American Economic Journal: Microeconomics 2023, 15(4): 146–182
https://doi.org/10.1257/mic.20210069

Supervised Machine Learning for Eliciting Individual Demand[†]

By John A. Clithero, Jae Joon Lee, and Joshua Tasoff*

The canonical direct-elicitation approach for measuring individuals' valuations for goods is the Becker-DeGroot-Marschak procedure, which generates willingness-to-pay (WTP) values that are imprecise and systematically biased. We show that enhancing elicited WTP values with supervised machine learning (SML) can improve estimates of peoples' out-of-sample purchase behavior. Furthermore, swapping WTP data with choice data generated from a simple task leads to comparable performance. We quantify the benefit of using various SML methods in conjunction with using different types of data. Our results suggest that prices set by SML would increase revenue by 29 percent over using the stated WTP, with the same data. (*JEL* C45, C91, D12)

Many economists seek to uncover aspects of consumers' latent preferences, such as their demand for a good, their tolerance of risk, or the way they discount the future. Direct elicitations incentivize individuals to make choices in such a way that they directly reveal their latent preferences. These direct elicitations could be either structure-free, such as in the Becker-DeGroot-Marschack (BDM) procedure, in which a person is incentivized to truthfully state her reservation price for a good (Becker, DeGroot and Marschak 1964), or structural, such as fitting choices between lotteries to a parametric model of a consumer's utility function to yield a risk aversion coefficient (for example, Holt and Laury 2002; Eckel and Grossman 2008; Andersen et al. 2008; Anderson and Mellor 2009). In both cases, theory provides a mapping between the data and preferences.

Direct elicitation is afforded a kind of elegance from its connection to a strong conceptual framework, but it often works better in theory than in practice. For example, BDM elicitations may suffer from biased reporting (Lehman 2015; Müller and Voigt 2010; Wertenbroch and Skiera 2002; Noussair et al. 2004; Kaas and Ruprecht

*Clithero: Department of Marketing, Lundquist College of Business (email: clithero@uoregon.edu). Lee: The Digital Economy Lab at the Stanford Institute for Human-Centered Artificial Intelligence (email: jaejoonl@stanford.edu) Tasoff: Department of Economic Sciences, Claremont Graduate University (email: joshua.tasoff@cgu.edu). Leeat Yariv was coeditor for this article. CGU IRB determined our experiment #2887 as exempt. We thank Jin Xu for valuable research assistance. We would like to thank participants at seminars at CMU, University of San Francisco, Case Western Reserve, Colegio de Mexico, Western Virginia University, The Wharton School, Karlsruhe Institute of Technology, Claremont Graduate University, USC, the WEAI Conference, and the ASSA Annual Meeting.

[†]Go to https://doi.org/10.1257/mic.20210069 to visit the article page for additional materials and author disclosure statement(s) or to comment in the online discussion forum.

2006; Berry et al. 2020) or confusion about how the mechanism works (Cason and Plott 2014). We present an alternative approach that can enhance direct elicitation using supervised machine learning (SML). We borrow the term from the computer science and statistics literature (James et al. 2015), but it is a concept familiar to all economists (Varian 2014). SML is simply the set of statistical methods that relate input variables to output variables. Regression in its many forms is a kind of SML.[1] In this paper we take an SML approach to recovering consumers' latent valuations for goods. The mapping between data and latent preferences is no longer set by theory as with a direct elicitation, but is instead uncovered statistically. Attempts to address the problems with the BDM procedure have thus far focused on fixing the front end by modifying the elicitation procedure to be simpler or more intuitive (for example, Wang et al. 2007; Miller et al. 2011; de Meza and Reyniers 2013; Mazar et al. 2014). The SML approach is to fix the problem on the back end by estimating the statistical relationship between the responses and the outcome behavior.

As a proof of concept, we apply SML to estimate a consumer's demand for a good. We show that SML outperforms raw BDM predictions of a person's purchase behavior. This is an important test, as the BDM is arguably the gold standard for eliciting WTP, the measured reservation value, for goods. Our paper presents a series of consumer choice tasks designed explicitly for an out-of-sample comparison of these approaches. Subjects in our experiment engage in three tasks involving choices over food items. The first is a task designed to elicit WTP (BDM-Task). The second is a Binary-Choice Task. We use data from these two tasks to predict behavior in the third task, which is a simple decision on whether to buy a good at a fixed price, which we will refer to as the Buy-Task. We thus have an outcome measure that approximates the kind of decisions that consumers make in the marketplace.

We find that BDM measures enhanced by SML predict purchase behavior better than the BDM measures alone, even with small amounts of data and using simple statistical methods such as logit. This is partially explained by the fact that the BDM has a downward bias in our sample. For example, subjects who state a WTP of $\$x$ will often still purchase the good when the price is $\$x + \$0.25$. However, this is not the whole story as there are considerable gains in performance as more sophisticated methods are used, such as random forest, suggesting that there are correlations in the choice data that can be leveraged to improve prediction. We also show that more data is not a substitute for better statistical methods. As sample size increases, low-dimension methods (e.g., logit) plateau at a higher mean squared error (MSE) relative to high-dimension methods (e.g., lasso and random forest). With random forest, we find that WTP data and Binary-Choice data perform about equally in predicting buy-decisions, out-of-sample, using a combined between- and within-subject analysis.

This is a notable result given that the Binary-Choice data contain only *ordinal* information about subjects' preferences between the various snacks and do not contain any *cardinal* information about the money-metric intensity of those preferences. WTP data contains exactly that *cardinal* information. Initially, the effectiveness of

[1] In contrast, factor analysis, clustering, and other statistical methods that do not have an outcome variable are forms of *unsupervised* learning.

SML with Binary-Choice Task data may appear counterintuitive. However, there is a parsimonious explanation. An array of Binary-Choice data can be used to construct an individual's ranking of the goods. When paired with the outcome variable of purchase decisions, these rankings can be linked to prices. For example, if a person prefers a granola bar to fruit juice and is willing to buy the fruit juice at $1.25, it is also likely that the person will purchase the granola bar at $1.25. We use SML to apply this kind of data and logic at scale.

Interestingly, SML with Binary-Choice data suffers considerable decreases in accuracy when we train the model on one sample of subjects and predict purchases on an entirely different sample of subjects. To be clear, this is not a consequence of overfitting, as our cross-fold validation procedure addresses that issue, but rather the data is not as generalizable for predictions of this kind. In contrast, SML with WTP suffers only minor decreases in accuracy when prediction is evaluated in this way. This suggests that patterns of preference within consumers are not effective at predicting the intensity of preference across consumers. For example, two people may have the exact same rankings of the food items, but the first person might be very hungry and have high demand for all items, while the second person may be satiated and have low demand for all items. The WTP data can reflect this but the Binary-Choice Task data cannot.

Our best predictions, though, come from using both data together, suggesting the existence of nonoverlapping information. Combining data from different elicitation methods improves performance. This result holds for both lasso and random forest.

We also show that revenue-maximizing prices predicted from our best SML models yield important deviations from a consumer's WTP. The two values are correlated, but the average absolute deviation is $0.58. Given that the average WTP of the good in our experiment is $2.03, this implies a 26.8 percent deviation in optimal pricing. Using our best SML model, we predict that our SML pricing would yield revenues 28.8 percent higher than directly setting prices to the WTP value, without using any additional data.

We view the main contributions of this paper as two-fold. First, our results draw attention to the surprisingly neglected approach of using SML to estimate latent preferences. Second, on a practical level, we show that SML can improve upon the existing canonical direct elicitation method, the BDM. Thus, we view our exercise as a necessary proof-of-concept that the SML approach to estimating latent preferences can be effective. While our particular elicitations and statistical methods are not necessarily the global optimum for estimating individuals' valuations for goods, our framework offers a template for future work in both theoretical and applied areas. For example, if highly predictive features are not well explained by theory this could motivate the development of new theories to account for the enhanced prediction. From a more applied standpoint, our elicitation method could identify latent parameters of consumers to subsequently be used in structural models. We discuss these applications and more in Section VIIIB.

Our paper is structured as follows. Section I discusses additional motivation and conceptual background. Section II presents our experimental design and Section III our empirical strategy. Sections IV, V, and VI report our results on direct elicitation, SML, and optimizing data use. In Section VII we quantify the benefits of

using SML. Section VIII discusses further implications for theory and practice, and Section IX concludes.

I. Eliciting Willingness to Pay and Predicting Behavior

Accurately forecasting consumer demand is a fundamental aim for all firms. There exist many innovations in both "indirect" and "direct" methods for elicitation (Miller et al. 2011), in both real (Ding 2007) and hypothetical choice scenarios (Schmidt and Bijmolt 2020). Our contribution builds on arguably the most popular direct elicitation method, the BDM. This section details some limitations of the BDM, why other simple choice data can help fill the gap, and why SML can make a meaningful contribution to building better predictions of individual consumer demand.

The BDM method has been hugely influential across marketing and economics. As of March 2022, Google Scholar lists more than 3,200 citations for the original paper (Becker et al. 1964). However, it has several well-established limitations. Consider the following three issues. First, the BDM produces measures that are systematically biased. Some papers find that WTP from the BDM are overstated, and other papers find that WTP from the BDM are understated.[2] We show additional evidence for this understatement of WTP.

A second issue is that direct elicitation requires specific kinds of choice tasks that may be complex and thus difficult to understand. For the case of measuring people's values for goods, many researchers have observed that the complexity of the BDM confuses subjects, which results in noisy and biased responses (see e.g., Cason and Plott 2014). Past research has found that procedural variations of the BDM that theoretically should have no effect on people's elicited responses instead have a profound effect (Bohm, Lindén and Sonnegård 1997; Urbancic 2011; Mazar, Kőszegi, and Ariely 2014; Tymula, Woelbert and Glimcher 2016).

Finally, practitioners of direct elicitation often (though not always, see Wang et al. 2007, as a counterexample) just-identify their parameter of interest as that is sufficient according to the conceptual framework of direct elicitation. However, the presence of noise in the outcome behavior (the behavior that the researcher wishes to predict) and stochasticity in the elicited behavior (the behavior used for predicting the main outcome) warrants more data. There are two potential sources of stochasticity: mistakes, such as if individuals misselect or misunderstand the choice structure, and stochastic preferences. These two sources can affect both the behavior to be predicted and the elicited behavior. So, stochasticity in one or both of the two behaviors can produce inconsistencies across choice environments.[3] In our data, we observe both purchases at prices that are above a person's elicited WTP and non-purchases at prices that are below a person's elicited WTP. Such consumer behavior is difficult to reconcile with a framework that does not build some flexibility into individual demand around a directly elicited WTP.

[2] We summarize these papers in Table A1 in the online Appendix.

[3] This could be problematic when the noise is on the elicited behavior, as this is measurement error, which could lead to mis-inference (see Gillen, Snowberg and Yariv 2019, for examples in experimental economics).

We believe SML offers a unique solution to the three issues with the BDM that we previously mentioned. First, if WTP values are systematically biased—as they are in our data—SML can debias these values by using a loss function that treats positive and negative errors symmetrically. In other words, by identifying any bias in WTP within the training sample, SML should not have such a bias in out-of-sample prediction. Second, SML can obviate the participant confusion sometimes caused by complex elicitation mechanisms, by instead relying on data generated from simple tasks. Input variables for SML are far less constrained than in direct elicitation. Potentially any kind of data (e.g., non-choice data) could be used by SML as long as the data is a good predictor for the outcome, including the data from a direct elicitation. This relaxation on data inputs allows for innovative data generation. In this paper, we estimate valuations using direct elicitation data produced using the BDM, and we illustrate how an SML approach can rely on simpler choice tasks and even non-choice data, such as response times. Third, the flexibility conferred by the SML approach also allows the researcher to leverage repeat observations, reducing measurement error.[4]

Some hurdles that all applications of SML face are worth mentioning. First, outcome data is required. The BDM can elicit a person's WTP without the need for observing any real purchase decisions. In contrast, SML needs that outcome data for model training. Second, direct elicitation, by assumption, identifies the latent preference with a single observation. Because SML is a statistical method it needs multiple observations, usually many, to generate accurate estimates. There is an additional reason why SML has been largely avoided as an approach for measuring latent preferences in favor of direct elicitation methods. Historically, a challenge facing SML was "knowledge discovery"; how does one find the statistical relationship between input variables and outcomes without overfitting? Advances in machine learning have effectively solved this problem (James et al. 2015). Algorithms search and test many specifications using brute-force computation and then settle on the best parameter values, based on out-of-sample prediction. Empirically, many popular algorithms are now effective at predicting outcomes in real-world social science data (see Lazer et al. 2009; Varian 2014; Hagen et al. 2020, for some examples).

Our paper adds to what is a growing literature applying machine learning to help predict consumer behavior, both in marketing and economics. In consumer behavior, SML has been used to understand heterogeneous responses to marketing images (Matz et al. 2019), customer retention (Ascarza 2018), and consumer finances (Netzer et al. 2019), among others. Fudenberg and Liang (2019) use SML to improve predictions in initial play in lab matrix games and to better develop an interpretable model of behavior. In the domain of price-setting, Bodoh-Creed, Boehnke and Hickman (2019) use SML to better predict variation in prices for

[4] Unsupervised learning (i.e., statistical methods for processing observations without the use of outcome data, such as factor analysis and clustering) could be used to leverage repeat observations too. For example, one can repeat a direct elicitation multiple times and then average the measures together. Or, one could also use structural unsupervised approaches, such as using nonlinear least squares or maximum-likelihood to estimate structural parameters (e.g., a CRRA risk preferences parameter or β and δ quasi-hyperbolic discounting parameters). Although we focus on SML, there are other available statistical methods to leverage repeat observations to reduce measurement error compared to nonstatistical methods.

homogenous products. For our purposes, there are also several recent papers using SML in economics. The closest to ours is Peysakhovich and Naecker (2017), who compare structural models of risk aversion and ambiguity aversion to SML predictions in order to evaluate the validity of the structural models. They find that structural models of risk aversion perform as well as SML but structural models of ambiguity aversion perform far worse. Fudenberg et al. (2022) extend this approach by constructing a metric for the "completeness" of a candidate model. SML is utilized to identify the maximum possible prediction performance, and the candidate theory's ability to approach that performance is indexed as a measure of "completeness." Finally, Camerer, Nave and Smith (2018) apply SML to predict outcomes in bargaining experiments.

