels, which plot the group average choice probabilities and RT for the YN-Task as a function of the (binned and normalized within subject) estimated true values (given by using $\hat{B}_i \hat{\mu}_i$). The circles denote the group averages. A basic test of the joint hypothesis that the DDM 1) describes the data generating process and 2) our estimation method does a reasonable job estimating the model parameters, is that we should be able to generate the observed data by simulating the model with the estimated parameters. To test this, we simulated 100 synthetic subjects with parameters sampled from those estimated for our real subjects (same procedure as the simulations in Section 5). The simulation results are depicted by

As outlier RT can cause estimation issues for the DDM [66], all trials with RT greater than 5 seconds were removed for both the Logit and DDM estimation. This constituted less than 5% of the YN-Task data.
the continuous curves, with the range representing the standard errors. As can be seen from the plots, this post-hoc check provides a good fit of both curves, with the standard error bars overlapping in most cases. The quality of the fit is not surprising given previous studies, as well as the fact that the parameters are selected to maximize the likelihood of this dataset.

Now consider the bottom two panels of Figure 9, where we plot the results of carrying out a similar exercise for the 2AFC-Task. Importantly, this test is more stringent than the previous one, since we use the best fitting parameters from the YN-Task, and not those optimized for the 2AFC-Task. As shown in Panel C, the DDM provides a strikingly good out-of-sample fit of the 2AFC choice curves. Panel D plots RT against estimated value differences and shows that the curve has the pattern predicted by the DDM: longer RT for trials in which the two options are closer in value. Unfortunately, we cannot use the YN-Task estimated DDM parameters to predict the RT in the binary choice task, as the perceptual requirements of the two tasks are different, so they are unlikely to have the same NDT. This is consistent with the fact that the response times are significantly longer for the 2AFC-Task than the YN-Task (paired $t(31) = 8.27$). Together, these results provide support for the assumption that the core DDM parameters (i.e., slope and barrier) remain constant, for each subject, across the two choice tasks.

6.3 Prediction Results

In this section we describe the results of carrying out our prediction exercise on the experimental data. The steps of the exercise are exactly as described in Section 4. For each of the 32 subjects, we use the data from the YN-Task to estimate the value of each item with the Logit method (denoted by $\hat{\beta}_i$), and to estimate the choice relevant parameters for the DDM ($\hat{B}$ and $\hat{\mu}_i$). We then use these values to predict the choices in the 2AFC-Task by plugging the estimated value differences into equations 6 and 8.

Figure 10A compares the MAE of the predictions generated by the Logit and the DDM. Each point describes the MAE for one subject, averaged across all trials of the 2AFC-Task. Note that since we do not know the true values or choice probabilities, the AE statistic in this exercise is slightly different from the one used in the simulations analyses. In particular, errors are now given by $1 - \hat{p}_t$ in trials in which Left is chosen, and by $0 - \hat{p}_t$ in trials in which it is not, where $\hat{p}_t$ denotes the estimated probability of choosing Left for the trial. As shown
Figure 9: Summary of experimental choice and RT data and fitted DDM parameters. A) Observed choice probabilities in the YN-Task. B) Observed response times in the YN-Task. C) Observed choice probabilities in the 2AFC-Task. D) Observed response times in the 2AFC-Task. In all cases, circles denote group means ($N = 32$), error bars denote group standard errors, and the bands denote the predictions obtained from simulating the experiment with 100 synthetic subjects with parameters sampled from the fitted DDM parameters for the real subjects. See text for more details. Summary of the fits are provided in Table 1 and summary statistics for individual data are also provided in Table 3.
in Figure 10A, for 29 out of 32 subjects the MAE is larger in the Logit predictions than in the DDM predictions. Across the group, the MAE is significantly smaller for the DDM than for the Logit (DDM mean = 0.251; Logit mean = 0.282; \( t(31) = 6.717, p < 0.0001 \), two-tailed pairwise \( t \)-test). An additional test also illustrates the enhanced predictive power of the DDM: the average number of correct predictions (using a fitted probability estimate greater than 0.5 to predict Left, and less than 0.5 to predict Right) was also greater for the DDM (DDM mean 79.5%, Logit mean 74.5%, paired \( t(31) = 4.69, p < 0.0001 \), see Table 2).

Recall that in the simulations, we found that the advantage of the DDM was most apparent at intermediate choice probabilities, near 0.25 or 0.75. Figure 10B plots the difference in MAE as a function of the choice probabilities predicted by the DDM, and shows that this pattern is also present in the experimental data. In contrast, as suggested by the intuition summarized in Figure 8, the advantage of the DDM is minimal for cases involving extreme choice probabilities, where the underlying difference in value between the two items shown in the trial is sufficiently large.