We focus on a different point: we provide a proof-of-concept for recovering latent preferences using SML. Bernheim et al. (2013) use SML and non-choice data to predict the behavior of populations in aggregate. They generate voluminous data on preferences for goods using non-choice survey data and then use this to predict aggregate demand for the good at a given price. Naecker (2015) applies this method to predict organ-donation registration. Smith et al. (2014) use SML on brain imaging data to make out-of-sample predictions on human choices. Our paper is in a similar spirit to these papers, but with different practical goals. First, we are predicting individual behavior instead of aggregate behavior. This is significant because it allows for entirely different kinds of analysis. For example, individual-demand curves allow for third-degree price discrimination, in which various observables of the consumer lead to promotions (e.g., discounts for seniors, free spring roll for frequent customers). Second, we are utilizing various kinds of choice data instead of explicitly focusing on non-choice data.

A common theme across many of these SML applications is that there are aspects of a consumer's preferences unobservable to the researcher. Direct elicitation may do a good job of identifying consumer WTP, but constructing how this maps to purchase decisions requires a mapping that is unseen. We believe SML provides one path for constructing such individual demand curves. The next section details our framework for testing this claim.

II. Experiment Design

Subjects engaged in three choice tasks for 20 different food items.[5] Food items have the advantage of being commonplace and therefore easy to assess. There were 20 BDM trials, 190 Binary-Choice trials, and 80 Buy trials, as described below. Though this may sound like a large number of tasks, each trial takes only a few seconds. All the tasks combined took less than 30 minutes on average (excluding waiting time, eating time, etc). Subjects were told that one trial would be randomly selected to count and they would get the food item depending on their choice.

[5] The ability to predict behavior becomes more valuable as the variability in the predicted behavior increases. Thus, to make our study interesting, we wanted food items for which preferences are heterogenous. We conducted a pilot study in which we elicited subjects' WTP for 40 food items. We selected the 20 food items that had the highest variances in WTPs for our main experiment. The list of 20 food items is available in Table B2 in the online Appendix.

Prior to engaging in the tasks, two of the 20 food items were randomly selected to be bonus "silver" and "gold" items. Subjects saw 20 covered cards on the screen and selected two. The cards were flipped over to reveal the bonus items. Subjects received bonus payments of \$2 for obtaining their silver item and \$4 for obtaining their gold item. The purpose of these bonus items was twofold. First, it provided a check on subjects' preferences, as the monetary bonuses should increase WTP (they did). Second, it created greater variation in peoples' valuations for the food items, making the prediction task harder.

The first task elicited the WTP for each of the 20 items (BDM-Task). The task was similar to a multiple-price list version of a BDM mechanism. On the computer screen, each subject is asked to choose her WTP for each of 20 items from \$0 to \$5.75 in increments of \$0.25. To help reduce concerns about participant confusion (Cason and Plott 2014), we provided detailed instructions about the BDM and also asked subjects to complete a quiz containing five questions before starting the task. Copies of experimental instructions and the quiz are provided in the online Appendix.

The second task was two-alternative forced choice (Binary-Choice Task). On each trial, subjects were shown a picture of two items: one placed on the left side of the computer screen and one on the right. Subjects had to decide which of the two items they would prefer to eat at the end of the experiment by clicking left or right. Subjects were free to take as long as they needed to make their decision. Subjects were asked to make a choice between all possible pairs of items, so each subject had 190 binary choices. In addition to each choice itself, each subject's response time was collected for each trial. We chose the Binary-Choice Task for several reasons. First, we thought it would be one of the simplest and most easily understood ways to get real choice data. Second, response time in the Binary-Choice Task has been shown to be predictive of future choice (Clithero 2018). Third, the Binary-Choice Task represents arguably the most common form of a choice task in all of behavioral science.

In the third task, subjects faced a single food item and a posted price and had to decide whether to buy each snack or not at the posted price (Buy-Task). Each snack was shown four times with four different prices. One of the four prices was the stated WTP in the first main task, and the other three prices were randomly selected from uniform distributions over the low, medium, and high supports $[0.25, 1.00]$, $[1.25, 2.00]$ and $[2.25, 5.75]$ respectively, each segmented in 25-cent increments. The order of the items and prices was presented randomly.[6] The Buy-Task is arguably most similar to what consumers actually face: a single good listed at a single price with the option to buy or not. The behavior in the Buy-Task is what we are trying to predict.[7] Figure 1 shows screenshots of the tasks. We use various subsets

[6] The order of the three tasks was kept constant over the course of the experiment. The Buy-Task was placed after the BDM-Task so that we could use the elicited price as one of the prices.

[7] One might argue that the BDM-Task and the Buy-Task are so similar as to render prediction using data from the first task for the latter task a trivial exercise. However, presentation and framing often does affect human decisions. A large experimental literature has shown that choice in the BDM using multiple-price lists can be inconsistent with binary choice (see Brown and Healy 2018; Freeman et al. 2019, for examples and a recent summary of the literature). Though prediction using data from one task to predict behavior in the other would seem trivial, the research shows that there can be substantial inconsistencies.

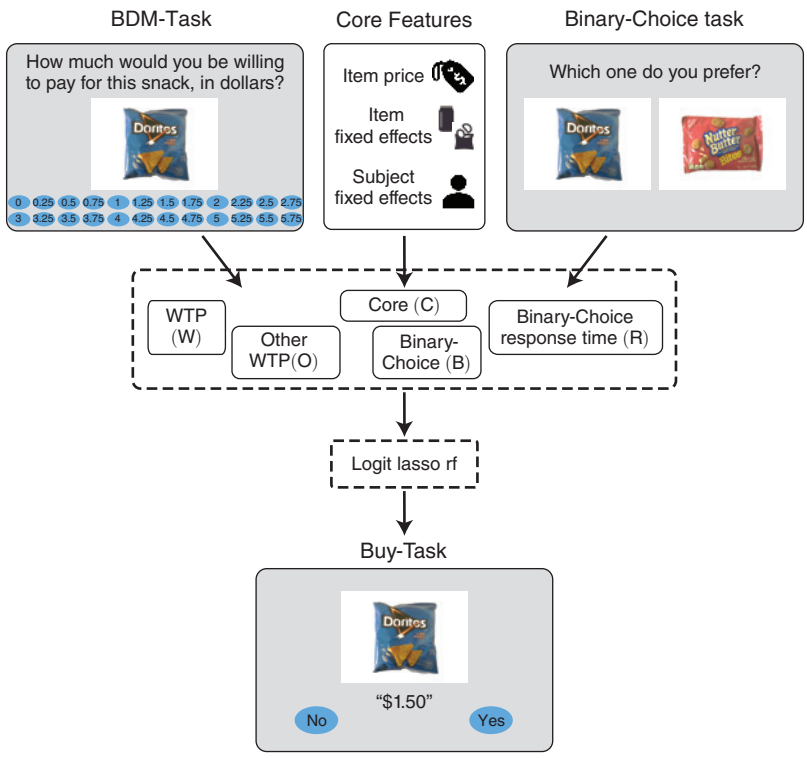

FIGURE 1. TASKS OVERVIEW AND MODELS

Notes: Summary of the empirical strategy. Three different tasks were completed by subjects: BDM-Task, Binary-Choice Task, and Buy-Task. Two different types of data from the first two tasks, WTP and Binary-Choice, along with Core features (top row), are used to construct feature spaces (second row). These feature spaces are subsequently inputted into one of three algorithms (third row) to make out-of-sample predictions on purchase decisions (bottom row).

of data from the BDM-Task and the Binary-Choice Task, fed into a variety of models to predict behavior in the Buy-Task. Details about the statistical methods are explained below in Section III.

There are two design tradeoffs that deserve discussion, both of which stem from our aim to predict purchase decisions in the Buy-Task. First, we include elicited WTP as one of the four prices per item in the Buy-Task. A reasonable question is if an aware subject would have an incentive to shade down their WTP to get a better price on the Buy-Task trial. We do not think this is a realistic concern in practice, as we did not inform subjects that their WTPs would be used as prices on a subsequent task later. This method of unexpected data use is not uncommon in the experimental economics literature (Charness et al. 2022). Given the large number of trials, as well as the usage of multiple price supports, we doubt that most participants ever realized that their WTPs had become prices on some of the Buy-Task trials. Second, as we wished to utilize BDM-Task and Binary-Choice data to predict Buy-Task data, and to have purchase decisions when faced with WTP, the design choice required placing the Buy-Task after the BDM. As a consequence, we did not randomize the order of the tasks. The lack of randomization could be a problem if preferences for

the goods changed over time. For example, subjects may have become hungrier over time or lost their appetite through overexposure. If so, the Binary-Choice trials, which were closer in time to the Buy-Task trials than the BDM-Trials trials were, may be unfairly advantaged in our analysis. We do not find evidence of any demand trends over time (results are in online Appendix C.2).

Participants were recruited through the maintained subject pool at a western university in the United States from March to August 2017. Each subject was asked to abstain from eating for three hours prior to the experiment. Subjects had to be fluent in English and have no dietary allergies or restrictions that would prevent them from consuming common snack foods. Upon arrival at the lab, all participants were consented and provided paper print-outs of the instructions. Instructions were then read out loud by the experimenter.

Each subject received the participation fee of $20, but the actual amount of cash received varied depending on their choices during the experiment and the trial randomly selected to count. In total, 55 subjects participated in the study over eight sessions. Subjects who received a food item at the end were required to take at least a single bite to prevent resale in a secondary market.[8] The experiment lasted approximately one hour.

III. Empirical Strategy for Prediction

A. *Basic Comparison*

Our main purpose is to compare our direct elicitation method, the BDM, to SML methods. Thus, the prediction of the BDM is our direct elicitation benchmark. We define the prediction of the BDM to be buy with probability 1 if the price is less than the WTP, and buy with probability 0 if the price is strictly greater than the WTP. If the price exactly equals the WTP, the person is indifferent and buys at probability q. We set q to the empirical average purchase frequency when $p = WTP$.

There are many SML methods available and two different datasets. In addition to assessing the performance of SML methods relative to the BDM, we also wish to characterize the comparative performance of the various SML methods, and the comparative performance of the two different data formats. We begin by smoothing the BDM prediction, allowing noise in the WTP via a logit according to equation (1):

$$
\begin{aligned}
(1)\quad \log\left[\frac{\Pr(buy_{ijt})}{1-\Pr(buy_{ijt})}\right] &= \beta_0 + \beta_1 p_{ijt} + \beta_2 p_{ijt}^2 + \beta_3 p_{ijt}^3 + \beta_4 WTP_{ij} + \beta_5 WTP_{ij}^2 \\
&\quad + \beta_6 WTP_{ij}^3 + \sum_{k=2}^{20} \mathbf{1}\{k = j\} \cdot \left(\gamma_k + \delta_k p_{ijt} + \zeta_k WTP_{ij}\right) \\
&\quad + \sum_{k=2}^{55} \mathbf{1}\{k = i\} \cdot \left(\eta_k + \iota_k p_{ijt} + \kappa_k WTP_{ij}\right) + \epsilon_{ijt},
\end{aligned}
$$

[8] Anecdotally, the authors observed that virtually all participants consumed the entire food item prior to leaving the experiment. This allows WTP measures to be interpreted in terms of a full item.

which constructs a probability the individual will purchase the item at the stated price, $\Pr(buy_{ijt})$. The buy-decision for person i on item j on trial t (where $t = 1,2,3,4$) is a function of a constant, the price p_{ijt}, person i's WTP for item j, item fixed effects γ_k, and item fixed-effect interactions with the price and WTP (δ_k and ζ_k, respectively), subject fixed effects η_k, subject fixed-effect interactions with the price and WTP (ι_k and κ_k), and an error term, ϵ_{ijt}. An alternative baseline comparison built around consumer surplus that performs similarly is outlined in online Appendix C.

This regression model does not make full use of all the WTP data available. Instead of merely including the WTP of the item in question we could also include the full vector of WTPs across all 20 items. Equation (2) is the same as equation (1) with two additions. First, on the second line, we include the WTP for the other 19 food items, indexed with k. Second, on the third line, the double summation term is item fixed-effects interacted with the WTP for the other 19 food items.

$$
\begin{aligned}
(2)\quad \log\left[\frac{\Pr(buy_{ijt})}{1-\Pr(buy_{ijt})}\right] &= \beta_0 + \beta_1 p_{ijt} + \beta_2 p_{ijt}^2 + \beta_3 p_{ijt}^3 + \beta_4 WTP_{ij} + \beta_5 WTP_{ij}^2 \\
&\quad + \beta_6 WTP_{ij}^3 + \sum_{k=2}^{20}\Big[\theta_k WTP_{ik} + \mathbf{1}\{k = j\} \\
&\qquad \cdot \left(\gamma_k + \delta_k p_{ijt} + \zeta_k WTP_{ij}\right)\Big] \\
&\quad + \sum_{l=1}^{20}\sum_{k=2}^{20} \lambda_{k,l}\mathbf{1}\{k = j\} \cdot WTP_{il} \\
&\quad + \sum_{k=2}^{55} \mathbf{1}\{k = i\} \cdot \left(\eta_k + \iota_k p_{ijt} + \kappa_k WTP_{ij}\right) + \epsilon_{ijt}.
\end{aligned}
$$

This allows for the WTP of a given good l to have a different effect on each good j. For example, a positive coefficient of $\lambda_{1,2}$ would imply that the WTP of item 2 increases the probability of buying good 1. The $\lambda_{k,l}$ may capture substitutability between goods, as a positively correlated valuation between two items may mean that they are very similar.[9]

We will refer to the first model as the logit(WTP), since it only uses the WTP for the item in question, and logit(WTP + OtherWTP) for the more complete model. The latter model is not obviously better than the former for prediction because it could overfit the data.

[9] Though the parameters in this model may relate to the concepts of complements and substitutes, the context here is different than in consumer theory. In the classic consumer theory problem, a consumer must decide how to allocate their budget across many different goods with a single choice set. When a compensated increase in the price of good l increases demand for k it is said that the goods are substitutes, and if it decreases demand for k it is said the goods are complements. In the BDM-Task consumers are faced with many choice sets, each with only two goods: the item and cash/numeraire. Since only one decision problem is randomly selected to count there cannot be any complementarities. For example, if there were a correlation in valuations between peanut butter and jelly in our BDM-Task, this does not directly represent a complementarity or substitutability because a subject will receive only peanut butter or jelly, not both.

B. *High-Dimensional Methods*

"High-dimensional" methods are methods that are well-suited to analyzing data with a large number of independent variables. We use two off-the-shelf methods that are widely adopted in the literature (all estimation was conducted using freely-available packages in R). We will describe the basic intuition and mechanics here. More detailed treatments can be found in textbooks such as Hastie et al. (2009) and James et al. (2015). The first method is a logistic regression with a least absolute shrinkage and selection operator (lasso) penalty (Tibshirani 1996). The lasso algorithm first normalizes all the independent variables. Then a regression is run, but for each independent variable included in the regression there is a marginal penalty applied to the objective function (e.g., mean squared error or binomial deviance, depending upon the outcome variable of interest), for each unit that a coefficient is away from zero. Because the marginal loss of increasing the magnitude of a coefficient is constant, coefficients on variables that do not reduce the error term enough lead to a corner solution, yielding coefficients of zero. This method is well-suited for high-dimensional data as it does not include variables in the model that do not substantially improve prediction. Since our outcome is dichotomous, the lasso we use is a penalized logit in which the objective is to minimize binomial deviance. Lasso is already popular in economics (Mullainathan and Spiess 2017) and is an intuitive modification to a logit prediction.

Tree methods take a different approach. A tree is a partition of the dataset that assigns the mean of the outcome within the partition to be the predicted value for all observations in the partition. In other words, it is a regression run on partition indicators and nothing else. The partitions are built using a myopic algorithm that first selects an independent variable and a value, then creates an indicator partitioning the dataset into all observations below this value and all observations above this value. This is repeated for every possible division and every independent variable. The indicator that improves the objective function by the most is selected to be the first "node" in the tree. The process is then repeated several times to create the tree.

A random forest is an algorithm that builds many trees on the same dataset and averages the predictions together to reduce variance. There are two randomizations in the generation of the trees. First, many bootstrap samples using the full sample size are produced. A tree will be produced for each bootstrap sample. Second, only a random subset of independent variables will be included in the generation of branches at each node of each tree. This restriction de-correlates the trees reducing the overall variance of the estimator (James et al. 2015). We use random forest because it is a widely used method, and has a record for high performance (Athey and Imbens 2019).

Each of these methods have tuning parameters. Lasso has a penalty parameter. Random forest has parameters for the number of nodes in the tree, the number of trees, and the size of the subset of variables to be considered at each node. We optimize these parameters according to the conventional methods in machine learning, which involves different tuning approaches for the different methods. For lasso we used an approach called 10-fold cross-validation. First we split the sample randomly into a 90 percent training set and a 10 percent test set. The training set is then randomly split

into 10 folds (i.e., subsets). A parameter set is chosen, and the model is estimated on 9 of the 10 folds with the loss function (e.g., MSE or binomial deviance) measured out-of-sample on the remaining validation fold. This is then repeated 10 times, each time rotating the validation fold. The loss averaged over the 10 folds is a performance metric for that given parameter set. The process is repeated iteratively over the entire parameter space. The parameter set that gives the least loss is selected. The model with the optimized tuning parameters is then re-estimated using the full training set and evaluated using out-of-sample loss on the test set, as detailed in the next section. Random forest, the most computationally intensive algorithm we used, employs a different tuning procedure. Random forest offers a straightforward way to estimate out-of-sample performance without cross-validation. Each bootstrap sample omits approximately $1/3$ of observations, referred to as out-of-bag (OOB) observations (James et al. 2015). One can predict a given observation by using the average of all trees for which the observation is OOB. The error is referred to as the OOB error. The algorithm then selects the parameter set to minimize the OOB loss (e.g., MSE or binomial deviance).

C. *Performance Metrics*

We evaluate performance based on out-of-sample prediction. As described above, for each analysis we have a randomly selected training set and test set. In our main analysis, we randomize at the observation level, where each observation is a buy-decision (four buy-decisions per subject-item). We reserve 440 observations (10 percent) of our data as the test set. The 440 observations were drawn randomly using stratified sampling. We divided the data into quintiles, based on a subject's percentage of purchase decisions. An equal number of observations were randomly drawn from each of these five bins. Due to the potential variance generated from a single random test set, we repeat the process N (here, $N = 50$) times. This repetition helps smooth out the prediction and ensures differences in performance are not attributable to particularities of one test set.

In the main body of the paper we use mean squared error (MSE) as our performance metric due to its widespread use. The MSE across the N test sets is then as follows:

$$MSE = \frac{1}{N}\sum_{n}\sum_{i,j,t}\left[\Pr\left(buy_{ijt}\right) - \hat{\Pr}\left(buy_{ijt}\right)\right]^2, \tag{3}$$

where $\Pr\left(buy_{ijt}\right)$ is the observed decision (i.e., ether 1 for "Yes" or 0 for "No") and $\hat{\Pr}\left(buy_{ijt}\right)$ is the predicted choice probability from SML. Here, each MSE test set n effectively captures the mean squared error across all subjects, items and prices. We then average across the N randomly constructed test sets to obtain our final measure of prediction performance.

Two other measures that appear in the literature are binomial deviance and "area under the curve" (AUC). Binomial deviance is the negative of the binomial log-likelihood. In a logit regression, the objective is to minimize this quantity. The AUC is a metric that summarizes the performance of a binary classifier and is often used in the machine learning literature. A perfect classifier will have an AUC equal

to 1, whereas a random classifier would have an AUC equal to 0.5. In our analyses, we focus on MSE. MSE has the most desirable properties as binomial deviance becomes infinite when a model makes a prediction with certainty and is wrong (as the BDM does), and AUC can be misleading for a binary predictor (Muschelli 2020) and requires large amounts of data. Due to the deficiency of AUC in our binary context, we also provide performance on classification accuracy with the threshold of 50 percent (to predict either "Buy" or "Not Buy") in Table C4 of online Appendix C.3. For completeness, we include AUC results in Figures C5 and C6 of online Appendix C.3. Both the classification results and the AUC results are largely consistent with our MSE results.

D. *Data and Features*

To perform our prediction exercise, we need to create the explanatory variables, a step commonly called feature construction. While modern SML methods have taken the art out of specification search and replaced it with an effective systematic approach, there remains a need for the researcher to generate the explanatory variables. The SML literature refers to these explanatory variables as features. We construct several sets of features, each stemming from one of our sources of data. We then leverage this modular design to understand how different kinds of data contribute to our ultimate prediction exercise. A summary of these feature modules is in Table 1 and a full list of all features can be found in online Appendix B.2.

The first set of features for us is a set of "Core" features that will be used in all predictive models. The features are constructed using price variables, fixed effects for subjects, and fixed effects for items (i.e., indicators for subjects and items). These features can also serve as a baseline for prediction, as a practitioner could construct this set of features without any additional data. In other words, if we want to predict purchase behavior in our Buy-Task, these are features we can use without considering any of our WTP-Task or Binary-Choice Task data.

As Section IIIA explains, we make use of the WTP dataset in two ways to construct features. "WTP" (W) models use data that only uses the WTP of the item in question in the analysis, including a polynomial expansion of WTP and the price, item fixed effects, interactions between WTP and the item indicators. The more expansive "WTP + OtherWTP" (WO) models include all those features as well as interactions between every WTP and every item indicator. These models include the full 20 item vector of WTP values in each regression. We begin by running simple logit models to predict the buy decisions as expressed in equation (1) and equation (2) giving us the models logit(W) and logit(WO). After logit, we run lasso and random forest on the W and WO feature spaces. Comparing logit(W) to the random forest model, rf(W), tests whether there is a gain in performance from using high-dimensional methods but with low-dimensional data. Comparing logit(WO) to random forest with the full feature set, rf(WO), tests whether there is a gain in performance using the high-dimensional method with high-dimensional data.[10]

[10] Our point here is to measure how much improvement there is from using high-dimensional data *only*, high-dimensional methods *only*, and both *simultaneously*. For exposition, we omit the lasso models using WTP as

Table 1—Summary of Feature Models

Groups of features	Total Features	Abbreviation
Core	149	C
Core + WTP	225	W
Core + WTP + OtherWTP	625	WO
Core + Binary-Choice	360	B
Core + Binary-Choice + Response Times	447	BR
Core + WTP + Binary-Choice	436	WB
Core + WTP + OtherWTP + Binary-Choice	836	WOB
Core + WTP + OtherWTP + Binary-Choice + Response Times	923	WOBR

Bodoh-Creed, Boehnke, and Hickman (2019) conduct a similar analysis in their exploration of price dispersion on eBay.

The second dataset is from the Binary-Choice Task, from which we have both the choice data and response times. On this dataset we run the same algorithms, logit, lasso, and random forest each with and without response-time data. The Binary-Choice data require more careful feature construction to have data that is in a form that the algorithms can use effectively. The "Binary-Choice" (B) models add 211 different features in our regressions based on the binary choice data. Our approach is to create item fixed effects and dummy variables for an item being chosen over a specific different item. For example, a dummy if the item in question was chosen over Godiva Dark Chocolate, another dummy if the item in question was chosen over KIND Nuts and Spices, and so forth. We also construct variables for the frequency with which the item in question was chosen.

As mentioned earlier, we recorded response-time data on each Binary-Choice trial. The existing literature suggests shorter response times should be correlated with a stronger preference for the chosen good relative to the unchosen good (Clithero 2018; Krajbich et al. 2010; Philiastides and Ratcliff 2013) and that response-time data can increase predictive performance in certain choice environments (Clithero 2018). We add another 87 features when we include the response-time data (R). The main type of feature in consideration with response time is an interaction term between dummies for whether the item in question was chosen over a specific item (e.g., it takes the value of one when the item in question was chosen over KIND Nuts and Spices). In principle, response-time data could mitigate the primary weakness of Binary-Choice data, which is a lack of data on preference intensity.

All combined Core, WTP, and Binary-Choice data lead to 836 features without response-time data and 923 features with response-time data. This partitioning of our feature space is an effort to combat a common criticism of SML, that it often does not have interpretable results. By creating modules that come from different kinds of data, we can isolate where there is added predicted value of the data.

they perform worse than random forest in this context. With the Binary-Choice data, which is higher dimension than the WTP, we devote more time to comparing the high-dimension methods, lasso, and random forest, to each other.

E. *Statistical Models and Research Questions*

Our empirical strategy is summarized in Figure 1 and Table 1. Figure 1 diagrams the tasks, data, and how the data feeds into the various statistical models. As discussed, we have several baseline prediction methods to use as benchmarks. We have the BDM prediction, which will not require any model fitting. As another simple prediction metric, we have the empirical purchase frequency of our sample, which can be used as the probability an individual purchases the item on each trial (which we will refer to as "Prob(Buy)"). Finally, we have the set of Core features, which we run in each of the algorithms used. We use logit, lasso, and random forest.

The nature of our dataset also allows us to vary the sample size of the data used to train algorithms. We are thus able to investigate if and how this impacts prediction performance. We are also interested in the extent to which high-dimensional methods (here, lasso and random forest) can improve predictions. We can compare algorithms with the same data as inputs. For example, comparing rf(WO) to logit(WO) reveals the improvement from using a high-dimensional method over logit.

As we have two different kinds of data, direct elicitation (WTP) and binary choice, we can compare the performance of these data using the various feature space modules summarized in Table 1. We can also look within the Binary-Choice data and determine if, given all of the other data we have, response-time data can be of additional predictive value for our prediction exercise.

Finally, we also wish to know whether the Binary-Choice data and the WTP data are redundant or if there is unique information in each dataset that would imply improved performance by using both? For this question we run models using feature space combinations of both, WB and WOB. Comparing rf(WOB) to rf(WO) and rf(B) reveals the improvement from combining WTP and Binary-Choice data together.

To summarize, the analysis will allow us to answer several important questions:

(*i*) What is the degree of bias in WTP using the BDM to measure reservation value? This can be done by measuring the predicted probability of purchase when the price is exactly at the WTP.

(*ii*) How does algorithm performance vary with sample size? Algorithms with more features may require higher sample sizes for good prediction. We can observe for what sample sizes a high-dimensional algorithm outperforms a low-dimensional model.

(*iii*) Which algorithm performs best with our data? We compare three different algorithms, as outlined in Section IIIB.

(*iv*) Which data best predicts purchases? We compare all feature spaces within a given algorithm.

(*v*) Are WTP and Binary-Choice data redundant? We can test this by constructing feature sets with various combinations of WTP and/or Binary-Choice data.

The next three sections answer these questions directly. We examine possible applications of the method in Section VIIIB.