As in the simulations, we can evaluate the calibration and resolution of the two prediction models. Figure 10C plots the estimated versus actual choice probabilities for the Logit and DDM. As shown in the figure, both methods are on average well resolved, and span the entire probability range, from 0 to 1. However, as suggested by the figure, a Hosmer-Lemeshow test across the average results shows that the DDM method is better calibrated (DDM: \( \chi^2(8) = 0.827 \), Logit: \( \chi^2(8) = 3.30 \)).

Together, the results in this section show that the ability of the DDM to make more accurate predictions also holds in real choice datasets. Furthermore, as demonstrated in the simulations, although the improvement in prediction accuracy is small for the case in which items have very similar or very different values, the DDM can reliably improve prediction accuracy for intermediate cases (recall Figures 6C and 10B) involving choices among items with somewhat different values.

\(^{18}\)Looking at the H-L \( \chi^2 \) at the individual subject level, we found that 21 of the 32 subjects had a sufficiently low \( \chi^2 \) to be considered to have predictions from the DDM that are significantly well-calibrated at \( p > 0.05 \). The median \( \chi^2(8) \) was 11.05.
Figure 10: Prediction results for the experimental data. A) Mean absolute error (MAE) for the DDM versus the Logit prediction methods. Each dot represents a subject (N = 32). Error bars denote standard errors. B) Difference in MAE across the two methods versus estimated DDM choice probabilities. Bands represent standard errors across subjects. Note the similarity between the MAE plots here for experimental data and those for the simulations in Figure 6. C) Observed versus estimated choice probabilities for both Logit (red) and DDM (blue). Bands represent standard errors across subjects. The black line is the 45°, denoting a path for perfectly calibrated predictions. Summary statistics are provided in Table 2.
7 Robustness

The DDM prediction method involves two changes from the canonical Logit approach: 1) the use of both response time and choice data, and 2) combining them in the DDM framework. One natural question is whether the prediction advantage of the DDM method is due to the introduction of new data, and not so much to the use of the structure proposed by the DDM. In this section we address this question by carrying out a robustness test of the Logit approach.

In particular, let $\bar{RT}_i$ denote the mean response time for item $i$ in the YN-Task, and consider the following version of the basic logistic regression model:

$$P(c_i = Yes) = \frac{1}{1 + e^{\beta_0 - (\hat{\beta}_i + \gamma \bar{RT}_i)}}. \quad (10)$$

This model provides a useful robustness test because it allows the estimated item values, given by $\hat{\beta}_i + \gamma \bar{RT}_i$, to be conditioned on average reaction times. Note that the model uses average response times, and not trial specific ones, in order to increase its predictive power (since reaction times are subject to trial specific noise). An additional property of the model is that it nests the basic Logit used in our analyses (if $\gamma = 0$, we then have equation 5).

Figure 11A compares the performance of the extended Logit approach with the DDM, again using the AE statistic. The DDM also outperforms this version of the Logit approach: across the group, the MAE is significantly smaller for the DDM than for the Logit (mean = 0.286; $t(31) = 5.328, p < 0.0001$, two-tailed pairwise $t$-test). Figure 11B provides an intuition for this result: for every subject, the MAE of the standard Logit approach are statistically indistinguishable from those of the extended Logit approach (paired $t$-test across subjects between the two Logit models, $t(31) = 1.50, p > 0.14$).

The results in this section are important because they demonstrate that the improved predicted power of the DDM method is not due solely to the use of RT since, as shown in Figure 11, adding RT to the Logit model does not improve its predictive power. We acknowledge that there might be other functional forms for the extended Logit approach that perform better, but we conjecture that such reduced forms would have to be consistent with the qualitative relationships predicted by the DDM.
Figure 11: Robustness check of Logit including RT data. A) Mean absolute error (MAE) for the DDM versus the Logit with RT prediction methods. Each dot represents a subject. Error bars denote standard errors. Dots above the 45° line represent subjects for which MAE was greater for the Logit with RT than the DDM. B) MAE for the Logit versus the Logit with RT prediction methods. Each dot again represents a subject. Here, the 45° line represents no change between the two Logit prediction methods.

8 Discussion

We have investigated if it is possible to improve out-of-sample predictions of simple economic choices by changing two aspects of the standard approach: 1) using response times in addition to the choice data, and 2) combining them using the DDM. In both simulated and real choice data, we found that the DDM outperforms the logistic regression model. Furthermore, although the improvement in prediction accuracy is small for the case in which items have very similar, or very different values, the DDM can improve prediction accuracy by about 50% for intermediate cases.

These results make several contributions to the economics literature. First, we have proposed and validated a new method for making out-of-sample predictions in simple discrete choice settings, such as the restaurateur’s problem described in the introduction. Second, we have demonstrated that, in simple choice settings, the method improves upon canonical methods like logistic regression. Third, our results provide a modest first step in answering Bernheim’s [6] challenge to show that neuroeconomic models can uniquely help in conducting the type of out-of-sample prediction exercises studied here.
We conclude with a discussion of several limitations of our analyses, which should be the focus of future research.