IV. Direct Elicitation

We begin our analysis with a look at the data. High-quality prediction is only interesting if there is a high degree of variation in the behavior being predicted. For the prediction task to be interesting, we want the dataset to have a high degree of variation in WTP within-item and within-consumer. If everyone values a Hershey's Bar at exactly \$1.25, prediction is trivial. Likewise, if demand for all goods can be easily characterized by a one-dimensional parameter, such as hunger, our elicitation devolves to just measuring hunger. A more challenging task is one in which there is heterogeneity in both hunger and tastes.

Figure 2, panel A shows the distribution of WTP by item. There is a high degree of heterogeneity across items, with average WTP ranging from \$1.10 for a 12 oz Coke (CK) to \$2.99 for Naked Mango Juice (NM). There is also considerable heterogeneity within item. The mean WTP is \$2.03 and the standard deviation is \$1.27. Figure 2, panel B shows the heterogeneity across individuals. Again there is ample heterogeneity ranging from a mean WTP of \$0.21 to \$3.40.[11] Also within person, the average standard deviation is \$1.11, indicating that people have different valuations across the products.

In Figure 3 we plot purchase frequency as a function of WTP minus price. The dots represent the mean purchase frequency for individuals at that implied surplus (WTP minus price). Purchase frequency fits a smooth S-shaped curve. This would seem to be a success for the BDM as a method. As a benchmark, we plot the dotted step function that represents the actual predictions of the BDM. On the one hand, the BDM step-function roughly approximates the curve, but on the other hand there is a considerable degree of error. Note that the average WTP of the goods in the sample is \$2.03. Consider a consumer surplus of \$|1| in Figure 3, which represents a roughly 50 percent change in the value of the average good. Even at such large surpluses relative to product value, purchases are still roughly 10 percentage points away from the BDM prediction.

The second important weakness of the BDM prediction is a systematic bias. This is evident by noting that the probability of purchase at 0 surplus is 62 percent. This suggests a systematic understatement of people's reservation value. Intuitively, this may make sense if people use a bargaining heuristic when stating their WTP (Plott and Zeiler, 2005). If people view their stated WTP as what a person would verbalize in a bargaining context, they might always state a WTP below their true reservation price in order to obtain positive surplus conditional on obtaining the item. In addition, research has shown that the way the options and distribution of prices is presented in the BDM affects people's choices (Urbancic 2011; Tymula et al. 2016). There may also be differences in the cognitive process underlying contingent choices versus

[11] One possible explanation for this is that some people are considerably hungrier than others, or that their hunger plays a larger role in their calculation of WTP. Histograms of WTP for each item are provided in the online Appendix in Figure C1.

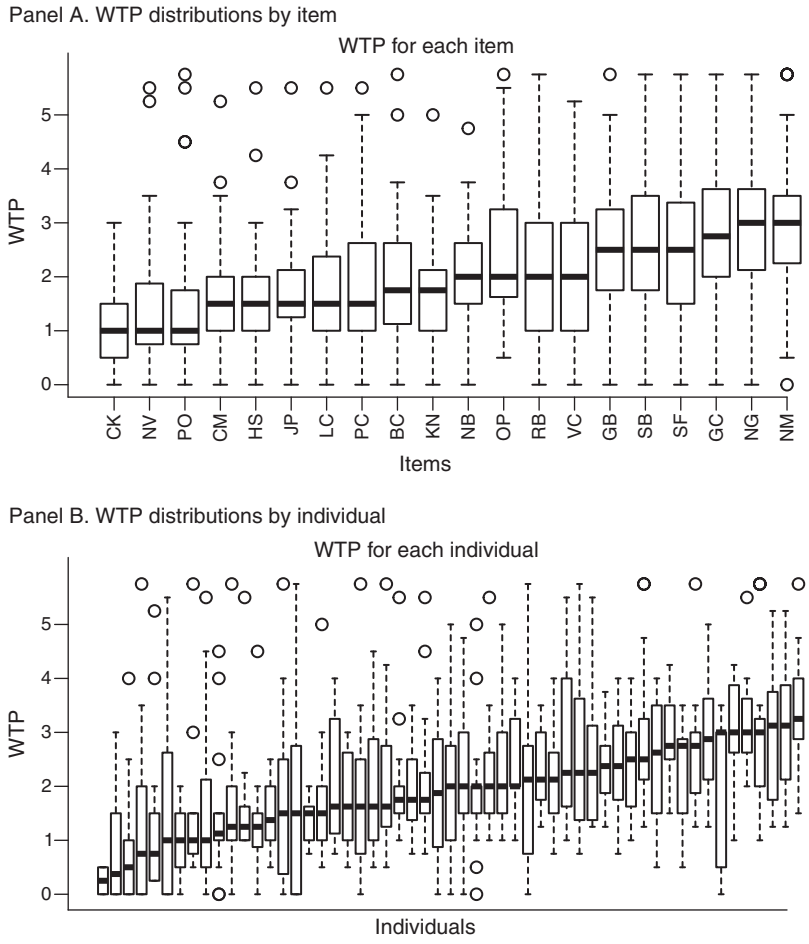

FIGURE 2. DISTRIBUTIONS OF WTP

Notes: All WTP are stated in US dollars. The line inside the box indicates the median, the top and bottom of the box indicate the seventy-fifth and twenty-fifth percentiles, the whiskers indicate the upper and lower adjacent values, and the circles indicate outside values. Food items and abbreviations are as follows. Bar Clif Peanut Crunch BC; Chex Mix CM; Coke CK; Godiva Dark Chocolate GC; Green and Blacks Organic Chocolate GB; Hershey's Bar HS; Justin's Peanut Butter JP; KIND Nuts and Spices KN; Luna Choco Cupcake LC; Naked Green Machine NG; Naked Mango NM; Naturally Bare Banana NB; Nature Valley Crunchy NV; Organic Peeled Paradise OP; Pretzel Crisps Original PC; Pringles Original PO; Red Bull RB; Simply Balanced Blueberries SB; Starbucks Frappuccino SF; Vita Coco VC.

noncontingent choices (Brandts and Charness 2011). Another interpretation starts from the view of stochastic choice. People may not actually have static reservation prices, but instead reservation prices that fluctuate over time. Obtaining the good in the future at even a known price is then a risky lottery. If people are risk averse and/or loss averse this could deflate people's stated WTPs as people may be concerned that their tastes will change by the time they actually receive the item. A third interpretation is that as the experiment progresses, demand for the food items goes up. The BDM-Task always took place before the Buy-Task. People may have become hungrier over time or the constant exposure to pictures of food may have increased

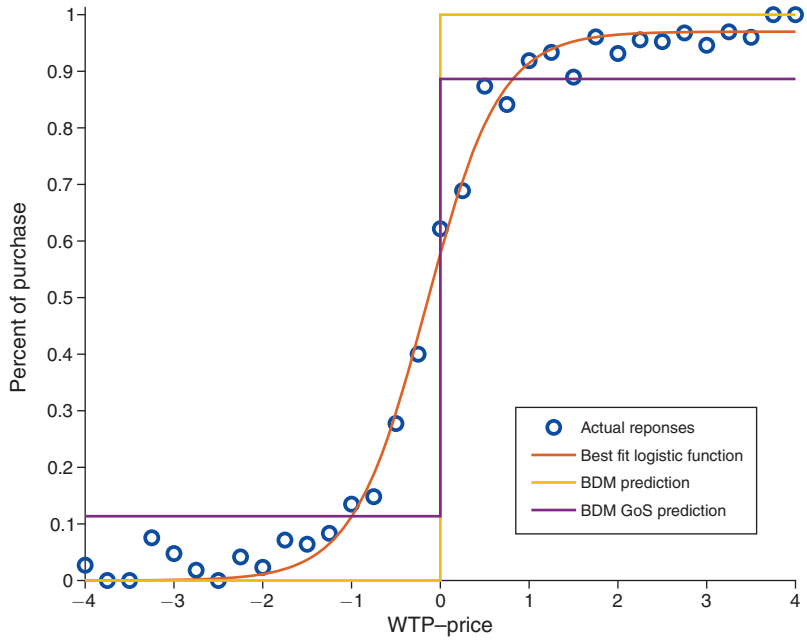

FIGURE 3: PURCHASE FREQUENCY AS A FUNCTION OF SURPLUS

Notes: Each dot represents the probability of purchasing for a given bin with the specified consumer surplus (i.e., WTP—Price). Both BDM (lighter) and BDM Grain of Salt (GoS, darker) predictions are plotted.

demand. While this is possible we do not observe robust evidence of this (see online Appendix C.2).

The experiment was not designed to test between these mechanisms. We document the downward bias here and note that it is an important justification for finding alternative methods for measuring individual valuation (Miller et al. 2011). One approach to "debias" BDM choice predictions would be to use a model $buy_{ijt} = \mathbf{1}\{WTP_{ij} - p_{ijt} + x \geq 0\}$ instead of the standard BDM prediction of $buy_{ijt} = \mathbf{1}\{WTP_{ij} - p_{ijt} \geq 0\}$. One could then find the x that minimizes MSE. The scale we employed to elicit WTP used intervals of 0.25, so this limits our level of granularity. We found that $x = 0$ performs better than $x = 0.25, x = 0.50$, or larger intervals. However, if we used smaller increments it is likely that a strictly positive x would minimize MSE given the apparent bias.

Most crucial for our purposes, visual inspection reveals that the logit curve representing the predicted values from a logit regression in Figure 3 appear to fit the data better than the BDM prediction. We can decompose these gains. The BDM prediction is a step function from 0 to 1, in which the step must be at exactly zero. The logit curve can fit better by (a) predicting interior probabilities between 0 and 1, and (b) using a more flexible and smoother shape.[12] We can improve the BDM predictions by also allowing it to predict interior probabilities. Rather than predict-

[12] Logit has the potential to suffer from overfitting whereas the BDM does not, but this is unlikely to happen given our sample size.

ing that $WTP > p$ implies purchase at 100 percent and 0 percent otherwise one can say that it implies purchase at the empirical probability of purchase. Empirically, this prediction is incorrect 11.36 percent of the time. Therefore, one can use a model that predicts purchase at 88.64 percent when $WTP > p$, and 11.36 percent when $WTP < p$. We call this model "BDM Grain of Salt" as it makes the BDM prediction, but without absolute certainty. Comparing BDM to BDM Grain of Salt then reveals the value of (1) predicting interior probabilities, and comparing BDM Grain of Salt to logit then reveals the value of (2), a more flexible and smoother shape.[13] We present quantitative evidence on these points in the next section.

V. Supervised Machine Learning

A. *WTP Data*

We now compare SML to the direct elicitation in predicting purchases on the Buy-Task. We use the same dataset, the output of the BDM-Task. We reserve 440 observations (10 percent of our sample) to compute out-of-sample MSE, as detailed in Section IIIC.

Consideration of the results begins with our benchmark models plotted in Figure 4, panel A. Like the rest of Figure 4, it displays the out-of-sample MSE as a function of the sample size in the training sample, progressing in intervals of 200, starting at 600. We benchmark performance against a model that predicts that any decision is "Buy" with probability r, where r is the mean purchase frequency. This sets a lower bound for performance. We find that $r = 0.556$ and the MSE equals 0.2470 for the full sample. The second benchmark, BDM, has a MSE of 0.1494, well below the MSE of Prob(Buy), and constant across sample size. This is because there is virtually no fitting of the model. The single parameter is the probability that the person purchases the good when the price exactly equals the WTP. We use the population average in the training sample, and, due to the law of large numbers, this doesn't change much as our sample increases. The only other variation comes from the hold-out sample, and here the variance is reduced since we average over 50 random samples.

Our third benchmark is BDM Grain of Salt. This is displayed as the lowest of the three horizontal lines in Figure 4. The MSE for BDM Grain of Salt is 0.1402. This represents a modest improvement over the BDM.

Several clear trends emerge for the models using WTP data in Figure 4, panels A and B.[14] First, at 1,400 observations and higher, logit(W), our most basic statistical model, beats the BDM, and at 1,800 observations and higher it beats BDM Grain of Salt. The improvement continues as sample size increases, but most of

[13] We thank an anonymous referee for suggesting this decomposition.

[14] For ease of comparison, we use the same axes in all plots in Figure 4. A companion pair of plots for the logit models with extended axes is provided in the online Appendix in Figure C4. We also want to note one limitation of the logit model. For training samples under 3,000 observations, we frequently encountered rank-deficient matrices during model fitting. This translated to greater variability in out-of-sample predictions. We include the results for all samples to illustrate this limitation of logit for smaller training samples. This applies for both the WTP and Binary-Choice data.

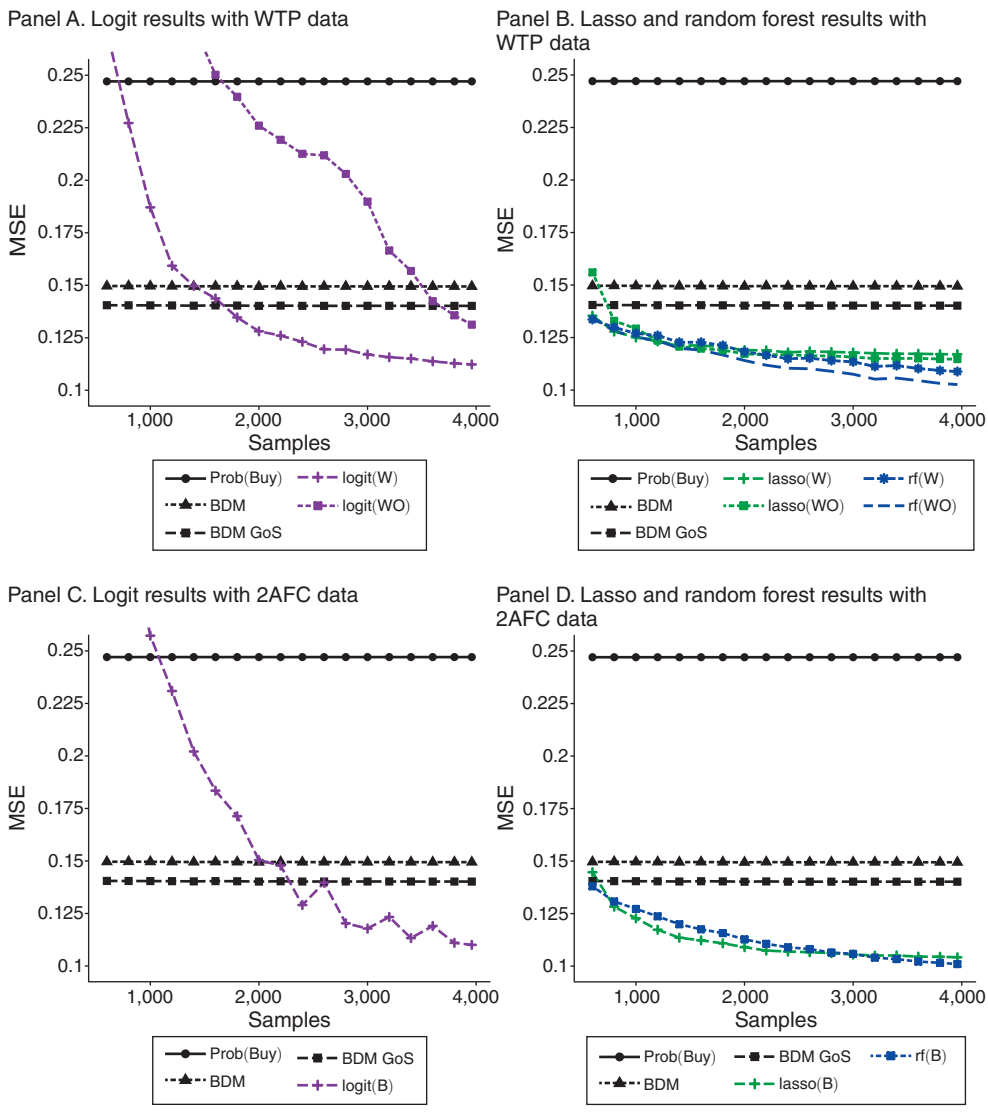

FIGURE 4. PERFORMANCE OF SUPERVISED MACHINE LEARNING WITH WTP AND BINARY-CHOICE DATA

Notes: Out-of-sample MSE estimated on 440 hold-out observations. Sample size increases from 600 to 3,960 in intervals of 200. Since samples are random, we repeat the estimation 50 times and MSE is averaged.

the gains occur before 3,000, ending with an MSE of 0.1123. We contrast this with our most sophisticated analysis of the same feature space, random forest. From the plot in Figure 4, panel B, it is clear that rf(W) outperforms the BDM and BDM Grain of Salt from the smallest sample size, 600. It exhibits continual improvement throughout our entire training sample, up to 3,960 observations, with a final MSE of 0.1088. This shows the considerable benefit of using high-dimensional data with high-dimensional methods.

We next decompose the gains by data and method. The logit(WO), shown in Figure 4, panel A, uses the full vector of WTP and begins with a performance of

0.4667 MSE at 600 observations, doing even worse than Prob(Buy). The MSE declines rapidly but only beats the BDM after 3,600 observations, and achieves a MSE of 0.1312 at the maximum sample, worse than its lower-dimension sister model, logit(W). This shows that using high-dimensional data with a low-dimensional method results in poor performance. As discussed previously, rf(W) outperforms the BDM and the BDM Grain of Salt at all sample sizes, and the same is true for rf(WO). Furthermore, rf(WO) outperforms rf(W) starting at a sample size of 800, with the gap growing until the full number of training observations is reached. The final MSE for rf(WO) is 0.1026, lower than that of lasso(WO). Likewise, rf(W) outperforms logit(W). This evidence shows that the gain in performance comes from using high-dimensional data with high-dimensional methods. Bodoh-Creed, Boehnke, and Hickman (2019) find similarly that high-dimensional data with high-dimensional methods are required to adequately explain price-dispersion on eBay. Cross matching the data with the method produces worse results than the more basic low-dimension data, low-dimension method, logit(W). Visual inspection of the lasso(W) and lasso(WO) results show that lasso(WO) also outperforms lasso(W), starting around 1,400 observations. The gains are more modest, though, with a final MSE of 0.1170 for lasso(W) and 0.1147 for lasso(WO).

B. *Binary-Choice Data*

Can we predict purchases at specific price levels using the Binary-Choice data—data which provides an ordinal ranking of alternatives but contains no cardinal information about reservation values? The answer is yes. All of the statistical models beat the baseline BDM benchmark. For the logit(B) model, Figure 4, panel C shows the MSE as a function of the sample size of the training sample. At roughly 2,200 observations and higher, the model performs better than the BDM, and at 2,400 observations and higher it beats BDM Grain of Salt. For the high-dimensional methods, shown in Figure 4, panel D, superior performance comes almost immediately, as prior to 1,000 observations all lasso and random forest models are better.

Surprisingly, we see no gain from using response-time data. All models using response-time data perform worse initially than their sister models without response times. As the sample size grows, the models with response times approach or converge to the performance of their sister model without response times. With the full sample, sister models without and with response times have almost the same MSE for both lasso and random forest. More details, including a corresponding plot in Figure C7, are provided in the online Appendix.

Both lasso and random forest with Binary-Choice data perform well, beating out logit(B). Inspecting the plot shows that the two high-dimensional models asymptotically approach different performance levels, suggesting more data is not a substitute for a better algorithm. For lasso, the MSE at the full sample is 0.1042, and for random forest it is 0.1010, compared to 0.1100 for logit(B).

Perhaps the most stark finding comes from comparing the results for Binary-Choice data to WTP data. The MSE (standard errors in the parentheses) for rf(B) at the full training sample, at 0.1010 (0.0012), is approximately the same as the MSE for rf(WO), 0.1026 (0.0012). For lasso, Binary-Choice data has a smaller MSE at 0.1042 (0.0012)

compared to 0.1147 (0.0013) for the full WTP dataset. Even a conservative interpretation would grant that the Binary-Choice data provides equivalent out-of-sample prediction to the WTP data for subsequent purchase behavior.[15] The Binary-Choice data produces high-dimensional information on each subject's ranking of the food items. Cardinal information on the value of each item is entirely lacking in the raw data.

It is worth noting that prediction here is both within-subject and between-subject. A given individual's Buy-Task decisions may be split up into the training sample and the test sample. The question we are answering is whether one can predict the behavior of individuals with Binary-Choice data given their past purchase behavior. A different question is whether one can predict the purchase behavior of a set of individuals using only data from a different set of individuals. This latter exercise would be an entirely between-subject prediction.

We posit that the strong performance of the rf(B) model comes from high-quality prediction within individuals since the Binary-Choice data represents ordinal-ranking data within person. We suspect that the between-subject performance with Binary-Choice data would do worse because there is no direct data on the relative valuations of goods between subjects. We return to this question in Section VID.

In summary, we find that high-dimensional data with a high-dimensional method offers the best predictions (i.e., rf(WO)). The WTP data and the Binary-Choice data are nearly equally predictive of buy decisions when one uses random forest (i.e., MSE of rf(WO) $\approx$ MSE of rf(B)). But much of the gain relative to the BDM benchmark is achieved through a simple logit (i.e., logit(W)).

VI. Optimizing Data Use

A. *Opening the Black Box*

The random forest algorithm yields models that escape conventional interpretation, such as comparing regression coefficients. These models generate many regression trees and then average over these trees. In the original paper, Breiman (2001) detailed other statistics that can shed light into the black box. We explore two such measures: the mean decrease in accuracy and the mean decrease in Gini index in the rf(WOB) model, the best performing model. The mean decrease in accuracy is the increase in prediction error from taking a feature and randomizing it without replacement across the observations. The Gini index is the total decrease in prediction error from splitting on the variable, averaged over all trees (James et al. 2015).

According to these metrics, the best predictors overall are price, polynomial expansions of price, the percentage of times the item was chosen in Binary-Choice trials, the rank of the item in terms of times chosen in Binary-Choice trials, WTP, and item fixed effects. In other words, the random forest algorithm leveraged distinct pieces of data

[15] The similarity between the BDM-Task and the Buy-Task should favor the predictiveness of the BDM data for the Buy-Task behavior. This holds for both the WTP predictions and the SML models that use WTP data. With this in mind, the equivalent performance of the Binary-Choice models is more surprising.

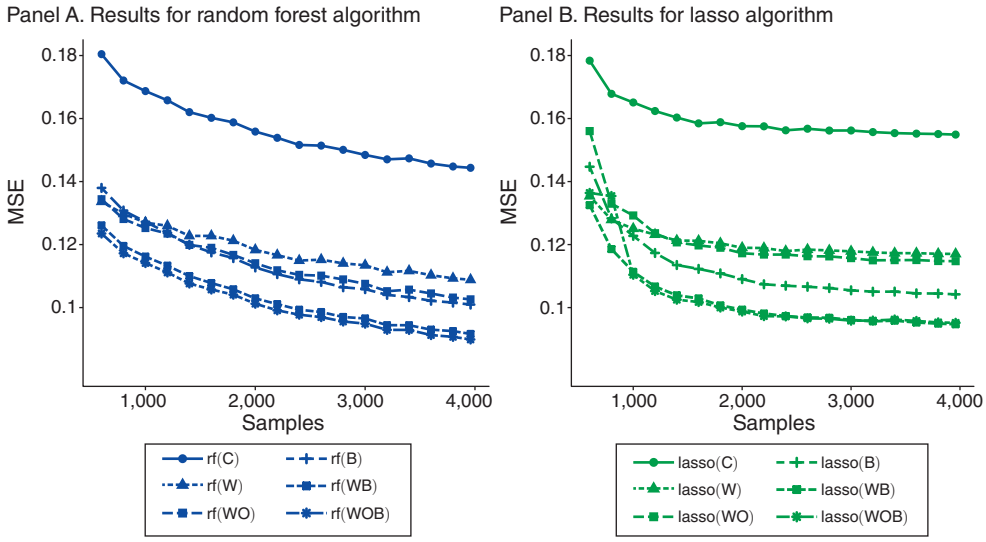

FIGURE 5. PERFORMANCE OF SUPERVISED MACHINE LEARNING WITH ALL COMBINATIONS OF DATA

Notes: Out-of-sample MSE estimated on 440 hold-out observations. Sample size increases from 600 to 3,960 in intervals of 200. Since samples are random, we repeat the estimation 50 times and MSE is averaged.

from the Core, WTP, and Binary-Choice data. This suggests that the WTP data and Binary-Choice data might not be redundant. We address this question next.

B. *Are WTP and Binary-Choice Data Redundant?*

One way we can consider the possible redundancy of WTP and Binary-Choice data is by including both datasets in a statistical model. Figure 5 shows the results for all training sample sizes. For the random forest algorithm, presented in Figure 5, panel A, we include several feature space combinations for comparison, starting with the baseline Core model, rf(C). The model of interest is random forest on the union of the WTP and Binary-Choice features called rf(WOB), which yields better performance at all sample sizes. MSE decreases continuously across the sample with a minimum of 0.0899 (0.0012) when the full training sample is used. This MSE is less than the 0.1026 (0.0012) for rf(WO) and 0.1010 (0.0012) for rf(B). This shows that there is indeed some non-redundant information in the two datasets.

Seeing the improvement from WO and B to the combined WOB, a practitioner might also be interested to know if a slightly smaller feature set could accomplish the same improved prediction. One option would be to simply focus on the WTP data for the item for which a practitioner wishes to predict purchase behavior. In other words, combine W with B feature spaces. From Figure 5, panel A we can see that rf(WB) outperforms rf(W) and rf(B) at all sample sizes, but underperforms relative to rf(WOB).

For comparison, lasso results for the same combinations of features are shown in Figure 5, panel B. In general, all of the trends observed with random forest are

also found in the lasso results. Clearly, the combination of WTP and Binary-Choice data improves predictions. The combination WOB has an out-of-sample MSE with the full training set of 0.0952 (0.0013), smaller than lasso(B) and lasso(WO). Interestingly, there is no meaningful gap between lasso(WOB) and lasso(WB). However, the MSE for lasso(WB) and lasso(WOB) are both greater than the respective MSE for rf(WB) and rf(WOB).

C. *Where Does the BDM Get Reservation Value Wrong?*

In Figure 6 we plot the MSE of the BDM as a function of the surplus as implied by WTP. Not surprisingly, most of the error comes from when the surplus is near zero. The interesting part is that MSE of the BDM hits a maximum at a surplus around −$0.25. This provides more evidence WTP from a BDM systematically understates reservation value. In other words, Figure 6 shows that WTP leads to more incorrect predictions of not buying (i.e., when price $>$ WTP) than buying (i.e., when WTP $>$ price).

We also plot the MSE of our best-performing SML model, rf(WOB). Random forest models rf(WO) and rf(B) are also plotted for comparison. As can be seen from Figure 6, the gain from the SML models is mostly in the surplus range of −$1 and $1.50. All models perform approximately the same outside this range, suggesting an irreducible unpredictability given the inherent stochasticity of the Buy-Task data.

D. *Between-Subject Analysis and Between-Item Analysis*

The Binary-Choice data predicts remarkably well given its lack of cardinal information. Thus far, we have been predicting out-of-sample purchase decisions using both data within-person and between-person. Since Binary-Choice gives detailed ordinal information about a persons' demand within person, we suspect that performance might be driven by strong within-person prediction and slightly weaker between-person prediction. For example, there is nothing directly in the Binary-Choice data that would indicate overall hunger or demand, whereas this may be evident from heterogeneity in WTPs across subjects. On the other hand, there may be specific choice patterns that are indicative of overall demand. For example, very hungry people may exhibit a preference for high-calorie items, and this might be correlated with high buy propensities on all items. Therefore it is worth testing how well models, in particular those employing Binary-Choice features, do when prediction is entirely between subject.