First, the food choices that we study entail real economic decisions, which people make dozens of times every day. However, they are not representative of many other salient economic decisions, such as the purchasing of a home, or the construction of an investment portfolio. These choices are substantially more complex, and they are made over hours or days, not over seconds. Furthermore, there does not currently exist a neuroeconomic model of these alternative types of decisions. Nevertheless, our results provide a template for how process or algorithmic models of decision making, when combined with new sources of data, might be able to improve out-of-sample predictions in these other domains.

Second, our analyses are based on the simplest form of the DDM, in which barriers are constant, attention is assumed to have no effect, there are only two items, and model parameters are deterministic and constant across trials.\textsuperscript{19} However, it is important to emphasize that there are many open questions regarding the precise details of the choice process [8, 42, 61, 76]. This includes whether the parameters change randomly across trials [64, 63], whether the barriers are stable or collapsing [22, 26, 56], what is the exact neural implementation of this general class of sequential-sampling models [8, 14, 80], what is the role of attention in the choice process [43, 48], and what happens as the number of choice items increases [21, 67]. For the purposes of this paper, we have worked with the simplest version of the DDM, as described in Figures 1 and 2. Obviously, future improvements in our understanding of these algorithms will allow us to refine the neuroeconomic model and out-of-sample prediction method proposed here.

Third, our prediction exercises are based on combining response times and choice data using the DDM framework. However, other types of process data, such as eye movements or computer mouse movements, might also increase out-of-sample predictive power. In fact, given progress in modeling DDM parameters with neural data [5, 19, 36], it is plausible that a combination of behavioral and neural data might further refine estimates of parameters for the choice algorithms.

\textsuperscript{19}As we have discussed elsewhere, there are several papers exploring the extension of including a parameter for attention in the DDM [43, 44, 45], as well as other related cognitive models of decision making [15, 41, 68].
Finally, we have focused on the relatively simple case of making predictions across choice contexts for the same group of subjects. An even more useful application for economics would be to improve our ability to make choice predictions across choice contexts and groups. Although this is beyond the scope of the current study, there is reason to hope that the exercise might be feasible. In particular, Krajbich and Rangel [45] have shown that it is possible to predict average choice data and response times in a trinary choice situation using the DDM parameters estimated in a different group and in a different task. However, they do not carry out the necessary comparative prediction exercise, and thus the feasibility of this approach in economic applications remains an open question.
Tables

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barrier</td>
<td>33.027</td>
<td>5.519</td>
</tr>
<tr>
<td>NDT</td>
<td>359.635</td>
<td>74.252</td>
</tr>
<tr>
<td>SSE</td>
<td>0.169</td>
<td>0.068</td>
</tr>
</tbody>
</table>

Table 1: DDM parameter estimates from YN-Task and sum of squared errors (SSE) for DDM fits on the YN-Task choices ($N = 32$).

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean AE</th>
<th>SD AE</th>
<th>Mean Correct</th>
<th>SD Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>DDM</td>
<td>0.251</td>
<td>0.257</td>
<td>0.795</td>
<td>0.059</td>
</tr>
<tr>
<td>Logit</td>
<td>0.282</td>
<td>0.244</td>
<td>0.745</td>
<td>0.068</td>
</tr>
</tbody>
</table>

Table 2: Summary of mean absolute error (AE) and prediction accuracy from predictions on 2AFC-Task ($N = 32$).

<table>
<thead>
<tr>
<th>Task</th>
<th>Stat</th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>YN</td>
<td>RT</td>
<td>0.990</td>
<td>0.736</td>
<td>0.913</td>
</tr>
<tr>
<td></td>
<td>Yes Frac</td>
<td>0.542</td>
<td>0.518</td>
<td>0.201</td>
</tr>
<tr>
<td>2AFC</td>
<td>RT</td>
<td>1.384</td>
<td>1.020</td>
<td>1.090</td>
</tr>
<tr>
<td></td>
<td>Left Frac</td>
<td>0.506</td>
<td>0.511</td>
<td>0.047</td>
</tr>
</tbody>
</table>

Table 3: Summary statistics for YN-Task and 2AFC-Task ($N = 32$).
Table 4: Summary of MAE for different observation totals, the number of YN choices per item, in the simulation exercises ($N = 1000$ for each row).

<table>
<thead>
<tr>
<th>Observations</th>
<th>DDM</th>
<th>Logit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean MAE</td>
<td>Mean SD MAE</td>
</tr>
<tr>
<td>5</td>
<td>0.115</td>
<td>0.095</td>
</tr>
<tr>
<td>10</td>
<td>0.080</td>
<td>0.070</td>
</tr>
<tr>
<td>15</td>
<td>0.066</td>
<td>0.057</td>
</tr>
<tr>
<td>20</td>
<td>0.058</td>
<td>0.050</td>
</tr>
</tbody>
</table>
References