We conduct the between-subject analysis by randomly selecting 5 out of 55 subjects, with all their observations, to be the test sample. We estimate our models on the remaining 50 subjects, using cross-validation procedures similar to those outlined in Section IIIB. We repeat the random sampling 50 times and average our results. The MSE results are displayed in Figure 7, using the same list of feature spaces as those displayed in Figure 5. Bars in Figure 7 are grouped by algorithm and feature space, with three bars (within and between subjects, between-subjects, between-items) per group. All between-subjects models (middle bar) perform

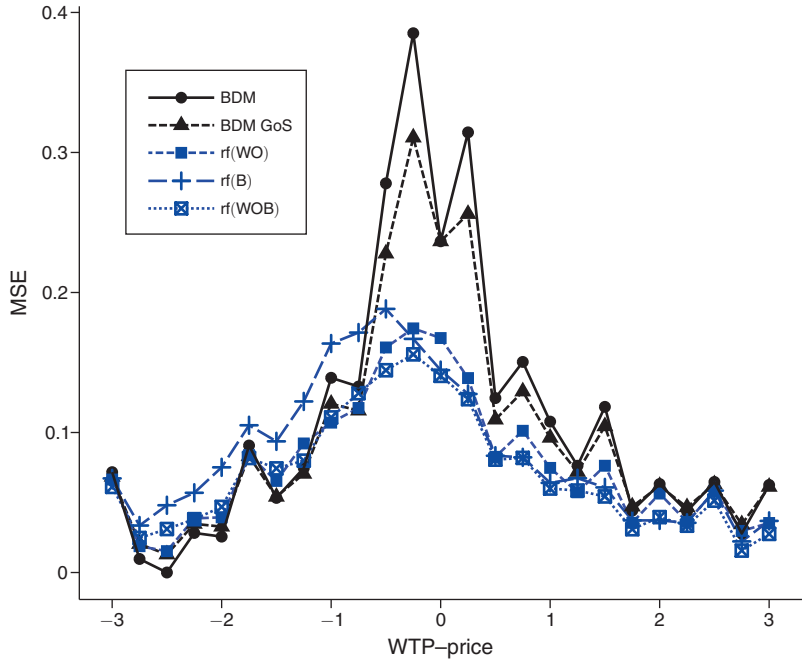

FIGURE 6. PERFORMANCE OF BDM AND SUPERVISED MACHINE LEARNING AS A FUNCTION OF CONSUMER SURPLUS

Notes: Out-of-sample MSE estimated on 440 hold-out observations. Since samples are random, we repeat the estimation 50 times and MSE is averaged.

worse than their respective baseline models (left bar), but the drop in performance varies. As the out-of-sample performance for logit with the larger feature spaces was much poorer than for either lasso or random forest, this section will focus on lasso and random forest. A complementary heatmap and a complete table with all results is provided in the online Appendix in Figure C3 and Table C3.

The emerging pattern is that between-subjects prediction increases MSE more for the Binary-Choice models than for the WTP models. Consider the lasso results first. The Binary-Choice model, lasso(B), performs better when using within- and between-subject data than the WTP model, lasso(WO), with respective full-sample MSE of 0.1042 (0.0012) and 0.1147 (0.0013). However, with only between-subject data, the lasso(B) MSE increases by 63.8 percent to 0.1707 (0.0040) while the lasso(WO) MSE increases only by 23.3 percent to 0.1415 (0.0039). We see the same pattern with the random forest models: rf(WO) and rf(B) perform almost equivalently at 0.1026 (0.0012) and 0.1010 (0.0012) MSE with within- and between-subject data. In the between-subject analysis rf(WO) increases by 29.2 percent to 0.1326 (0.0027) and rf(B) increases by 42.2 percent to 0.1436 (0.0027). So, while the WTP data and Binary-Choice data perform about equally in within and between analysis, Binary-Choice data performs worse in between-subject analysis.

Nonetheless, all of the SML models suffer when moving from within and between subjects, to between-subjects. The BDM benchmark MSE of 0.1458 and the BDM Grain of Salt benchmark MSE of 0.1372 pose more challenging hurdles. Whereas

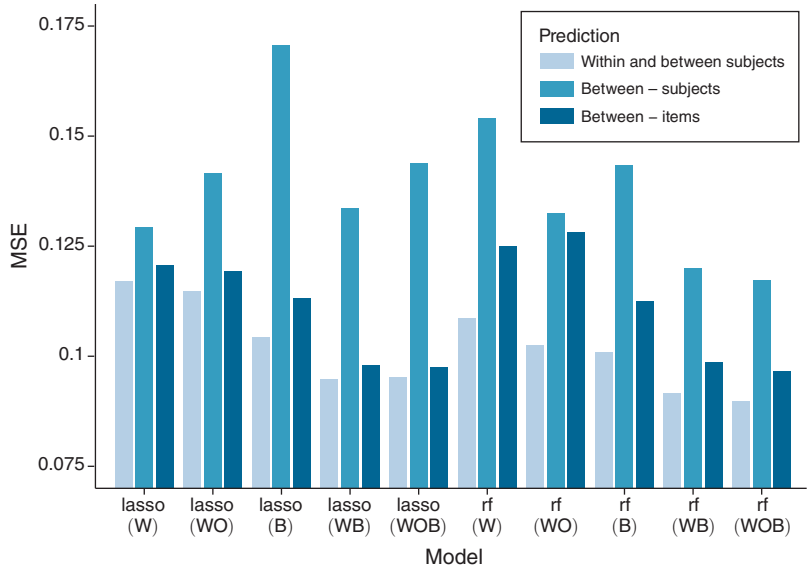

FIGURE 7. TEST SAMPLE—HOLDING OUT OBSERVATIONS VERSUS HOLDING OUT SUBJECTS

Notes: Summary plot of out-of-sample MSE results for high-dimensional methods, using the full available sample of features.

in the within and between models (left bars in Figure 7) all 10 models do better than BDM Grain of Salt, only half of the models do better than BDM Grain of Salt in the between-subject models. There is an important takeaway here. When the prediction objective is entirely between subject, "adding a grain of salt" to the direct elicitation method raises the bar for SML. This could reassure economists who wish to use the BDM for prediction purposes in certain scenarios.

We conduct a similar between-item exercise in which we leave out two items and all their observations in the test sample. The only necessary change in feature spaces involves dropping the item fixed effects. We estimate our models on the remaining 18 items. As before, we repeat the random sampling 50 times and average our results. This would be akin to having data on an existing set of customers as well as their incentivized preferences for a novel good and then predicting whether they buy this novel good at various prices. These results are summarized in the set of right bars in Figure 7. Again, the best-performing model is rf(WOB). The models that use Binary-Choice data, lasso(B) and rf(B), perform better than their respective between-item analysis with the models that use WTP data, lasso(WO) and rf(WO). Overall, it appears that while WTP data is better for predicting between-subject, Binary-Choice data is slightly better for predicting between-item.[16]

[16] In online Appendix C.4 we also consider rf(WO) and rf(WOB) further in between-subject comparisons. Specifically, we discuss some scenarios in which Binary-Choice data might be more useful for helping with between-subject prediction.

Visual inspection of the bar plot in Figure 7 reveals a clear pattern between columns. Whereas the average increase in MSE across all listed random forest models for the between-subjects analysis is 35.0 percent, the average increase in MSE for the between-item analysis is 13.4 percent. In other words, the increase in MSE is *smaller* for between-item than between-subject. A parallel result holds for lasso (38.0 percent for between-subjects, and 4.3 percent for between-items). In these cases, both algorithms face a greater challenge for prediction on new individuals compared to new items.

E. *Which Elicitation to Use?*

The results lead to a natural question: "When would a practitioner prefer to use WTP elicitations, Binary-Choice elicitations, or both?" WTP data predicts slightly better in our main analysis and much better in the between-subjects analysis. It is also important to take into consideration the costs of these elicitations. The average amount of time required for a Binary-Choice question is about two seconds, and the average amount of time for a WTP question is eight seconds. The time required to read the instructions for the Binary-Choice Task was 1 minute 20 seconds, and for the BDM it was 4 minutes 30 seconds. There were a total of 190 Binary-Choice questions and 20 WTP questions, which leads to an average length of 7 minutes 40 seconds for the Binary-Choice Task and 7 minutes 10 seconds for the BDM-Task. At the scale we used, the time costs are approximately equivalent. However, there are still some reasons to prefer the Binary-Choice Task. If a researcher wants to keep the elicitation very short, the Binary-Choice Task has a much shorter set of instructions and is more intuitive. Likewise, if a researcher has concerns that the population will struggle to understand the BDM rules, a Binary-Choice Task may be preferred. If time is not an issue, a researcher may wish to conduct both elicitations as they may have different sets of strengths and weaknesses and, as several of our results have demonstrated, the information contained in the two is not redundant.

Perhaps there are other elicitations that generate even more predictive data. We speculate that there may be considerable gains in prediction from other elicitation methods and by the use of non-choice survey data (Netzer et al. 2008). Bernheim et al. (2013) show that this can be a productive avenue when trying to predict aggregate behavior.

VII. Implications

A. *Revenue-Maximizing Prices*

Ultimately, the value of better prediction comes in the form of better action. There are many reasons why one may want to uncover a person's valuation for goods. Practitioners may wish to measure aggregate demand curves and heterogeneity of demand for a wide range of purposes, not limited to profit maximization, optimal taxation, and to measure the effects of an event or policy change. Mechanisms that include SML could potentially allocate resources more efficiently given that SML

more accurately uncovers latent valuations. Perhaps the most obvious application would be to use the results of direct elicitation and SML for revenue maximization.

The BDM has an immediate implied revenue-maximizing price—the reported WTP. We can use our predictive model to estimate a revenue-maximizing price by simply maximizing $p \times \Pr(buy|p)$. To control for data quality, in both the BDM prediction and SML prediction we use the same WTP data. We therefore use the top-performing algorithm with this data, the rf(WO) model. Figure 8, panel A displays a scatter plot of computed revenue-maximizing prices as a function of WTP, at the person-item level.

The revenue-maximizing prices are highly correlated with WTP ($\rho = 0.7576$). Nonetheless there is an obvious discrepancy. The average difference between WTP and the estimated optimal price is $-\$0.2206$, and the mean WTP is \$2.03, leading to an average difference of -10.8 percent. This difference again represents a systematic downward bias in WTP, as evidenced by more points being above the 45-degree line in Figure 8, panel A. It may be tempting to simply correct the BDM predictions by adding the average bias to the WTP to obtain a WTP′, but this would not lead to the correct conclusions as the price elasticities of the goods may vary. Indeed, the scatter plot in Figure 8, panel A shows that this bias is not evenly distributed.

Looking at average absolute difference reveals the scale for potential improvement: the average absolute difference across all person-item pairs is \$0.5788, representing a 26.8 percent difference in the pricing of the product. Given that the SML model has substantially lower MSE, it is likely that the earnings of the rf(WO) would be considerably higher. We estimate the increased revenue from rf(WO) in the next section.

B. *Estimated Revenue*

In order to compare revenue under the revenue-maximizing prices, we need to make some assumptions about the true probability of purchase as a function of price. Ideally, we want each individual's item-specific stochastic demand curves. In our setting this cannot be perfectly observed, only estimated. We use rf(WOB), our best-performing model, to estimate these stochastic demand curves. This model uses random forest on all of the Binary-Choice and WTP data. This model yields 1,100 individual $\times$ item stochastic-demand curves. We then compare the expected revenue of setting the price equal to the person's WTP for an item, to the expected revenue of setting the price equal to the revenue-maximizing price under the rf(WO) model.

The average expected revenue of setting the price to the WTP is \$1.259 and the average expected revenue of setting the price equal to the revenue-maximizing price under the rf(WO) model is \$1.621. Thus, setting the price according to random forest yields an estimated increase in revenue by 28.8 percent over the raw WTP values without the use of any additional data. Figure 8, panel B shows a histogram of the gain in revenue by using random forest across the 1,100 individual-item pairs. It is evident that the distribution's mass is predominantly positive (95.2 percent > 0). As a benchmark, the average gross margin across all sectors in the United States is

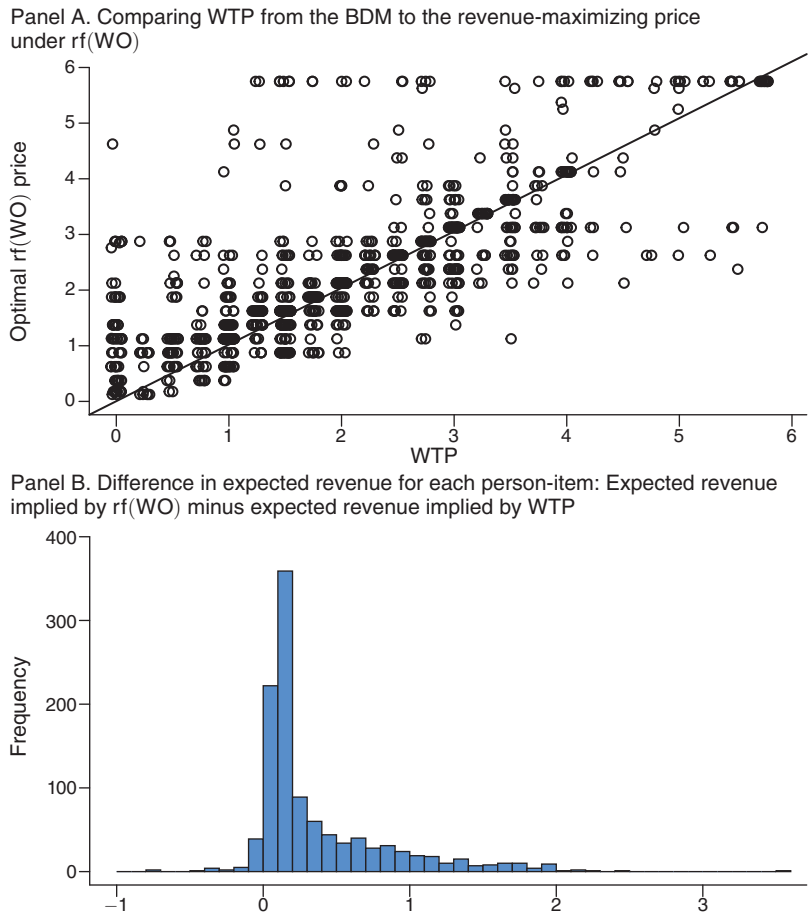

FIGURE 8. COMPARING REVENUE-MAXIMIZING PRICES

approximately 37 percent as of January 2019 (Damodaran 2019). Thus, the improvement from using SML over the BDM price is substantial.[17]

C. *Aggregate Demand*

Our within-subject analysis can be seen in the context of price discrimination. Many firms, particularly those online, monitor consumer behavior. This data could be used to offer individual-specific prices. But what if price discrimination is not

[17] As a robustness check on our choice of rf(WOB) as the "ground truth" for individual stochastic demand curves, we used a second statistical model. An important qualification of any alternative technique here is the ability to perform as well as random forest. Gradient boosted regression often performs as well or nearly as well as random forest in many applications. Using boosted regression with the WOB features, we find an MSE of 0.092, which is better than any nonrandom forest model in our analysis, though still worse than rf(WOB). With boosted(WOB) in place of rf(WOB) as the assumed "ground truth," the average expected revenue of using the WTP is $1.21 while it is $1.31 using rf(WO), yielding a 8.25 percent increase in revenue. Out of 1,100 individual-item pairs, the expected revenue from rf(WO) exceeds the expected revenue from the BDM 79.2 percent of the time.

available as a strategy to the firm? Perhaps prices are required by the market context to be uniform. Do the SML models imply an aggregate demand curve that matches the aggregate demand curve derived from the BDM? On the one hand, the systematic understatement of valuations in the BDM would suggest a leftward shift of the demand curve relative to an SML-derived demand curve. On the other hand, if the bias in the BDM is sufficiently small, and if the valuations are sufficiently noisy, the aggregation may wash out much of the differences between the BDM-derived aggregate demand and the SML-derived aggregate demand.

We can construct a useful comparison with our data. In Figure 9 we display the aggregate demand curves for all 20 products. We display both the aggregate demand implied from the BDM in black and the aggregate demand implied from rf(WOB) in blue. Some aggregate demand curves have high overlap between the BDM and rf(WOB) versions, such as for Luna Choco (Figure 9, panel I) or Simply Blueberries (Figure 9, panel R). Other curves have larger discrepancies such as with Hershey's chocolate (Figure 9, panel F) or Naked Green juice (Figure 9, panel J).

There are a few salient patterns across these figures. Demand for rf(WOB) is shifted outward for medium to high prices and shifted inward for low prices. For most of the items there is a single intersection point, and it appears as if one curve is a rotation of the other curve around the intersection point. This intersection point varies across the products (e.g., near $1.50 for Clif Bar, but near $2.75 for Green and Black's). So, there does not seem to be any single transformation of the BDM curve that would consistently produce the rf(WOB) curve over the 20 items. Additionally, upon careful inspection, one can observe that many of the rf(WOB) demand curves bend backward slightly for high prices. For example, the Hershey's curve in panel F might be the most distinctive. This is an instructive point, as the SML algorithms are theory-free and hence do not impose the law of demand.

VIII. Discussion

A. *Connection to Theory*

Direct elicitation is appreciated for the way it tightly connects theory to data. SML approaches are often promoted or disparaged, depending on the audience and context, for being "theory free." However, theory and SML are not inherently in conflict. In fact, our view is that these two approaches are highly complementary. Direct elicitation, because it incentivizes behavior, and because it tightly links that behavior to the latent object of interest, may produce data that is highly predictive. When married to SML, direct elicitation data can be debiased, and patterns in the data could be more fully exploited. For example, we found that the WTP values of the 19 other goods were predictive of buying a given good.

One potential criticism of our SML approach is that we did not generate a reservation value but instead generated a stochastic demand curve. If one's purpose is to construct a reservation value for each individual-item then our exercise is incomplete. This is a reasonable objective if one is testing a theory that makes predictions about reservation values, or if stochastic demand curves are too unwieldy objects for comparison. Our SML approach can deliver this with one additional step: by

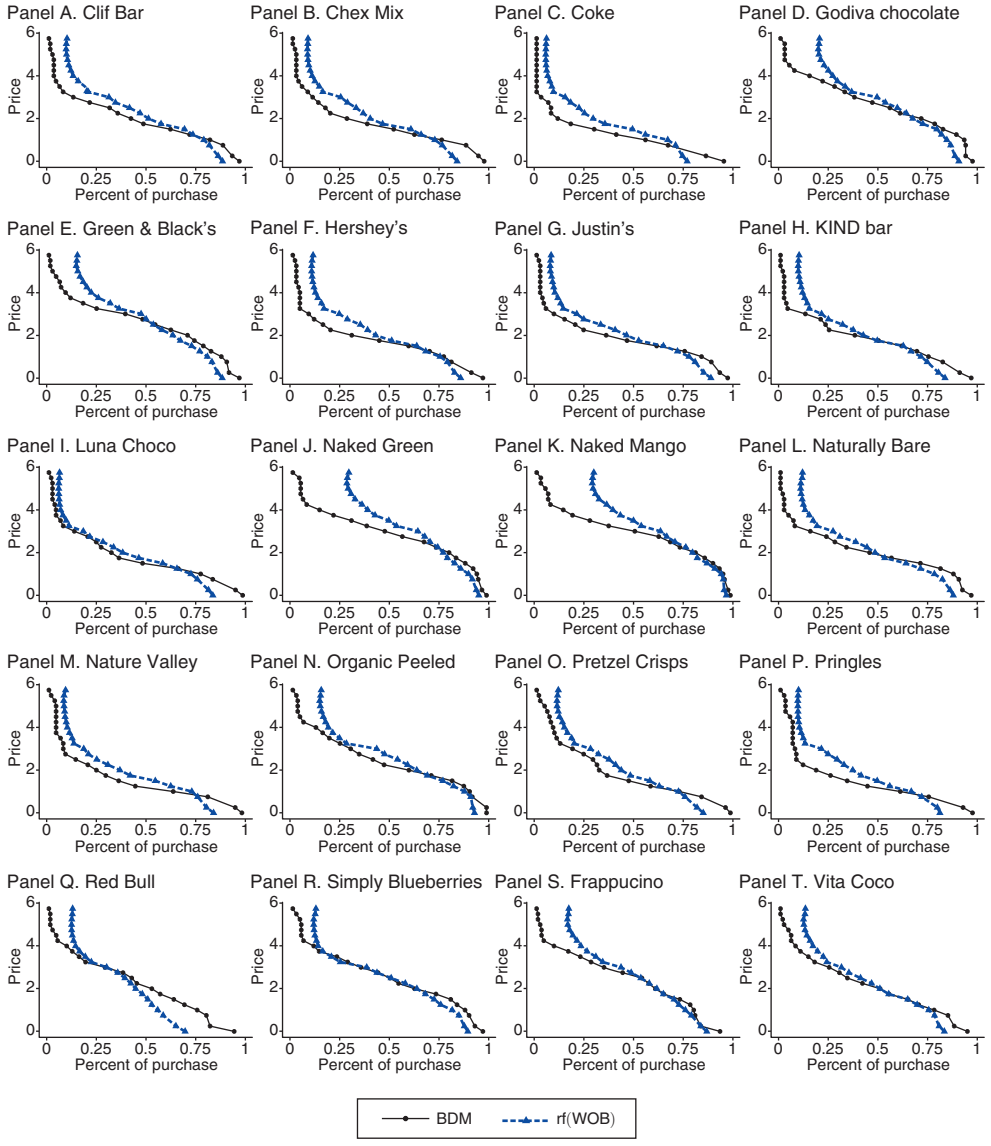

FIGURE 9. AGGREGATE DEMAND CURVES BY ITEM

Notes: Black curve represents the aggregate demand implied by the BDM responses. Blue curve represents the aggregate demand implied by the model rf(WOB).

constructing a mapping from stochastic demand curves to reservation values. For example, one can define the mapping, "the reservation value is the price at which the person is exactly 50 percent likely to buy the good."

As is common in experiments with a large number of choices, we observe inconsistent behavior across trials. There are several plausible explanations. Some possibilities might be that people make mistakes, or individuals do not know precisely their preferences, or randomize when they are close to indifference. Framing might

also cause systematic shifts in behavior across tasks. Identifying the source is beyond the scope of this paper, but all of these are likely happening to some degree (Brown and Healy 2018; Agranov and Ortoleva 2017). To the extent that people randomize on the Buy-Task, this should inflate MSE for all of our predictions. To the extent that people randomize on the BDM or Binary-Choice Task, this would cause measurement error. In OLS, classical measurement error on the independent variables attenuates the coefficients. In the context of our study, this too would inflate the MSE. It is also possible that some subjects are far noisier than other subjects. Our high-dimensional SML models provided data to address this as best as is possible. Indeed, we interact individual fixed-effects with WTP and the Binary-Choice features (see online Appendix B.2), so we allow for heterogenous effects across individuals. This affords the predictive model the ability to ignore the behavior of exceptionally noisy individuals.

B. *Applications*

Applied more broadly, SML could potentially be used to identify specific aspects of purchase decisions, such as product attributes (e.g., nutritional data or branding) or the context in which a purchase is made, which could be influential. Including additional data such as this may improve SML predictions of purchase decisions. So long as the computational burden is manageable, our modular feature space could be expanded to identify the most predictive kinds of data. If highly predictive features are not well explained by theory or not captured by WTP elicitation, a gap in theory has been identified. This could facilitate the development of new theories (Fudenberg and Liang 2019).

Perhaps an under-explored use of measuring consumer valuations is for welfare analysis. Applications could be in many domains, including cost-benefit analysis or measuring the welfare effects of a policy or event. For example, the prevalence and growing importance of the digital economy presents a measurement challenge. Many digital goods are free and have no clear market price. Several recent efforts attempt to identify the value of these goods, such as Facebook (Brynjolfsson et al. 2019; Allcott et al. 2020; Mosquera et al. 2020). Both Allcott et al. (2020) and Mosquera et al. (2020) ran experiments and utilized the BDM to estimate willingness-to-accept the deactivation of personal Facebook accounts. Our work can serve as a template for extending these attempts to measure consumer surplus in the digital economy.

How does this approach of using lab tasks with SML relate to widely used methods for demand-system estimation, such as BLP (Berry et al. 1995, 2004)? BLP is a structural estimation method that uses data on consumer purchases of goods within an industry and instruments on price to estimate price elasticities for the goods, including both own and cross-price elasticities. The output of this estimation allows one to conduct powerful counterfactual analysis. In addition to being able to estimate the effect of a price change of one product on the demand for all other products in the market, BLP allows the economist to estimate the optimal price-markups on products, and predict how the introduction or elimination of a product would affect demand on competing products. One of the key assumptions in BLP, and many other frameworks for estimating consumer demand

(Nevo 2011), is that products are bundles of characteristics and consumers choose products based on their demand for those characteristics.

As is, our approach in this paper cannot generate the same kind of output as BLP, that is cross-price elasticities for a demand system. However, with more structure and more tasks an approach similar to ours could generate cross-price elasticities. Viewed in a very broad sense, all of our tasks are situated within the context of binary menus. The Binary-Choice Task is a choice between two goods, but without any prices. The Buy-Task is a choice between obtaining a specific good at a price or getting nothing. The BDM can also be viewed in this way, as ultimately a randomly drawn price will result in a purchase or not. Hence, we have no tasks in which we vary the price of one good and observe the effect on the demand for another. What we can estimate are own-price elasticities, which are already visualized in the stochastic-demand curves in Figure 9. To estimate cross-price elasticities one could include additional choice tasks with larger menus. For example, if subjects chose from a series of triplets, two goods at specific prices or nothing, one could use the data to estimate cross-price elasticities.

More generally, our approach does allow for counterfactual analysis (akin to what we do in Section VII), but only within the domain of the kinds of tasks we ran on subjects. The approach can be augmented with more structure to allow for counterfactual predictions outside of the domain of the tasks. For example, to estimate a cross-price elasticity for items A and B we can make additional assumptions such as: a consumer's choice responsiveness can be explained by a logistic curve as a function of the relative consumer surplus between A and B. Figure 3 shows a person's responsiveness between buying A at price p_A or not as a function of the person's surplus as implied by the BDM. One could impose a common logistic structure to imply a purchase probability of A over B as the difference in the surpluses vary, and this would imply a cross-price elasticity. The takeaway here is that SML alone cannot do counterfactual analysis outside the domain of the tasks elicited, but its estimates could be integrated into structural models.

Another potential application of the framework used here could be as a supplement to structural estimation of demand systems. A major challenge for structural estimation of demand is finding good instruments for the price (Nevo 2011). We speculate that, in contexts for which good instruments are not available, elicitation with SML could identify parameters such as own-price elasticity, as described above. These parameters could then be fed into a more comprehensive structural model, such as BLP.

Another natural domain to consider is the growing literature on theories of stochastic choice. Along the lines of Fudenberg et al. (2022), our SML approach can help test theories or identify gaps. He and Natenzon (2020) develop a model of stochastic choice that includes a function of the similarity between products.[18] However, the similarity of two products is not always readily apparent or observable. Our SML methods and utilization of WTP might help. The WTP data that we elicited in our experiment may be one way to uncover some of the similarity between

[18] The literature shows that people are more likely to substitute a product with a similar product. For example, if the price on a coupe goes up, consumers are likely to substitute with a different model of coupe than with a minivan.

products. People who are likely to value a kind of product are also likely to value very similar products; SML could search for such structure in correlations between elicited WTPs. So, while our paper has primarily focused on how our Binary-Choice data can complement the BDM data via SML, it is possible for SML to enhance our understanding of Binary-Choice tasks using other kinds of data like WTP.[19]

Finally, researchers may wish to apply similar methods for estimating beyond unit demand (Hanemann 1984). Consumers often not only decide whether or not to purchase a soda, but how much soda they would like to buy. Determining how SML can improve predictions in this domain is a promising avenue for future work. There may be many viable paths, but one could be an extension of our approach. To begin, a researcher would need an elicitation method for multiunit demand. An obvious option would be to use the BDM elicitation first for one unit, separately for two units, and so on. Binary-Choice data also could be collected, wherein the subject chooses to buy $n > 0$ units of a good at price p, or not. An outcome measure, such as purchase decisions for multiple units at a specific price, would also be needed. Last, the researcher would need a statistical method (i.e., a method such as OLS or random forest) to connect the elicitation data to the outcome data.

IX. Conclusion

In this paper, we demonstrate that SML can be an effective alternative for recovering individual-level latent preferences. SML offers sizable advantages over direct elicitation. First and foremost, we show that using SML facilitates the combination of distinct kinds of preference-elicitation data to improve predictions of consumer demand. Nested within this contribution is the result that SML can enhance BDM predictions of purchase decisions. Second, we demonstrate that data generated from direct elicitation is not necessary for prediction of purchase decisions. Specifically, we indirectly elicit demand using Binary-Choice data, which is fast, easily comprehended by subjects, and predicts better than the BDM.

Our results also show that more data cannot substitute for a superior statistical algorithm. With our data, for instance, random forest does asymptotically better than logit, while logit does asymptotically better than the BDM. Indeed, statistical methods are not substitutes for each other, and neither are data types. Conversely, we find that elicitation (i.e., WTP) and choice (i.e., Binary-Choice) data are complementary, and can be combined to improve predictions.

The basic approach presented in this paper can be extended to estimating other individual-latent preference parameters such as risk preferences, time preferences, or option attributes of importance that are not directly observable to the researcher. We think that SML may be especially useful in capturing unobserved consumer heterogeneity given the difficulty consumers will have in directly stating such differences within most traditional elicitations.

The finding that SML models predict better than direct elicitations could have useful implications for theory. For example, the systematic bias in the BDM prediction

[19] We thank an anonymous reviewer for directing our attention to this paper.

suggests that there are other cognitive processes at play during the BDM that are not at play in different-context purchase decisions. This suggests that there may be more predictive theories waiting to be discovered that are also more parsimonious than SML models. Future work that could explain this gap may lead to improved theories of choice and innovative hypothesis generation.

REFERENCES

Agranov, Marina, and Pietro Ortoleva. 2017. "Stochastic Choice and Preferences for Randomization." *Journal of Political Economy* 125 (1): 40–68.

Allcott, Hunt, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. 2020. "The Welfare Effects of Social Media." *American Economic Review* 110 (3): 629–76.

Anderson, Lisa R., and Jennifer M. Mellor. 2009. "Are Risk Preferences Stable? Comparing an Experimental Measure with a Validated Survey-Based Measure." *Journal of Risk and Uncertainty* 39: 137–60.

Andersen, Steffen, Glenn W. Harrison, Morten I. Lau, and E. Elisabet Rutström. 2008. "Lost in State Space: Are Preferences Stable?" *International Economic Review* 49 (3): 1091–1112.

Ascarza, Eva. 2018. "Retention Futility: Targeting High-Risk Customers Might Be Ineffective." *Journal of Marketing Research* 55 (1): 80–98.

Athey, Susan, and Guido W. Imbens. 2019. "Machine Learning Methods that Economists Should Know About." *Annual Review of Economics* 11: 685–725.

Becker, Gordon M., Morris H. DeGroot, and Jacob Marschak. 1964. "Measuring Utility by a Single-Response Sequential Method." *Behavioral Science* 9 (3): 226–32.

Bernheim, B. Douglas, Daniel Bjorkegren, Jeffrey Naecker, and Antonio Rangel. 2013. "Non-choice Evaluations Predict Behavioral Responses to Changes in Economic Conditions." NBER Working Paper No. 19269.

Berry, James, Greg Fischer, and Raymond Guiteras. 2020. "Eliciting and Utilizing Willingness to Pay: Evidence from Field Trials in Northern Ghana." *Journal of Political Economy* 128 (4): 1436-73.

Berry, Steven, James Levinsohn, and Ariel Pakes. 1995. "Automobile Prices in Market Equilibrium." *Econometrica* 63 (4): 841–90.

Berry, Steven, James Levinsohn, and Ariel Pakes. 2004. "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market." *Journal of Political Economy* 112 (1): 68–105.

Bohm, Peter, Johan Lindén, and Joakim Sonnegård. 1997. "Eliciting Reservation Prices: Becker–DeGroot–Marschak Mechanisms versus Markets." *Economic Journal* 107 (443): 1079–89.

Brandts, Jordi, and Gary Charness. 2011. "The Strategy versus the Direct-Response Method: A First Survey of Experimental Comparisons." *Experimental Economics* 14: 375–98.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45: 5–32.

Brown, Alex L., and Paul J. Healy. 2018. "Separated Decisions." *European Economic Review* 101: 20–34.

Brynjolfsson, Erik, Avinash Collis, and Felix Eggers. 2019. "Using Massive Online Choice Experiments to Measure Changes in Well-Being." *Proceedings of the National Academy of Sciences* 116 (15): 7250–55.

Cason, Timothy N., and Charles R. Plott. 2014. "Misconceptions and Game Form Recognition: Challenges to Theories of Revealed Preference and Framing." *Journal of Political Economy* 122 (6): 1235–70.

Charness, Gary, Anya Samek, and Jeroen van de Ven. 2022. "What is Considered Deception in Experimental Economics?" *Experimental Economics* 25: 385–412.

Clithero, John A. 2018. "Improving Out-of-Sample Predictions Using Response Times and a Model of the Decision Process." *Journal of Economic Behavior and Organization* 148: 344–75.

Clithero, John A., Jae Joon Lee, and Joshua Tasoff. 2023. "Replication data for: Supervised Machine Learning for Eliciting Individual Demand." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.38886/E180561V1.

Damodaran, Aswath. 2019. "Operating and Net Margins – Margins by Sector (US)," January. http://pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/margin.html.

de Meza, David, and Diane Reyniers. 2013. "Debiasing the Becker-DeGroot-Marschak Valuation Mechanism." *Economics Bulletin* 33 (2): 1446–56.

Ding, Min. 2007. "An Incentive-Aligned Mechanism for Conjoint Analysis." *Journal of Marketing Research* 44 (2): 214–23.

Eckel, Catherine C., and Philip J. Grossman. 2008. "Forecasting Risk Attitudes: An Experimental Study Using Actual and Forecast Gamble Choices." *Journal of Economic Behavior and Organization* 68 (1): 1–17.

Freeman, David J., Yoram Halevy, and Terri Kneeland. 2019. "Eliciting Risk Preferences Using Choice Lists." *Quantitative Economics* 10 (1): 217–37.

Fudenberg, Drew, and Annie Liang. 2019. "Predicting and Understanding Initial Play." *American Economic Review* 109 (12): 4112–41.

Fudenberg, Drew, Jon Kleinberg, Annie Liang, and Sendhil Mullainathan. 2022. "Measuring the Completeness of Economic Models." *Journal of Political Economy* 130 (4): 956–90.

Gillen, Ben, Erik Snowberg, and Leeat Yariv. 2019. "Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study." *Journal of Political Economy* 127 (4): 1826–63.

Hagen, Linda, Kosuke Uetake, Nathan Yang, Bryan Bollinger, Allison J. B. Chaney, Daria Dzyabura, Jordan Etkin, and et al. 2020. "How Can Machine Learning Aid Behavioral Marketing Research?" *Marketing Letters* 31: 361–70.

Hanemann, W. Michael. 1984. "Discrete/Continuous Models of Consumer Demand." *Econometrica* 52 (3): 541–61.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Minding, Inference, and Prediction.* New York: Springer.

He, Junnan, and Paulo Natenzon. 2020. "Moderate Expected Utility." Unpublished.

Holt, Charles A., and Susan K. Laury. 2002. "Risk Aversion and Incentive Effects." *American Economic Review* 92 (5): 1644–55.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2015. *An Introduction to Statistical Learning*. New York: Springer.

Kaas, Klaus Peter, and Heidrun Ruprecht. 2006. "Are the Vickrey Auction and the BDM Mechanism Really Incentive Compatible? – Empirical Results and Optimal Bidding Strategies in Cases of Uncertain Willingness-to-Pay." *Schmalenbach Business Review* 58 (1): 37–55.

Krajbich, Ian, Carrie Armel, and Antonio Rangel. 2010. "Visual Fixations and the Computation and Comparison of Value in Goal-Directed Choice." *Nature Neuroscience* 13 (10): 1292–98.

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, and et al. 2009. "Computational Social Science." *Science* 323 (5915): 721–23.

Lehman, Sebastian. 2015. "Toward an Understanding of the BDM: Predictive Validity, Gambling Effects, and Risk Attitude." Unpublished.

Matz, Sandra C., Cristina Segalin, David Stillwell, Sandrine R. Müller, and Maarten W. Bos. 2019. "Predicting the Personal Appeal of Marketing Images Using Computational Methods." *Journal of Consumer Psychology* 29 (3): 370–90.

Mazar, Nina, Botond Koszegi, and Dan Ariely. 2014. "True Context-Dependent Preferences? The Causes of Market-Dependent Valuations." *Journal of Behavioral Decision Making* 27 (3): 200–08.

Miller, Klaus M., Reto Hofstetter, Harley Krohmer, and Z. John Zhang. 2011. "How Should Consumers' Willingness to Pay be Measured? An Empirical Comparison of State-of-the-Art Approaches." *Journal of Marketing Research* 48 (1): 172–84.

Mosquera, Roberto, Mofioluwasademi Odunowo, Trent McNamara, Xiongfei Guo, and Ragan Petrie. 2020. "The Economic Effects of Facebook." *Experimental Economics* 23 (2): 575–602.

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2): 87–106.

Müller, Holger, and Stefen Voigt. 2010. "Are There Gambling Effects in Incentive-Compatible Elicitations of Reservation Prices? - An Empirical Analysis of the BDM-Mechanism." Unpublished.

Muschelli, John. 2020. "ROC and AUC with a Binary Predictor: A Potentially Misleading Metric." *Journal of Classification* 37: 696–708.

Naecker, Jeffrey. 2015. "The Lives of Others: Predicting Donations with Non-Choice Responses." Unpublished.

Netzer, Oded, Alain Lemaire, and Michal Herzenstein. 2019. "When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications." *Journal of Marketing Research* 56 (6): 960–80.

Netzer, Oded, Olivier Toubia, Eric T. Bradlow, Ely Dahan, Theodoros Evgeniou, Fred M. Feinberg, Eleanor M. Feit, and et al. 2008. "Beyond Conjoint Analysis: Advances in Preference Measurement." *Marketing Letters* 19: 337.

Nevo, Aviv. 2011. "Empirical Models of Consumer Behavior." *Annual Review of Economics* 3 (1): 51–75.

Noussair, Charles, Stephane Robin, and Bernard Ruffieux. 2004. "Revealing Consumers' Willingness-to-Pay: A Comparison of the BDM Mechanism and the Vickrey Auctions." *Journal of Economic Psychology* 25 (6): 725–41.

Peysakhovich, Alexander, and Jeffrey Naecker. 2017. "Using Methods from Machine Learning to Evaluate Behavioral Models of Choice under Risk and Ambiguity." *Journal of Economic Behavior and Organization* 133: 373–84.

Philiastides, Marios G., and Roger Ratcliff. 2013. "Influence of Branding on Preference-Based Decision Making." *Psychological Science* 24 (7): 1208–15.

Plott, Charles R., and Kathryn Zeiler. 2005. "The Willingness to Pay - Willingness to Accept Gap, the 'Endowment Effect,' Subject Misconceptions, and Experimental Procedures for Eliciting Valuations." *American Economic Review* 95 (3): 530–45.

Schmidt, Jonas, and Tammo H.A. Bijmolt. 2020. "Accurately Measuring Willingness to Pay for Consumer Goods: A Meta-Analysis of the Hypothetical Bias." *Journal of the Academy of Marketing Science* 48: 499–518.

Smith, Alec, B. Douglas Bernheim, Colin F. Camerer, and Antonio Rangel. 2014. "Neural Activity Reveals Preferences Without Choices." *American Economic Journal: Microeconomics* 6 (2): 1–36.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society* 58 (1): 267–88.

Tymula, Agnieszka, Eva Woelbert, and Paul Glimcher. 2016. "Flexible Valuations for Consumer Goods as Measured by the Becker-DeGroot-Marschak Mechanism." *Journal of Neuroscience, Psychology, and Economics* 9 (2): 65–77.

Urbancic, Michael B. 2011. "Testing Distributional Dependence in the Becker-DeGroot-Marschak Mechanism." Unpublished.

Varian, Hal R. 2014. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives* 28 (2): 3–28.

Wang, Tuo, Ramaswamy Venkatesh, and Rabikar Chatterjee. 2007. "Reservation Price as a Range: An Incentive-Compatible Measurement Approach." *Journal of Marketing Research* 44 (2): 200–13.

Wertenbroch, Klaus, and Bernd Skiera. 2002. "Measuring Consumers' Willingness to Pay at the Point of Purchase." *Journal of Marketing Research* 39 (2): 228–41.